ALLC: The European Association for Digital Humanities (ALLC)
Association for Computers and the Humanities (ACH)
Canadian Society for Digital Humanities / Société
canadienne des humanités numériques (CSDH/SCHN)
centerNet
Australasian Association for Digital Humanities (aaDH)
Japanese Association for Digital Humanites (JADH)

Digital Humanities 2013

Conference Abstracts

University of Nebraska–Lincoln, USA

16-19 July 2013

Primary Encoder: Laura Weakly
Designer: Karin Dalziel
Encoders and Proofreaders:
Matt Bosley
Matthew Lavin
Elizabeth Lorang
Keith Nickum
Erin Pedigo
Hannah Vahle
Illustration: University of Nebraska–Lincoln Communications
Available online at: http://dh2013.unl.edu

ISBN: 978-1-60962-036-3 (paperback)
Published by the Center for Digital Research in the Humanities

The 24th Joint International Conference of the Association for Literary and Linguistic Computing and Association for Computers and the Humanities
and
The 5th Joint International Conference of the Alliance of Digital Humanities Organizations

International Program Committee

Bethany Nowviskie, chair

Craig Bellamy
John Bradley
Paul Caton
Carolyn Guertin
Ian R. Johnson
Sarah Potvin
Jon Saklofske
Sydney Shep
Tomoji Tabata
Melissa Terras
Deb Verhoeven
Ethan Watrall

Local Organizing Committee

Katherine Walter, co-local organizer
Kenneth Price, co-local organizer

Joan Barnes
Karin Dalziel
Laura Weakly
Kim Weide
Annette Wetzel
Jeremy Wurst

Conference Sponsors

University of Nebraska–Lincoln
University of Nebraska–Lincoln Office of Research and Economic Development
Senior Vice Chancellor of Academic Affairs Offices at the University of Nebraska–Lincoln
University of Nebraska–Lincoln Libraries
University of Nebraska–Lincoln Department of History

Conference Volunteers

DeeAnn Allison
David Arredondo
Molly Bass
Brett Barney
Brent Baum
Rebecca Bernthal
Janel Cayer
James Coltrain
Kiyomi Deards
Laura Dimmit
Mary Ellen Ducey
Jacob Friefeld
Amanda Gailey
Peter Gillon
Jolie Graybill
Patrick Graybill
Jaci Groves
Jennifer Isasi
Andrew Jewell
Matt Jockers
Matthew Lavin
Courtney Lawton
Sue Leach
Elizabeth Lorang
Courtney Lawton
Brandon Locke
Kevin McMullen
Melissa Moll
Joseba Moreno
Keith Nickum
Kayla Nielson
Erin Pedigo
Brian Pytlik Zillig
Stephen Ramsay
Svetlana Rasmussen
Janet Salvati
Robert Shepard
Frank Smutniak
Kristin Sorensen
Grace Thomas
William Thomas
Maggie Van Diest
Laura Weakly
Jean Williss
Judith Wolfe

Special thanks to Computer Operations and Research Services at UNL Libraries for technical support.

Welcome to Digital Humanities 2013

Katherine Walter
kwalter1@unl.edu
Co-director, Center for Digital Research in the Humanities, University of Nebrask–Lincoln

Kenneth Price
kprice2@unl.edu
Co-director, Center for Digital Research in the Humanities, University of Nebrask–Lincoln

Welcome to the University of Nebraska–Lincoln and to Digital Humanities 2013. The theme we have chosen for this year's conference is **"Freedom to Explore."** This theme seems appropriate given our Great Plains location and our belief that Digital Humanities is a new frontier in the Humanities.

In the 1840s, many pioneers traveled across Nebraska in wagon trains and hand-carts along the Great Platte River Road and wagon trails, such as the Oregon Trail–moving west to seek riches, land or religious freedom. Some people stayed, especially in response to the 1862 Homestead Act signed by President Abraham Lincoln. At that time many Civil War veterans, women and blacks homesteaded, as well as immigrants from such countries as Germany, Bohemia, Sweden and Denmark. Forty miles south of Lincoln, Nebraska today is the Homestead National Monument of America—worth seeing if you have time. Another place worth visiting is Scout's Rest, Buffalo Bill's ranch outside of North Platte.

Also in 1862, the last piece of Trans-Continental Railroad legislation was signed by President Lincoln. From Council Bluffs (just across the river from Omaha, Nebraska) entrepreneurs hastened to build the railroad west to Promontory Point in Utah. Today, the Union Pacific Railroad headquarters is in Omaha and Warren Buffett, the owner of the Burlington Northern Santa Fe (BNSF) railroad, lives in Omaha, so railroading is very much part of Nebraska's heritage. Those of you who will be going on the Nebraska History tour will see the Durham Western Heritage Museum in Omaha, located in a former Union Pacific station. If you are driving further east, we recommend visiting the Union Pacific Railroad Museum in Council Bluffs, Iowa.

For many years, Mexicans and Latinos from Central America have been migrating to Nebraska; and now Lincoln, Nebraska, is a refugee resettlement community with about 53 languages-of-origin spoken—languages of Southeast Asia, former Soviet bloc countries, the Middle East and Africa, as well as Spanish. Indian tribes in Nebraska today are the Omaha, Ponca, Dakota Sioux and the Winnebago, with other tribes having been relocated to reservations in Oklahoma or South Dakota during the nineteenth-century.

The University of Nebraska–Lincoln (UNL) was formed by the third major piece of legislation signed by President Lincoln in 1862—the Morrill Act. UNL is both a land-grant university and a university designated a "research university—very high activity" by the Carnegie Foundation. It has over 24,000 students from 130 countries. In 2010, we joined the Committee on Institutional Cooperation (CIC), a consortium of universities enrolling approximately half a million students each year, with approximately $7 billion in funded research, over 79 million library volumes, and 46,000 faculty. The CIC Digital Humanities Committee is now exploring how our institutions can collaborate on digital humanities research and teaching.

In 2005, the University approved the formation of the Center for Digital Research in the Humanities, and the commitment from this institution has been wonderful. We hope you can come to our open house during the conference and meet some of our faculty and graduate students.

Sixty plus years after Father Roberto Busa began working with IBM on the Index Thomisticus, digital humanities continues to explore and create new approaches for examining the humanities. Like our conference logo—the Western Meadowlark taking wing–DH is soaring!

Welcome from the Program Committee Chair

Bethany Nowviskie
bethany@virginia.edu
Scholars' Lab, University of Virginia, USA

2013 is a banner year, as we celebrate the 25th joint international conference of the Association for Computers and the Humanities (ACH) and ALLC: the European Association for Digital Humanities, organized in the past six years with a growing number of key collaborators and partnering professional organizations. These now include the Canadian Society for Digital Humanities (CSDH/SCHN), the Australasian Association for Digital Humanities (aaDH), and centerNet. As planning for next year begins, we welcome the contributions of the Japanese Association for Digital Humanities (JADH). But the conference hosted by our strong Alliance of Digital Humanities Organizations is not just an international affair: it is also wildly interdisciplinary and—like the community of practice we call "DH"—thoroughly inter-professional in character.

As in past years, the International Program Committee, aided by a host of volunteer peer reviewers, faced the problem of an abundance of riches. I would like to thank our splendid Nebraskan local organizers for accommodating a greater number of parallel sessions than usual, so that we could increase the diversity of offerings at *Digital Humanities 2013*. We look forward to hearing from all of you, and from our plenary speakers. These are David Ferriero, 10th Archivist of the United States; Isabel Galina of the Instituto de Investigaciones Bibliográficas at the National University of Mexico; and the winner of ADHO's highest honor, the Roberto Busa Award for outstanding lifetime achievements in the application technology to humanistic research: Professor Willard McCarty.

My sincere thanks go to fellow members of our stalwart Program Committee: Craig Bellamy (ACH); John Bradley (ALLC); Paul Caton (ACH); Carolyn Guertain (CSDH/SCHN); Ian Johnson (aaDH); Sarah Potvin (cN); Jon Saklofske (CSDH/SCHN); Sydney Shep (aaDH); Melissa Terras (ALLC, vice-chair); Tomoji Tabata (ALLC); Deb Verhoeven (aaDH); and Ethan Watrall (cN). I am likewise grateful to: ADHO's infrastructure committee (ably chaired by Chris Meister) for sometimes-heroic support of the conference system; to members of our Multilingual and Multicultural Issues Committee (led by Elisabeth Burr) for organizing translations of the CFP and advising the PC through the call and review process; and to the conference's local organizers (most especially Kay Walter and Karin Dalziel) for their responsiveness and invariable good cheer. I also wish to thank our 53 session chairs and other volunteers who will enrich the intellectual program in Lincoln, as well as the hundreds of active peer reviewers from around the world who generously provided up to six independent assessments for each of the nearly 350 submissions made to this year's conference.

Finally, I would like to thank ADHO's Conference Coordinating Committee (chaired by Ray Siemens), the ADHO Steering Committee (chaired by Neil Fraistat with very helpful contributions from John Nerbonne and Julia Flanders), and the wider DH community for support of a number of procedural changes I undertook this year with the endorsement and good energy of the International Program Committee. The goal of our experimentation was to advance an inclusive, fair, and welcoming peer review system, and to make a rigorous *Digital Humanities* vetting process as transparent, constructive, and collegial as possible. I hope that the openness of the system and breadth of its results are a good match for the big skies of America's heartland, and wish all of this year's attendees the freedom to explore.

Obituary for Prof. Lisa Lena Opas-Hänninen

Prof. Lisa Lena Opas-Hänninen of the University of Oulu and chair of the ALLC: The European Association for Digital Humanities, passed away in Helsinki on Feb. 2, 2013 after a long illness. She is survived by her husband, Prof. Heikki Hänninen of the University of Helsinki, and she will be remembered by generations of digital humanists.

Lisa Lena loved her work and her colleagues. She traveled a great deal in order to stay in personal contact with others in, around and beyond our disciplines. Conferences such DH 2013, which this book of proceedings describes, were her normal fields of activity, where she inevitably arrived early and stayed late, engaged everyone interested in new opportunities for international collaboration, attending innumerable meetings and talks, always with words of encouragement to younger scholars, with witty side remarks to those sitting nearby, and with invitations to discuss it all at more leisure over a drink later in the evening. The invitations were delivered in a collegial, almost conspiratorial manner! Those who accepted them were always delighted to find a good number of colleagues engaged in friendly banter and argument. We shall miss her for her contributions, for her welcoming and encouraging way, and for the feeling she gave us that we were together part of a large and important movement.

— The present and past members of the executive committees of the (Alliance of) Digital Humanities Organizations.

Bursary Winners

Digital Humanities 2013 Student Conference Bursaries

Two of the 2013 bursary awards have been enabled by a generous donation from Patrick Juola

Hamed M. Alhoori (Texas A&M University)
Adam Anderson (Harvard University) and David Bamman (Carnegie Mellon University)
Drayton Callen Benner (University of Chicago)
Alberto Campagnolo (University of the Arts, London)
Alexandra Chassanoff (University of North Carolina at Chapel Hill)
Constance Crompton (University of British Columbia-Okanagan)
Courtney Evans and Ben Jasnow (University of Virginia)
Paul Matthew Gooding (University College London)
Andrew Hankinson (McGill University)
Simon Rowberry (University of Winchester)
Graham Alexander Sack (Columbia University)
Ayush Shrestha (Georgia State University)
Dana Ryan Solomon (UC Santa Barbara)
Lindsay Thomas (University of California, Santa Barbara)

Consortium on Institutional Cooperation (CIC) Graduate Student Scholarships

The University of Nebraska-Lincoln received a generous donation to make competitive awards to the following students from CIC universities

Terry Brock (Michigan State University)
Mattie Burkert (University of Wisconsin)
Matt Burton (University of Michigan)
Trey Conatser (The Ohio State University)
Christopher Leeder (University of Michigan)
Grant Simpson (Indiana University)
Dawn Taylor (Penn State University)

Plenary Sessions

Pre-Conference Workshops and Tutorials

Panels

Papers

Posters

List of Reviewers

Alexander, Dr. Marc
Alsop, Peter Roger
Alvarado, Dr. Rafael
Anderson, Dr. Deborah
Anderson, Sheila
Andreev, Vadim Sergeevich
Andrews, Dr. Tara Lee
Antonijevic, Dr. Smiljana
Appleford, Simon James
Arneil, Stewart
Arthur, Dr. Paul
Ashton, Andrew Thomas
Audenaert, Dr. Michael Neal
Baayen, Prof. Rolf Harald
Baker, Drew
Bamman, David
Bański, Dr. Piotr
Barney, Brett
Bartsch, Dr. Sabine
Baudoin, Dr. Patsy
Bauer, Jean Ann
Bauman, Syd
Baumann, Ryan Frederick
Beavan, David
Bellamy, Dr. Craig
Bennis, Prof. Hans
Bentkowska-Kafel, Dr. Anna
Biber, Dr. Hanno
Bingenheimer, Dr. Marcus
Blanke, Dr. Tobias
Bodard, Dr. Gabriel
Bode, Dr. Katherine
Boggs, Jeremy
Bol, Prof. Peter Kees
Boot, Dr. Peter
Bordalejo, Dr. Barbara
Borgman, Prof. Christine L.
Borin, Prof. Lars
Boschetti, Dr. Federico
Bosse, Arno
Bouchard, Matthew
Bowen, Prof. William
Boyd, Dr. Jason Alexander
Bradley, John
Brisac, Anne-Laure
Brown, Prof. James
Brown, Prof. Susan
Brown, Travis Robert
Brughmans, Tom
Büchler, Marco
Buchmueller, Sandra
Burghart, Marjorie
Burnard, Lou
Burr, Prof. Elisabeth
Burrows, Dr. Toby Nicolas
Buzzetti, Prof. Dino
Byron, Dr. Mark Stephen
Canteaut, Dr. Olivier
Caton, Dr. Paul
Cayless, Dr. Hugh
Chamberlain, Dr. Daniel David
Cheesman, Dr. Tom
Chen, Dr. Shih-Pei
Chorney, Dr. Tatjana
Chue Hong, Neil
Ciula, Dr. Arianna
Clavaud, Florence
Clavert, Dr. Frédéric
Clement, Dr. Tanya
Clivaz, Prof. Claire
Cohen, Dr. Daniel J.
Connors, Louisa
Conway, Prof. Paul
Cooney, Dr. Charles M.
Cordell, Dr. Ryan
Cowan, William
Craig, Prof. Hugh
Crawford, Tim
Cream, Prof. Randall
Crompton, Dr. Constance
Croxall, Dr. Brian
Crymble, Adam H.
Cummings, Dr. James C.
Cunningham, Dr. Richard
Czmiel, Alexander
Dacos, Marin
Dahlstrom, Dr. Mats
Dalmau, Michelle
Dalziel, Karin
Davis, Dr. Rebecca Frost
Dawson, Dr. John
Day, Shawn
Deegan, Prof. Marilyn
Delve, Dr. Janet
Devlin, Dr. Kate
DiNunzio, Joseph
Dobrin, Dr. Lise M.
Dombrowski, Quinn Anya
Douglass, Prof. Jeremy
Downie, Prof. J. Stephen
Dubin, Dr. David S.
Dunn, Dr. Stuart
Dunning, Alastair
Earhart, Dr. Amy
Eberle-Sinatra, Dr. Michael
Eckart, Thomas
Eder, Dr. Maciej
Edmond, Dr. Jennifer C
Egan, Dr. Gabriel
Eide, Øyvind
Ell, Dr. Paul S.
Engel, Prof. Deena
Esteva, Dr. Maria
Everaert, Prof. Martin
Fendt, Dr. Kurt E.
Finn, Prof. Edward
Fiormonte, Dr. Domenico
Fischer, Dr. Franz
Fitzpatrick, Dr. Kathleen
Flanders, Dr. Julia
Forest, Dr. Dominic
Fraistat, Prof. Neil R.
France, Dr. Fenella Grace
French, Dr. Amanda
French, Dr. Scot
Funkhouser, Prof. Chris
Furner, Dr. Jonathan
Furuta, Dr. Richard
Gallet-Blanchard, Prof. Liliane
Gants, Prof. David
Garces, Dr. Juan
Garfinkel, Dr. Susan
Gärtner, Prof. Kurt
Gartner, Richard
Gibbs, Prof. Fred
Gibson, Dr. Matthew
Gil, Alexander
Gilbert, Joseph
Gillies, Sean
Gist, D. Chris
Goetz, Dr. Sharon K.
Gold, Dr. Matthew K.
Goldfield, Dr. Joel
Gouglas, Dr. Sean
Gow, Ann
Gradmann, Prof. Stefan
Graham, Wayne
Green, Prof. Harriett Elizabeth
Gregory, Dr. Ian
Groß, Dr. Nathalie
Gueguen, Gretchen Mary
Guertin, Dr. Carolyn
Guiliano, Dr. Jennifer Elizabeth
Hankins, Gabriel Anderson
Hanlon, Ann
Harbeson, Prof. Eric

Harris, Dr. Katherine D.
Hawkins, Kevin Scott
Heath, Sebastian
Hedges, Dr. Mark
Heiden, Dr. Serge
Heuvel, Dr. Charles van den
Heyer, Prof. Gerhard
Hill, Dr. Timothy
Hirsch, Dr. Brett
Hoffmann, Daniel
Holmes, Martin
Hoover, Prof. David L.
Horton, Prof. Tom
Hu, Dr. Xiao
Hughes, Prof. Lorna
Hui, Dr. Barbara
Huitfeldt, Claus
Hunyadi, Prof. László
Inaba, Prof. Mitsuyuki
Isaksen, Dr. Leif
Ivanovs, Prof. Aleksandrs
Jannidis, Fotis
Jewell, Dr. Andrew Wade
Johnsen, Dr. Lars
Johnson, Eric D. M.
Johnson, Dr. Ian R.
Johnston, Kelly Gene
Jones, Prof. Jason B.
Jones, Prof. Steven Edward
Juola, Prof. Patrick
Kaislaniemi, Samuli
Kansa, Dr. Eric Christopher
Keating, Dr. John Gerard
Kelber, Nathan Patrick
Kelleher, Prof. Margaret
Keramidas, Dr. Kimon
Kermanidis, Dr. Katia Lida
Khosmood, Foaad
Kibbee, Prof. Douglas
Kirschenbaum, Prof. Matthew
Knight, Gareth
Knox, Douglas
Koh, Prof. Adeline
Körner, Fabian
Koster, Dr. Elwin
Kowal, Kimberly
kraus, Dr. Kari Michaele
Krauwer, Steven
Kretzschmar, Dr. William
Krot, Michael Adam
Kuemmel, Dr. Christoph
Lana, Dr. Maurizio
Lancaster, Dr. Lewis Rosser
Lang, Dr. Anouk
Lavagnino, Dr. John
Lavrentiev, Dr. Alexei
Lawless, Prof. Séamus
Lawrence, Dr. Katharine Faith
Ledezma, Prof. Domingo
Lendvai, Dr. Piroska
Leon, Dr. Sharon
Lester, Dave
Lindemann, Prof. Marilee
Litta Modignani Picozzi, Dr. Eleonora
Llewellyn, Clare
Lombardini, Dr. Dianella
Lorang, Dr. Elizabeth M.
Losh, Elizabeth
Lucic, Ana
Lüngen, Dr. Harald
Luyckx, Kim
Lyman, Dr. Eugene W.
Maeda, Prof. Akira
Mahony, Simon
Makinen, Dr. Martti
Maly, Dr. Kurt
Mandell, Prof. Laura C.
Manovich, Prof. Lev
Marino, Prof. Mark Christopher
Martin, Prof. Worthy N.
Martín Arista, Dr. Javier
McCarty, Prof. Willard
McDaniel, Dr. Rudy
McDonald, Prof. Jarom Lyle
McPherson, Tara
Meece, Stephanie
Meeks, Elijah
Meister, Prof. Jan Christoph
Meschini, Federico
Miles, Adrian
Mimno, Dr. David
Miyake, Dr. Maki
Monteiro Vieira, Jose Miguel
Morrison, Prof. Aimée
Mostern, Dr. Ruth
Moulin, Prof. Claudine
Moulthrop, Prof. Stuart
Mueller, Prof. Martin
Muller, Prof. A. Charles
Muñoz, Trevor
Munson, Matthew Aaron
Muralidharan, Aditi
Murphy, Dr. Orla
Mylonas, Elli
Nagasaki, Kiyonori
Nelson, Prof. Brent
Nelson, Dr. Robert K.
Nerbonne, Prof. John
Newton, Greg T.
Nieves, Dr. Angel David
Norrish, Jamie
Nowviskie, Dr. Bethany
Nyhan, Dr. Julianne
O'Donnell, Prof. Daniel Paul
Ohya, Prof. Kazushi
Olsen, Prof. Mark
Opas-Hänninen, Prof. Lisa Lena
Ore, Dr. Christian-Emil
Ore, Espen S.
Organisciak, Peter
Paolillo, Prof. John
Pasanek, Brad
Perdue, Susan Holbrook
Perez Isasi, Dr. Santiago
Pierazzo, Dr. Elena
Piez, Dr. Wendell
Pitti, Daniel
Porter, Dorothy Carr
Posner, Dr. Miriam
Potvin, Sarah
Prescott, Prof. Andrew John
Priani, Dr. Ernesto
Puschmann, Dr. Cornelius
Pytlik Zillig, Prof. Brian L.
Rahtz, Sebastian
Ramsay, Dr. Stephen
Rapp, Prof. Andrea
Redwine, Gabriela Gray
Rehbein, Prof. Malte
Rehberger, Prof. Dean
Rehm, Dr. Georg
Renear, Prof. Allen H.
Reside, Dr. Doug
Rhody, Dr. Jason
Riddell, Allen Beye
Ridge, Mia
Ridolfo, Dr. Jim
Robey, Prof. David
Robinson, Prof. Peter
Rochester, Dr. Eric
Rockwell, Prof. Geoffrey
Rodríguez, Dr. Nuria
Roe, Dr. Glenn H.
Roeder, Torsten
Rohrbach, Prof. Augusta
Romanello, Matteo
Romary, Laurent
Roorda, Dr. Dirk
Rosner, Prof. Lisa
Ross, Claire Stephanie
Roueché, Prof. Charlotte
Roued-Cunliffe, Henriette

Rudman, Prof. Joseph
Ruecker, Dr. Stan
Ruotolo, Christine
Russo, Dr. Angelina
Rybicki, Dr. Jan
Sahle, Patrick
Saklofske, Dr. Jon
Salciute-Civiliene, Gabriele
Sample, Prof. Mark
Sanderson, Dr. Robert
Sayers, Dr. Jentery
Schaßan, Torsten
Scheinfeldt, Dr. Joseph Thomas
Schlitz, Dr. Stephanie
Schmidt, Dr. Desmond
Schmidt, Harry
Schmidt, Sara A.
Schöch, Dr. Christof
Schreibman, Prof. Susan
Schubert, Prof. Charlotte
Scifleet, Dr. Paul Anthony
Seppänen, Prof. Tapio
Sewell, David Robert
Shaw, Dr. Ryan Benjamin
Shaw, William Stewart
Shep, Dr. Sydney
Shimoda, Prof. Masahiro
Short, Prof. Harold
Siemens, Dr. Lynne
Siemens, Prof. Raymond George
Simons, Prof. Gary F.
Sinclair, Prof. Stéfan
Singer, Prof. Kate
Smith, Prof. Martha Nell
Smithies, Dr. James Dakin
Snyder, Dr. Lisa M.
Sokól, Dr. Malgorzata
Spence, Paul Joseph
Sperberg-McQueen, Dr. Michael
Spiro, Dr. Lisa
Steggle, Prof. Matthew
Sternfeld, Dr. Joshua
Stokes, Dr. Peter Anthony
Sukovic, Dr. Suzana
Sula, Dr. Chris Alen
Suzuki, Dr. Takafumi
Swanstrom, Dr. Elizabeth Anne
Tabata, Prof. Tomoji
Tanner, Simon
Tasovac, Toma
Teehan, Aja
Teich, Prof. Elke
Terras, Dr. Melissa
Thaller, Prof. Manfred
Thawonmas, Prof. Ruck
Theibault, Dr. John Christopher
Todirascu, Amalia
Tomasek, Dr. Kathryn
Tomić, Marijana
Tonra, Dr. Justin Emmet
Travis, Dr. Charles Bartlett
Trettien, Whitney
Trippel, Dr. Thorsten
Tupman, Dr. Charlotte
Underwood, Prof. Ted
Unsworth, John
Valverde Mateos, Dr. Ana
van Dalen-Oskam, Prof. Karina
Van den Branden, Ron
van den Herik, Prof. H. J.
Van Elsacker, Bert
van Erp, Dr. Marieke
van Hooland, Prof. Seth
Van Zundert, Joris Job
Varner, Dr. Robert Stewart
Venecek, John T.
Verdu Ruiz, Silvia
Verhoeven, Prof. Deb
Vertan, Dr. Cristina
Vetch, Paul
Viglianti, Raffaele
Visconti, Amanda
Walsh, Prof. John
Walter, Katherine L.
Warwick, Prof. Claire
Watrall, Dr. Ethan
Weidman, Robert William
Weingart, Scott
Welger-Barboza, Prof. Corinne
Wernimont, Prof. Jacqueline D.
Wheeles, Dana
Wiesner, Dr. Susan L.
Wilkens, Prof. Matthew
Willett, Perry
Williams, Dr. George H.
Winder, Dr. William
Witt, Dr. Andreas
Wittern, Prof. Christian
Wolff, Dr. Mark
Worthey, Glen
Wulfman, Dr. Clifford Edward
Wynne, Martin
Zafrin, Dr. Vika
Zeldenrust, Douwe
Zimmerman, Matthew
Zöllner-Weber, Dr. Amélie

Plenary Sessions

Opening Keynote

Ferriero, David S.
Archivist of the United States

Harnessing the Wisdom of the Crowd: The Citizen Archivist Program at the National Archives. (July 16, 2013 5:30 PM)

In articulating his commitment to transparency, collaboration, and participation in his administration, President Obama said to his senior staff, on his first day in office:

"Our commitment to openness means more than simply informing the American people about how decisions are made. It means recognizing that Government does not have all the answers, and that public officials need to draw on what citizens know. And that's why, as of today, I'm directing members of my administration to find new ways of tapping the knowledge and experience of ordinary Americans—scientist and civic leaders, educators and entrepreneurs…"

One member of that administration, David S. Ferriero, Archivist of the United States, describes how his agency had embraced the President's message and engaged the American public in the work of the National Archives.

Biography

The Honorable David S. Ferriero was sworn in as 10th Archivist of the United States on November 13, 2009.

The Archivist of the United States, appointed by the President of the United States, is the head of the National Archives and Records Administration, an agency of the Executive Branch of the Government. The agency is responsible for providing guidance to the White House and the Executive Branch agencies and departments on the creation and maintenance of their records. It oversees the transfer to the National Archives of the permanently valuable records of the federal government and makes them available for study. Those records include the Oaths of Allegiance signed by George Washington and his troops at Valley Forge, the Declaration of Independence, the Constitution, and the Bill of Rights.

This collection translates into about 12 billion sheets of paper, 40 million photographs, miles and miles of video and film, and more than 5.3 billion electronic records. The records are housed in facilities around the country, from Anchorage, Alaska to Atlanta, Georgia — including 2 Washington, DC, area buildings, 14 Regional Archives, 17 Federal Records Centers, 13 Presidential Libraries, and the National Personnel Records Center.

Previously, Mr. Ferriero served as the Andrew W. Mellon Director of the New York Public Libraries (NYPL). In this position he was part of the leadership team responsible for integrating the 4 research libraries and 87 branch libraries into one seamless service for users; and was in charge of collection strategy; conservation; digital experience and strategy; reference and research services; and education, programming, and exhibitions.

Before joining the NYPL in 2004, Mr. Ferriero served in top positions at two of the nation's major academic libraries, the Massachusetts Institute of Technology in Cambridge, MA, and Duke University in Durham, NC.

Mr. Ferriero earned bachelor's and master's degrees in English literature from Northeastern University in Boston and a master's degree from the Simmons College of Library and Information Science, also in Boston. He served as a hospital corpsman in the Navy during the Vietnam War.

Busa Award Lecture

McCarty, Willard
Professor of Humanities Computing and Director of the Doctoral Programme in the Department of Digital Humanities at King's College London; Professor in the Digital Humanities Research Group, University of Western Sydney; and Fellow of the Royal Anthropological Institute (London)

Getting there from here: Remembering the future of digital humanities. (July 18, 2013 3:30 PM)

In this talk I look back on a life of learning in digital humanities and on the past of literary computing. Personal retrospection yields a moral; another arises from my highly referential style of writing. But my overall aim is to pick out from the incunabular period (1949-1991) clues pointing to the trajectory of a discipline that has survived decades of neglect to become popular but is not yet able to articulate an agenda of as well as in the humanities. The Web and what we have made from it were sufficient for popularity. Many have been well served, are grateful and want to do more; others see opportunity to extend the reach of their native research. But beyond providing material for that which happens elsewhere by other means by other people we remain largely as we were before the Web. In this lecture I argue that the key to a scholarly life in digital humanities worth living has been visible all along, by taking seriously what happens where we stand: at the traumatic cross-roads of the humanities and computing.

Biography

Willard McCarty is Professor of Humanities Computing and Director of the Doctoral Programme in the Department of Digital Humanities at King's College London; Professor in the Digital Humanities Research Group, University of Western Sydney; and Fellow of the Royal Anthropological Institute (London). He is Editor of the British journal, *Interdisciplinary Science Reviews* (2008-) and founding Editor of the online seminar *Humanist* (1987-). He is recipient of the Canadian Award for Outstanding Achievement, Computing in the Arts and Humanities (2005), and the Richard W. Lyman Award, National Humanities Center (2006). He is currently at work on *Machines of Demanding Grace*, a book concerned with the interrelation of the humanities and computing. He lectures occasionally in Europe, North America, and Australia. See www.mccarty.org.uk.

Closing Keynote

Galina, Isabel
Researcher at the Instituto de Investigaciones Bibliográficas at the National University of Mexico (UNAM)

Is there anybody out there? Building a global DH community. (July 19, 2013 4:00 PM)

Digital Humanities has come a long way towards establishing itself as a dynamic and innovative field of study. It has been pointed out however that an important area for further development is a broader internationalization of the Digital Humanities community. Currently most of the available literature is centered on DH projects in a handful of mainly English speaking countries. An important challenge for the DH community is to extend its international reach and integrate work from a wider range of languages and academic institutions. Presently there are numerous initiatives under way that aim to broaden Digital Humanities regional and linguistic diversity. In this talk I will discuss some of the issues with identifying DH scholars and establishing networks of collaboration drawing on specific examples of setting up the Red de Humanidades Digitales (RedHD).

Biography

Isabel Galina is currently a researcher at the Instituto de Investigaciones Bibliográficas at the National University of Mexico (UNAM). With a background in English Literature and Electronic Publishing, her Ph.D. research at University College London (UCL) was on the impact of electronic resources on scholarly communication and publishing. This led to a particular interest in new modes of scholarship and digital projects within the Humanities.

At the UNAM she has been involved in numerous initiatives related to institutional repositories, digitization projects, electronic publishing and the use and visibility of digital resources. She is a founding member and current president of the Red de Humanidades Digitales (RedHD) which aims to promote and strengthen Digital Humanities with special emphasis on research and teaching in Spanish as well as the Latin American region in general. She is Associate Editor of *LLC: The Journal of Digital Scholarship in the Humanities* and Honorary Research Fellow at the UCL Department of Information Studies.

Workshops

Looking for needles in DH haystacks: efficient querying of complex data

Banski, Piotr
banski@ids-mannheim.de
Assistant Professor of linguistics at the Institute of English Studies of the University of Warsaw

Diewald, Nils
niewald@ids-mannheim.de
Researcher at the Institut für Deutsche Sprache (IDS) in Mannheim

Witt, Andreas
witt@ids-mannheim.de
Head of the Research Infrastructure Group at the IDS

The rapid development of the discipline (or, more precisely, disciplines) known as Digital Humanities resulted in the ever wider accessibility of digitization methods and, consequently, the steadily growing amount of digitized and interlinked data. However, as in many other disciplines that have followed a similar pattern of development, it turns out that, while the amount of information is growing, the methods for quick, easy and successful retrieval of that information are either not yet established, or not yet sufficiently widespread.

In view of the massive amount of available data, an average DH scholar is confronted with the task of finding a needle in a haystack: while, seemingly, everything is there: structured, interlinked and ready to be used, and while wellknown query mechanisms exist and have been used for years in other disciplines, the fundamental questions still concern the best way to formulate the particular research questions, the best method appropriate to the task at hand, or a friendly tool that would provide the relevant results in the desired format and without too steep a learning curve.

The tutorial is going to present stateoftheart methods in querying data, from textual to multimodal, with a focus on use cases commonly found in Digital Humanities, or envisioned for the near future of this expanding field. It will be taught by two specialists in markup languages and corpus linguistics, currently involved in the process of creating a new analysis platform designed to handle large amounts of linguistic data. This is not meant to be a tutorial just for linguists, however: we intend to provide an opportunity to carry over some wellknown methods and techniques from linguistic research, where they have been used for years, onto the broader area of Digital Humanities, where queries target not only texts, but also nontextual objects, such as binary streams, ontologies, prosopographic databases or GIS data (these latter types of objects will be discussed to the extent to which they can be linked from textual resources).

We shall focus primarily on the search in metadata, nonannotated data, and structured annotated data (especially TEIencoded).

Part of the way to ensure closer cooperation among DH researchers may be to provide them with a common language in which they can specify questions asked of a variety of datasets in a variety of structures. The tutorial shall investigate to what extent the creation of such a *Corpus Query Lingua Franca* is a realistic endeavour, what basic elements such a language would have to possess, what kind of objects it would have to query, and what set of constraints it will have to unavoidably obey. This is also one of the current foci of the Special Committee of ISO that addresses language resource management (TC37 SC4), in which both presenters actively participate.

Description of target audience and expected number of participants

We expect up to ca. 35 people. We intend the tutorial for general DH audience: variety is a virtue in this case, because we want to address actual use cases, some of which will surely come from the participants themselves.

Background information

Piotr Bański is an Assistant Professor of linguistics at the Institute of English Studies of the University of Warsaw, and a researcher at the Institut für Deutsche Sprache in Mannheim, where he is the Project Manager of the “Corpus Analysis Platform of the Next Generation” (KorAP), a project financed by the Leibniz Association (LeibnizGemeinschaft). He served as an elected member of the TEI Technical Council for term 20112012 and since 2010 has been involved in the work of the ISO TC37 SC4 committee for Language Resource Management. His latest project within the scope of ISO is work on *Corpus Query Lingua Franca*, within TC37 SC4 Working Group 6, convened by Andreas Witt. His current interests focus mostly on text encoding as well as the creation and use of robust language resources.

After graduating from Bielefeld University in 1996, Andreas Witt started at this university as a researcher and instructor in Text Technology. He was heavily involved in

the establishment of the Magister and BA programmes in Text Technology at Bielefeld Universität in 1999 and 2002 respectively. After completing his Ph.D. in 2002 he became an assistant lecturer with the Text Technology group in Bielefeld. In 2006 he moved to Tübingen University, where he was involved in a project on “Sustainability of Linguistic Resources” and in projects on the interoperability of language data. Since 2009 he has been a senior researcher at the Institut für Deutsche Sprache (Institute for the German Language) in Mannheim. Andreas is a member of numerous research organizations including the TEI Special Interest Group "TEI for Linguists". His major research interests deal with questions of the use and limitations of markup languages for the linguistic description of language data.

Outline

Main issues addressed by the tutorial:

- What should a text query system for DH in the 21st century look like?
- What kinds of queries should a query system be able to deal with?
- How to define a modern query language?
- How should a text corpus be structured in the future?

List of topics (some of them may receive only cursory attention; much depends on the composition of the audience and the demand)

- Digital text
- Annotation of text
- Annotation formats (HTML, TEI, others)
- Text corpora
- Corpora of written languages
- Corpora of spoken language
- Aligned corpora
- Trees, Graphs, feature structures
- Web as a Corpus
- Characters and character encoding
- Metadata
- Simple search
- Search with regular expressions
- Search in XMLData (Xquery, XPATH)
- Complex Annotations
- Multilevel annotations
- Relations between annotations
- Existing corpus query systems

From 2D to 3D: An Introduction to Additive Manufacturing and Desktop Fabrication

Boggs, Jeremy
jkb2b@virginia.edu
Design Architect for Digital Research and Scholarship in the Scholars’ Lab, at the University of Virginia Library

Elliott, Devon
devonelliott@gmail.com
PhD candidate in history at Western University

Sayers, Jentery
jentery@uvic.ca
Assistant Professor of English at the University of Victoria

Desktop fabrication is the digitization of analog manufacturing techniques. Comparable to desktop publishing, it affords the output of digital content (e.g., 3D models) in physical form (e.g., plastic). It also personalizes production through accessible software and hardware, with more flexibility and rapidity than its analog predecessors.

Additive manufacturing is a process whereby a 3D form is constructed by building successive layers of a melted source material (at the moment, this is most often some type of plastic). There is little waste, as material is not removed from a block to create the shape, as in traditional machine milling, for example. The method also affords the materialization of forms not possible to construct from traditional subtractive processes of manufacturing. The technology driving additive manufacturing in the desktop fabrication field is the 3D printer, tabletop devices that materialize digital 3D models.

In this workshop, we will introduce technologies used for desktop fabrication and additive manufacturing, and offer a possible workflow that bridges the digital and physical worlds for work with threedimensional forms. We will begin by introducing 3D printers, and demonstrate how they operate by printing things throughout the event. The software used in controlling the printer and in preparing models to print will be explained. We will use free software sources so those in attendance can experiment with the tools as they are introduced.

The main elements of the workshop are

- Acquisition of digital 3D models — from online repositories to creating your own with photogrammetry, scanning technologies, and modelling software
- Software to clean and reshape digital models in order to make them print-ready and remove artifacts from the scanning process.
- 3D printers and the software to control and use them

Those attending are asked to bring, if possible, a laptop computer to install and run the software introduced, and a digital camera or smartphone for experimenting with photogrammetry. Workshop facilitators will bring cameras, a 3D printer, plastics, and related materials for the event. By the end of the conference, each participant will have the opportunity to print an object for their own use.Those attending are asked to bring, if possible, a laptop computer to install and run the software introduced, and a digital camera or smartphone for experimenting with photogrammetry. Workshop facilitators will bring cameras, a 3D printer, plastics, and related materials for the event. By the end of the conference, each participant will have the opportunity to print an object for their own use.

Instructors

Jeremy Boggs, University of Virginia Library

Jeremy Boggs is the Design Architect for Digital Research and Scholarship in the Scholars' Lab, at the University of Virginia Library. His dissertation, entitled "The Designing Historian," explores design as a methodology for doing digital history. Other research interests include the history of design, the history of technology, and social/cultural history. He has conducted workshops and introduced 3D printing at the Scholar's Lab.

Devon Elliott, Western University

Devon Elliott is a PhD candidate in history at Western University. In addition to his application of digital fabrication methods to his dissertation project on the history of stage magic, he has conducted workshops on 3D technologies at universities, conferences, galleries, and hackerspaces. He is also an instructor on digital fabrication and physical computing at the Digital Humanities Summer Institute.

Jentery Sayers, UVic

Jentery Sayers is an Assistant Professor of English at the University of Victoria, with research interests in comparative media studies, digital humanities, AngloAmerican modernism, computers and composition, and teaching with technologies. He is the director of the Maker Lab at UVic, and teaches digital fabrication and physical computing at the Digital Humanities Summer Institute.

Audience

Targeted towards scholars interested in learning about technologies surrounding 3D printing and additive manufacturing, and for accessible solutions to implementing those technologies in their work. Past workshops have been for faculty, graduate and undergraduate students in the humanities; librarians; archivists; GLAM professionals; digital humanities centers. This is an introductory workshop, so little prior experience is necessary, only a desire to learn and be engaged with the topic.

Length and format of the workshop

The one day workshop will have three major components that all participants will engage with. These core components are:

1) Where to get digital 3D models. This can be from online repositories that are freely available (and also serve as potential archival solutions or distributive channels) to creating one's own models with photogrammetry or scanning technologies, or drawing virtual models in design software.
2) Software solutions that mediate between the virtual forms and the corresponding physical technologies are necessary to clean and postprocess models to make them suitable for printing. There isn't one tool to do all this, so this section will introduce those various software platforms, and explain both how the virtual model needs to be transformed and the specific elements of those software packages that enable one to make those changes.
3) 3D printing What is it? How does it work? What machines are available? Pros and cons of using the devices. Software to interface with the machines. Materials that they can print with. Limits of the printable forms.

Using Open Annotation

Cole, Timothy W.
Mathematics and Digital Content Access Librarian and Professor of Library and Information Science at the University of Illinois

Gerber, Anna
Technical Project Manager with the ITEE eResearch group at The University of Queensland, Australia

Sanderson, Robert
Co-chair of the W3C Open Annotation Community Group and co-editor of the specification

Smith, James
Software Architect at the Maryland Institute for Technology in the Humanities (University of Maryland)

Brief Description

Annotation is a long-established scholarly primitive[1] supporting digital humanities scholarly workflows and practices. As the humanities scholars use of retrospectively and born-digital materials grows so too does the need for robust, standards-based annotation tools and services that can span content repositories and Web application boundaries.

Over the course of this half-day workshop we will examine the current and prospective role of annotation in digital humanities scholarship and investigate the potential utility of the Open Annotation data model specification[2] recently released by the W3C Open Annotation Community Group.[3] Participants will consider whether this specification can help encourage the extension of existing tools and the development of new more robust, interoperable Web-based annotation tools and services in ways that can better meet the needs of the digital humanities scholarly community. Prior to the Workshop, each participant will be asked to submit a brief (1 to 2 page) summary giving their initial assessment of the Open Annotation data model in the context of a specific annotation application, research requirement or use case drawn from his or her own scholarship. All summary assessments submitted will be posted on a Workshop Website (to be maintained by the Open Annotation Collaboration)[4] and a subset of these assessments will be presented by participants and discussed during the Workshop. In addition the Workshop Leaders will present results from several of the concrete annotation demonstration experiments conducted over the last 18 months by the Open Annotation Collaboration. The current status of Web-based digital annotation tools, services, practices and communities will be reviewed with the goal of illuminating critical facets of infrastructure beyond the scope of the Open Annotation data model or areas of the data model which require further refinement. Outcomes from the Workshop will help identify potential priorities and future directions for the W3C Open Annotation Community Group. Participants will gain a better understanding of the Open Annotation data model, its implementation, and its potential as a resource supportive of their future work.

If accepted, budget for this proposed Workshop will be underwritten by the Open Annotation Collaboration, a project based at the University of Illinois at Urbana-Champaign and funded by a generous grant from the Andrew W. Mellon Foundation. OAC was one of the initiatives, along with the Annotation Ontology initiative,[5] that founded the W3C Open Annotation Community Group. This workshop represents an opportunity to explore the annotation needs of digital humanists and ensure that these needs are factored into the future plans of the W3C Open Annotation Community Group.

Additional Contributors

- Paolo Ciccarese, Biomedical Informatics Research & Development, MIND Informatics, Instructor at the Harvard Medical School and Assistant in Neurology at the Massachusetts General Hospital.
- Jane Hunter, Professorial Research Fellow & Leader of the eResearch Lab, School of ITEE, The University of Queensland.
- Jacob Jett, Visiting Project Coordinator, Center for Informatics in Science and Scholarship, GSLIS, University of Illinois at Urbana-Champaign.
- Herbert Van de Sompel, Information Scientist at the Los Alamos National Laboratory and leader of the LANL Research Library Digital Library Research & Prototyping Team

Target Audience

We anticipate an audience of about 25 digital humanities tool and Web service developers and technology managers responsible for digital library services, scholarly discourse services, note-taking software and similar Web-based applications. Registration is open (i.e., non-competitive) until full; however, registrants will be advised that we will ask each of them to submit a position paper (see below).

Half-day Workshop Agenda Outline

1. Open Annotation Introduction & Overview (30 minutes)
2. Example implementations / demonstrations from OAC Experiments (60 minutes)
3. Q & A Session (30 minutes)
4. [break] (15 minutes)
5. Participant presentations (90 minutes)
6. Discussion (30 minutes)

Writing your First Digital Humanities Grant

Guiliano, Jennifer
guiliano@umd.edu
Assistant Director of the Maryland Institute for Technology in the Humanities (MITH)

Appleford, Simon
simonja@clemson.edu
Associate Director for Humanities, Arts, and Social Sciences at the Clemson CyberInstitute, Adjunct Lecturer in History at Clemson University

Abstract

Designed for humanities scholars seeking assistance with their first grant, this workshop introduces participants to best practices in writing and submitting a grant. Participants will be provided with a series of online resources, including presentations, exemplar successful grants, and podcasts to help them complete a first draft of a proposal before they arrive in Lincoln. Those drafts will be circulated to other participants prior to the workshop and will serve as the core basis of our workshop discussions with the anticipation being that each participant will receive clear feedback from other attendees that will aid them in the revision of their proposal. Drafts will be encouraged to emulate the popular National Endowment for the Humanities Digital Humanities Startup Grant competition in order to provide the most flexibility for participants in their digital humanities endeavors. This seminar will be limited to 15 participants; additional seminars may be made available should demand necessitate.

Target Audience

The target audience for this workshop are primarily early career digital humanists, including graduate students and junior scholars, who will be submitting their first grant in the coming year. We choose to limit the workshop to 15 participants as the substantive nature of the discussion necessitates a smallgroup atmosphere. A session similar to this was successfully held at the Digital Humanities Winter Institute as part of the instructors' project development class. That class had 20 participants, the majority of whom were digital humanists undertaking their first project. Additionally, Jennifer coled a similar session with Dr. Lisa Rhody at the Modern Studies Association that included neophyte digital humanists. That session benefited from its small size. Both sessions at DHWI and MSA were well reviewed and would be appropriate for firsttime grant writers in the humanities.

Workshop Logistics

Applications: Participants will be admitted on a firstcome, firstserved basis. Registration for this workshop will close on May 1st (or when the workshop reaches capacity). Participants will be notified of their admission on May 1st. Papers are due to the full group no later than June 15th. We set this deadline a bit early to allow participants time to not only read all papers but to complete any supporting reading they might elect to familiarize themselves with.

Papers should be 5 to 7 pages in length, doublespaced. We encourage participants to produce shorter papers to allow for greater commenting/consideration. Additionally, we want to encourage these papers to emulate the narrative section of the National Endowment for Humanities Digital Humanities Start Up Grants solicitation, as this competition focuses on humanities significance and innovation and is a likely funding source for early development projects.

We ask that every paper include the major elements of a project proposal, namely that the narrative should not assume specialized knowledge, and it should be free of jargon. It should clearly define technical terms so that they are comprehensible to a nonspecialist audience.The narrative should address the longterm goals for the project as well as the activities that the project would support.

Within the narrative, you should:

1.) Provide a clear and concise explanation — comprehensible to a general audience — of the project activities and the ultimate project results, noting their value to scholars, students, and general audiences in the humanities. Describe the scope of the project activities, the major issues to be addressed, and their *significance*

to the humanities. Show how the project will meet its objectives in innovative ways.
2.) Provide a rationale for the compatibility of your *methodology* with the intellectual goals of the project and the expectations of those who would make use of the project products.
3.) Provide a clear and concise summary of an *environmental scan* of the relevant field. The goal of an environmental scan is to take a careful look at similar work being done in your area of study. For example, if you are developing software to solve a particular humanities problem, please discuss similar software developed for other projects and explain how the proposed solution differs. If there are existing software products that could be adapted and reused for the proposed project, please identify them and discuss the pros and cons of taking that approach. If there are existing humanities projects that are similar in nature to your project, please describe them and discuss how they relate to the proposed project. The environmental scan should make it clear that you are aware of similar work being done and should explain how your proposed project contributes to and advances the field.
4.) Provide a *concise history* of the project, including information about preliminary research or planning, previous related work, previous financial support, publications produced, and resources or research facilities available.
5.) Complete a brief *workplan* that describes the specific tasks that will be accomplished during the grant period, identify the computer technology to be employed, and identify the staff members involved. Indicate what technical resources will be required.
6.) Identify potential *staff and collaborators*.
7.) Describe the plans to *disseminate* the project results through various media (printed articles or books, presentations at meetings, electronic media, or some combination of these).

Instructors

Jennifer Guiliano received a Bachelors of Arts in English and History from Miami University (2000), a Masters of Arts in History from Miami University (2002), and a Masters of Arts (2004) in American History from the University of Illinois before completing her Ph.D. in History at the University of Illinois (2010). She currently is an Assistant Director at the Maryland Institute for Technology in the Humanities at the University of Maryland and a Center Affiliate of the National Center for Supercomputing Applications. She has served as a PostDoctoral Research Assistant and Program Manager at the Institute for Computing in Humanities, Arts, and Social Sciences at the National Center for Supercomputing Applications (20082010) and as Associate Director of the Center for Digital Humanities (20102011) and Research Assistant Professor in the Department of History at the University of South Carolina. Jennifer currently serves on the Association for Computing in the Humanities (ACH) Executive Council (20132016), as codirector with Trevor Muñoz of the Digital Humanities Winter Institute (DHWI), and has served as an instructor in Project Development and Grant writing by invitation at the Digital Humanities Summer Institute, the Digital Humanities Winter Institute, as well as a number of individual universities. She has been primary author or primary investigator of over $3 million in externally funded grants.

Simon Appleford is Associate Director for Humanities, Arts, and Social Sciences at the Clemson CyberInstitute, and an Adjunct Lecturer in History at Clemson University. Simon received a Masters of Arts in Modern History and a Masters of Literature in Modern American History from the University of St. Andrews, Scotland, and is currently completing his PhD in History from the University of Illinois at UrbanaChampaign. Prior to joining Clemson University in 2011 he was Assistant Director at the University of Illinois' Institute for Computing in Humanities, Arts, and Social Science. His interests in digital technologies and American history have led to several publications including articles in *CTWatch Quarterly* and *Toward the Meeting of the Waters: Currents in the Civil Rights Movement in South Carolina* (University of South Carolina Press, 2007). He has served as primary author or primary investigator of over $3.5 million in externallyfunded research.

Introduction to the TXM content analysis platform

Heiden, Serge
slh@ens-lyon.fr
Project manager of the TXM platform development

The objective of the "introduction to TXM" tutorial is to introduce the participants to the methodology of textometric content analysis (http://textometrie.ens-lyon.fr/?lang=en) through working with the TXM software directly on their own laptop computers. At the end of the tutorial, the participants will be able to input their own textual corpora (Unicode encoded raw texts or XML tagged texts)

into TXM and to analyze them with the panel of content analysis tools available : word patterns frequency lists, kwic concordances and text browsing, rich full text search engine syntax (allowing to express various sequences of word forms, part of speech and lemma combinations constrained by XML structures), statistically specific sub-corpus vocabulary analysis, statistical collocation analysis, etc.).

During the tutorial, each participant will install TXM (from http://sourceforge.net/projects/txm) and the TreeTagger lemmatizer (http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger) on her Windows, Mac or Linux laptop and will leave the tutorial with a ready to use environment.

The tutorial will also introduce the participants to the TXM community ecosystem (users mailing list and wiki, bug reports, etc.) and to the TXM portal version server software (See for example http://portal.textometrie.org/demo) for on line corpus distribution and analysis. Time permitting, TEI encoding aspects of corpora related to TXM could also be introduced, as well as speech transcriptions or parallel corpora encoding and analysis.

The objective of the “introduction to TXM” tutorial is to introduce the participants to the methodology of textometric content analysis (http://textometrie.ens-lyon.fr/?lang=en) through working with the TXM software directly on their own laptop computers. At the end of the tutorial, the participants will be able to input their own textual corpora (Unicode encoded raw texts or XML tagged texts) into TXM and to analyze them with the panel of content analysis tools available : word patterns frequency lists, kwic concordances and text browsing, rich full text search engine syntax (allowing to express various sequences of word forms, part of speech and lemma combinations constrained by XML structures), statistically specific sub-corpus vocabulary analysis, statistical collocation analysis, etc.)

The tutorial will be taught in English for the first time in DH2013 (the TXM User Graphical Interface is already available in English), and will complement two accepted communications introducing the TXM platform given during the conference:

—“TXM Platform for analysis of TEI encoded textual sources” #391 long paper;
— “TXM Portal: Providing Online Access to Textometric Corpus Analysis” #399 poster with live demo.

Tutorial Instructor

Serge Heiden

Project manager of the TXM platform development (http://textometrie.ens-lyon.fr/spip.php?article9). S. Heiden develops the textometry content analysis methodology through the development of tools able to process richly encoded corpora. Working on the relation between analysis tools and XML-TEI encoded corpora, he is involved in the TEI consortium activities as the TEI Tools SIG convener (http://www.tei-c.org/Activities/SIG/Tools).

Target audience and expected number of participants

The ideal number of participants is about 12-15 people, the maximum number of participants is about 20.

Each participant should come with her own laptop computer. The tutorial needs to run at least for a full day(*): typically half day for TXM tools fundamentals and half day for main corpus formats fundamentals (TXT and XML) and input procedures into the platform.

(*) The regular TXM tutorials run for two days (one day TXM introduction, one day corpus formating and import into TXM).

Brief Outline

9am – 12pm
1pm – 5pm

— Install & introduction: 45'
— TXM user interface & windows, corpus Description command

— Main tools: 2h15
—Lexicon analysis & spreadsheet export
—Index building for distributional semantics & Corpus Query Language syntax
— Concordance & Reading, Progression graphics
—Partitions, Subcorpus & Specificity/Factorial analysis
—Coccurrence analysis
—TXM portal demo (optional)
—TXM community: mailing lists, web sites and documentation

— TXM import strategy and main corpus formats: TXT-Unicode+CSV, XML+CSV, XML-TEI: 1/2h
— TXT-Unicode sample corpus and TXT+CSV import into TXM, sample analysis: 1h15
— introduction to XML and to TXT2XML conversion tools: 1/2h

— XML sample corpus and XML/w+CSV import into TXM, sample analysis: 1h45

Fast-Tracking a research database using Heurist

Johnson, Ian
Director, Arts eResearch, University of Sydney

In this workshop participants will learn how to build a Heurist online database to support a small research project, and end up with a fully operational, web-accessible, private or shared database on a hosted server. From this they will be able to freely create additional databases and/or download the Open Source software for installation on an institutional or cloud-based server. No programming or technical skills are required to complete the workshop.

Heurist is a database abstraction which hides the complexity of database design behind a simple web interface, allowing researchers to rapidly define quite complex databases through a researcher-oriented interface without the need for programming. Heurist has been developed across a number of Australian Research Council research grants in the Humanities to support a wide range of research, from the production of large historical encyclopaedias with an editorial team, such as the Dictionary of Sydney (dictionaryofsydney.org), through archaeological survey and excavation databases to individual text markup projects. We use it internally for project management and task tracking as well as for image collection and PDF document management.

New databases can be set up on a web service within minutes, using templates developed by other Humanities projects. From the start these can include rich interlinking of records, spatial and temporal data, text markup and annotation. Heurist allows incremental development of the database structure as a project develops, without altering data which is already in the system. This allows fast-tracking of database setup because one can start with a few categories and work up to a more complex design, rather than having to do complete planning at the start which locks in a design and creates limitations or additional costs as projects evolve. Heurist is unique in allowing the import of existing record structures from any other Heurist database registered with a central index. We are actively building a number of 'community servers' which provide structural templates for different disciplines. These templates promote good database design and common standards without being prescriptive.

Heurist is being incorporated into two national research infrastructure projects. It will provide database-on-demand services on the NeCTAR Research Cloud for HuNI (Humanities Networked Infrastructure), allowing user-generated databases to expose their content to the HuNI Virtual Laboratory search, web linking and annotation services and to Research Data Australia, the national index of research databases. For FAIMS (Federated Archaeological Information Management System) it will provide a schema repository and data ingest from field data collection devices (Android tablets, GPS, cameras etc.), data refinement, data analysis/visualisation, web publishing and repository output services.

The workshop will be relevant to a wide range of Humanities researchers, but particularly to scholars who deal with collections of richly interlinked heterogeneous entities. These may include historical individuals, organisations and events, inscriptions, manuscripts and archival records, theatrical performances, places, buildings, artefacts, images and bibliographic information. It will be particularly relevant to scholars who do not have good institutional backing for eResearch tools and database development.

By the end of the workshop participants will be confident to use the web interface to create new databases online, import templates, make changes to database structure, edit and import data, search for and save subsets of the data, map, export and transform data, and publish data feeds within a website.

Ian Johnson’s research interests focus on methodologies for managing and publishing research data and the application of GIS and mobile devices in archaeological field research. He has developed a number of eResearch tools including the Minark database for archaeologists (1980-1987), TimeMap web mapping application (1995–2003), The Electronic Cultural Atlas Initiative data clearinghouse (1998–2003), FieldHelper (2004–2007) and Heurist (2005–present).

Target audience

Scholars who deal with collections of richly interlinked heterogeneous entities, particularly those who have limited access to institutional eResearch support, but also Digital Humanists who need to provide eResearch support to other scholars. I can handle up to around 20 people.

URL

HeuristScholar.org/h3/index.html

Keywords to Keyframes: Video Analytics for Archival Research

Kuhn, Virginia
vkuhn@cinema.usc.edu
Associate director of the Institute for Multimedia Literacy in the School of Cinematic Arts at the University of Southern California

Simeone, Michael
mpsimeon@gmail.com
Institute for Computing in Humanities, Arts, and Social Science, University of Illinois at Urbana — Champaign

The **target audience** for this workshop consists of scholars with research interests related to the way that visual media impacts culture.

Description + Schedule

This workshop will serve scholars of any level of technical expertise who are interested in studying images as part of their work in the digital humanities using a hybrid method that combines machine analytics (keyframes) and crowd-sourced tagging (keywords). Facilitating a discussion and training session featuring up to twenty participants, we will demo the Large Scale Video Analytics (LSVA) workbench for moving and still image analysis and archiving. The LSVA is a web portal developed through a collaboration among the National Center for Supercomputing Applications (NCSA) and the Extreme Science and Engineering Discovery Environment (XSEDE), the IML (Institute for Multimedia Literacy), and ICHASS (the Institute for Computing in the Humanities, Arts and Social Science). The LSVA has customized the prominent Medici content management system, a multimedia database which has served scholars worldwide. The LSVA requires no software installation, though we do require online access. Further, we will provide access to IM2Learn, a free software package for image analysis developed by the NCSA. There will be no CFP associated with this workshop; rather, we would like participants to self select.

The LSVA deploys the application of various algorithms for image recognition and visualization into the workflow that allows real-time analysis of video, as well as crowd-sourced content labeling such that the system becomes more valuable the more it is used.

In addition, the LSVA team has created and customized visualization tools that enhance research in several ways: novel visualizations employ spatial and temporal simultaneity, revealing unique aspects of a single film sequence; comparative visualizations represent relationships among multiple films within an archive; and, finally, the integration of visualization imagery becomes an input tag and a front end process that feeds the *Medici* content management system and enhances word-based labels, helping to close the semantic gap that occurs when words are applied to images.

8:00-8:30 am: Introduction to basic concepts of computer vision and image retrieval
8:30-9:00 am: Overview of standard research methodologies and those the LSVA extends
9:00-9:15 am:BREAK
9:15-10:00 am: Demo of LSVA system and walkthrough of interpreting output and requerying
10:00-11:00 am: Hands-on with LSVA system: uploading, sorting, and analyzing moving and still images

Workshop Leader Bios

BIO: **Virginia Kuhn** is associate director of the Institute for Multimedia Literacy, an organized research unit in the School of Cinematic Arts at the University of Southern California. She was the 2009 recipient of the USC provost's Award for Teaching with Technology, she co-chairs the Scholarly Interest Group on Media Literacy and Pedagogical Outreach for the Society for Cinema and Media Studies, and she serves on the editorial boards of several journals of media and technology. She joined USC in 2005 after successfully defending one of the first born-digital dissertations in the US, challenging archiving and copyright conventions. Committed to helping shape open source tools for scholarship, she published the first article created in Scalar, which appeared in the *International Journal of Learning and Media* and titled "Filmic Texts and the Rise of the Fifth Estate. She also serves on the editorial boards of several journals of media and technology.

BIO: **Michael Simeone** is the Associate Director for Research and Interdisciplinary Studies at ICHASS at the University of Illinois at Urbana-Champaign. Aside from his work on projects that engage the computational study of video, image-analysis of Great Lakes area historical maps, and the significance of social network analytics for the humanities, his research focuses on the intersection of humanities research procedures and data science. He received his PhD in English from the University of Illinois at Urbana-Champaign.

Collating Texts with Juxta WS in Ruby

Laiacona, Nick
nick@performantsoftware.com
President of Performant Software Solutions LLC

Middell, Gregor
gregor@middell.net
Literary Scholar & Software Developer

Full Description

In this workshop, participants will learn how to develop a customized collation pipeline in Ruby using Juxta WS. Juxta WS is a web service that can collate variant texts and visualize the differences between them. It is free, open-source software comprising a pipeline of "micro-services" that may be valuable to any project that deals with digital texts, especially texts encoded in XML. Juxta WS can take in a variety of source file types including plain text, HTML, and XML, with special handling for texts encoded in TEI. Results can be output at any point along the Juxta WS pipeline, so it can be used, for example, to convert XML to plain text or HTML and output an XSLT style sheet; isolate text from an HTML file; tokenize a text stream; annotate the differences between texts; align fragments of text between sources; visualize the collation of texts encoded in TEI parallel segmentation; or generate a single TEI parallel segmented output from multiple input files

Juxta WS can serve any project where the differences between variant texts are of interest. The Carolingian Canon Law Project at the University of Kentucky was an early adopter of Juxta WS, establishing a private collation workspace where registered users can collate variant texts from a Latin corpus. The Melville Electronic Library at Hofstra University is using Juxta WS to compare variants of Melville's writings, including the new manuscript transcriptions being created with the TextLab fluid-text editing tool. At Texas A&M University, the Early Modern OCR Project is collating the output of OCR engines with hand-typed versions of the same texts, using the change index calculated by Juxta WS as a metric of OCR performance. Juxta WS also powers Juxta Commons, a site sponsored by NINES where anyone may create a free account, upload source files for collation, and share the resulting visualizations. We think Juxta could be used creatively by diverse projects and we are looking forward to learning about projects participants bring to the workshop.

In the first half of the workshop, participants will follow along on their laptops as we introduce Juxta WS and the constituent parts of the collation pipeline. In the second half of the workshop, we will begin by showing participants how to obtain texts for experimentation from online sources. We will then have a hacking session where participants will work in small groups to leverage Juxta WS on texts from their own projects or obtained online.

During the hacking session, Gregor and Nick will float and answer questions. Participants will be able to access a Juxta WS server for use during this class and throughout DH 2013. Documentation for the Juxta WS API is available online and we will add information about the Ruby bindings there. There will also be a public Github repo where participants can push their work. We will help anyone not familiar with Git with a quick review of the necessary commands.

At the end of the workshop there will be a brief show and tell period. Each group will be invited to show what they worked on, either as working code or as the sketch of an idea.

Workshop Outline (3 hours, not including the break)

1. Introduction of Speakers and Topic (5 min.)
2. Demonstration of Juxta (10 min.)
3. Working with Juxta WS in Ruby (75 min.):
    a. Sources
    b. Tokenization
    c. Filtering XML Tags
    d. Collating
    e. Presenting Visualizations
4. Break
5. Finding Interesting Data (10 min.)
    a. online sources of humanities texts for collation
6. Break into small groups (5 min.)
7. Hacking (60 min.)
8. Show and Tell (15 min.)

Workshop Leaders

Nick Laiacona

Nick Laiacona is the President of Performant Software Solutions LLC. Under his leadership, Performant has cultivated a specialty in building custom software and websites for digital humanities projects. In recent years, Laiacona and the Performant team have worked on Juxta, a program for visualizing textual collations; TypeWright,

a tool for crowd-sourcing the correction of "dirty OCR" in databases of early modern books; and TextLab, an NEH-funded web application for fluid text editing. Performant Software also provides ongoing development support to the scholarly websites NINES and 18thConnect, and to their forthcoming peer site, MESA. With more than fifteen years of professional experience, Laiacona has acted as technical lead on digital projects funded by the National Endowment for the Humanities, the Andrew W. Mellon Foundation, and the National Institutes of Health.

Gregor Middell

Gregor is a scholar in the field of humanities computing currently contributing to a genetic digital edition of Goethe's Faust. Before joining the Faust project, he earned a Master's degree in Berlin, majoring in both Modern German Literature and Computer Science. Gregor's main research interests are firstly computer-supported collation with the aim to semi-automatically correlate electronic texts, analyze textual variance and determine intertextual relationships; secondly he is interested in markup theory/practice. To this end he participates in the development of two open-source projects, Juxta and the CollateX. As a lecturer in the Digital Humanities Program at Julius-Maximilans-Universität Würzburg, he also gained experience in teaching DH skills like text encoding, data modeling and programming.

Target Audience and Expected Number of Participants

This workshop is intended for software developers with a basic working knowledge of XML and Ruby. Knowledge of TEI is optional. It will be of interest to developers working on any text-based project, but especially editorial projects, projects handling variant texts, and projects dealing with XML encoded texts, including TEI.

This workshop is designed for between 5 and 20 participants.

Prerequisites

Workshop participants should bring:

- A working knowledge of Ruby. Suggested reading: (http://www.ruby-doc.org/docs/ProgrammingRuby/)
- A working knowledge of XML. Suggested reading: (http://www.w3schools.com/xml/default.asp)
- A laptop with Ruby installed, a working Internet connection, and a favorite code editor.

Built to Last: Sustainability Strategies for Digital Humanities Projects

Maron, Nancy
nancy.maron@ithaka.org
Program Director, Ithaka S+R

Scholars, librarians, and publishers today are building digital resources that are valuable for scholarship and teaching in the humanities, from multi-format research projects to digitized collections to born-digital works and innovative software tools. While some may be experiments and are valuable for the experience they offer or the capacity they build, others create collections of content, dynamic websites, or other resources that are intended to continue well beyond their initial creation. As these projects continue, their creators often face the challenge of identifying the financial and non-financial resources that will permit them to maintain their value to users over time.

For the past several years, the team at Ithaka S+R has been studying how project leaders develop successful sustainability plans, learning from hundreds of project leaders around the world who have spoken with us about the challenges they face in building their projects and in finding ongoing support for the activities they feel need to be sustained post-grant.

This half-day tutorial will introduce project leaders to the basics of sustainability planning, help them establish ambitious but realistic sustainability goals, define the challenges they face, and sketch out a hypothesis of their ideal funding model. The workshop will include group participation and will share real-world examples, illustrated by case studies of projects that really worked, or …didn't. The session will also allow participants to review the 'Funding Model Framework,' a tool designed by Ithaka S +R to help guide those leading digital resource projects in choosing and testing the funding strategies that will work best for them

We hope that by introducing new some ideas and practical tools in a supportive and engaging setting, this tutorial will encourage digital humanities project leaders in developing and testing new ideas to support their work.

Ithaka S+R is a not-for-profit research, training, and consulting service that has been studying the sustainability

of digital resources for several years. We are currently engaged in an NEH-funded research project on *Sustaining the Digital Humanities*, exploring institutional strategies for supporting this work. Other recent reports on this topic include *Sustaining Our Digital Future: Institutional Strategies for Digital Content* (2013); *Revenue, Recession, Reliance: Revisiting the Case Studies in Sustainability* (2011); *Funding for Sustainability: How Funders' Practices Influence the Future of Digital Resources* (2011); *Sustaining Digital Resources: An On-the-Ground View of Projects Today* (2009). Our reports and tools are freely available on the Ithaka S+R website at http://www.sr.ithaka.org/.

Instructors

Facilitator: Nancy Maron, Program Director, Ithaka S+R

Nancy Maron leads Ithaka S+R's program in Sustainability and Scholarly Communications, developing research, tools, and training to assist those responsible for funding, leading, or otherwise supporting digital resources in higher education and the cultural sector. She has led Ithaka S+R's recent studies on sustainability, including *Sustaining Digital Resources: An On-the-Ground View of Projects Today* (2009), *Ithaka S+R Case Studies in Sustainability* (2009), *Revenue, Recession, Reliance: Revisiting the Case Studies in Sustainability* (2011), and *Funding for Sustainability: How Funders' Practices Influence the Future of Digital Resources* (2011). Prior to joining Ithaka S+R, Nancy spent more than a decade in the book publishing industry, at Harry N. Abrams, Macmillan Library Reference, and the Perseus Books Group, where she was Director of Academic and Library Marketing. She holds a B.A. in Humanities from Yale University and an M.A. in French Studies and History from New York University.

Target Audience

Participants of this tutorial should be those with interest in and/or responsibility for charting a course for the development of a digital scholarly project or resource. This could include:

Academic project leaders who are leading or have created a digital resource.
Managers of digital collections and digitization units at cultural organizations, including museums, libraries, archives and other institutions.

Those in early stages of considering sustainability strategies for their projects are encouraged to attend. The maximum class size for this workshop is 40, to allow for best discussion and sharing of experience.

Brief Course Outline

9:00–9:20 Introductions; presentation of projects and sustainability challenges
9:20–10:00 Introduction to sustainability basics; defining your sustainability 'goal'
10:00–10:15 Sketching a funding model: The Funding Model Framework
10:15-10:45 Group work: your sustainability hypothesis
10:45-11:00 Break
11:00-11:30 Testing your hypothesis: The research phase
11:30–12:00 Defining next steps

Tutorial: Designing successful digital humanities crowdsourcing projects

Ridge, Mia
mia.ridge@open.ac.uk.
PhD candidate in Department of History, Open University, United Kingdom

Brief description of content or topic

Successful crowdsourcing projects help organisations connect with audiences who enjoy engaging with their content and tasks, whether transcribing handwritten documents, correcting OCR errors, identifying animals on the Serengeti or folding proteins. Conversely, poorly-designed crowdsourcing projects find it difficult to attract or retain participants.

This workshop will present international case studies of best practice crowdsourcing projects to illustrate the range of tasks that can be crowdsourced, the motivations of participants and the characteristics of well-designed projects. Attendees will learn about the attributes of well-designed humanities crowdsourcing projects and will be able to apply these lessons by designing and critiquing a simple crowdsourced task based on their own materials or projects.

Sample outline for half-day workshop:

9:00-10:30 Introductions, definitions, history and examples of types, tasks
10:30-10:45 Break
10:45-11:30 Ethics, participation and motivations in crowdsourcing; design tips for crowdsourcing projects
11:30-11:45 Break
11:45-12:30 Working in pairs/small groups: design a crowdsourcing project; optionally discuss results with the group

Contact information for workshop leader, including a one-paragraph statement of research interests and areas of expertise

Mia is currently researching a PhD in digital humanities (Department of History, Open University, United Kingdom), focusing on historians' use, evaluation of and contributions to scholarly crowdsourcing projects. She has published and presented widely on her research including user experience research and design for engagement and participation in cultural heritage and is editing a book called Crowdsourcing our Cultural Heritage (Ashgate, forthcoming). Formerly Lead Web Developer at the Science Museum Group, Mia has designed successful crowdsourcing projects, advised prestigious cultural organisations on usability and design for audience participation, and delivered full-day training workshops on crowdsourcing and data visualisation for scholarly research for the British Library's Digital Scholarship programme.

Description of target audience and expected number of participants (based, if possible, on past experience)

I have run this workshop before as a full-day workshop and have modified it for a half-day version for DH2013. Based on past experience, this tutorial can accommodate up to 24 participants. No technical or design experience is necessary but knowledge of potential or existing audiences for any relevant datasets or related tasks would be helpful in the design exercise. There are no special requirements for technical support.

Teaching Text Analysis with Voyant

Rockwell, Geoffrey
Professor of Philosophy and Humanities Computing at the University of Alberta, Canada

Sinclair, Stéfan
Associate Professor of Digital Humanities at McGill University

Introduction

One of the common skills covered in introductory digital humanities courses at both the undergraduate and graduate levels is computer-assisted text analysis. This workshop will introduce participants to ways of teaching text analysis with the online text analysis environment, *Voyant Tools* (voyant-tools.org). Unlike previous workshops that have been focused on using Voyant Tools for research, this workshop is aimed at participants who want to introduce text analysis into their teaching. Participants are not expected to know much about text analysis or Voyant; this workshop will include a brief hands-on component to introduce text analysis with Voyant as an example of what can be done.

Outline of the workshop

The workshop will take the following form.

Introductions: where the instructors and participants introduce themselves and their teaching context.
Brief example of teaching text analysis with Voyant: where participants are led through a hands-on introduction to Voyant as if they were students.
Discussion of example: where we discuss how the hands-on example tutorial could work (or not) in an undergraduate class.
Models for text analysis: where we break into groups that develop models for how they might use text analysis in a course. Some might develop a model for introducing text analysis in a literature course, some in a digital humanities course.
Managing the module: where we discuss what can go wrong and what learning resources there are.

Why bother: where we conclude with a discussion of the place of text analysis in the digital humanities curriculum.

What is text analysis and why teach it?

Text analysis is about asking questions of a text with the help of a computer. As such it is a research method of interest to any who interpret texts. Computer-assisted text analysis tools evolved out of the concordance as a tool for studying a text by searching it for patterns and gathering passages that agree in some way. Early text analysis tools like COCOA and OCP were designed to produce print concordances from the early electronic texts being entered in the 1960s and 70s (Lancashire 1986). With interactive programs like TACT in the late 1980s we saw tools designed to support research on the computer screen (Lancashire 1996). Now text analysis tools like Voyant are available as online web services that you can upload a text to. The newer tools can handle large e-texts and they provide text mining and visualization features. Tools like HyperPo and Voyant can be thought of as reading tools designed to provide multiple interfaces for interpretation (Sinclair 2003). These online tools make it possible to teach text analysis without having students struggle with the complexities of pre-indexing tools, special markup, or installing software.

Teaching text analysis has been part of introductory digital humanities courses because of the important place of electronic texts in computing in the humanities. Humanities computing grew out of early concording efforts like the Index Thomisticus of Father Busa (Busa 1980). As we developed models for how to represent texts in electronic form we began to ask what might be learned from these electronic texts. What questions might be asked of etexts that we couldn't ask in close reading of a text? Teaching text analysis is a way of letting students see the opportunities and limits to algorithmic criticism, as Stephen Ramsay calls it (Ramsay 2008). Teaching text analysis lets them engage with what the computer can really do in the way of analyzing (taking apart or tokenizing) information and synthesizing new views on information like visualizations. It is also a way of introducing students to research methods that they can use in their studies of texts.

Models for integrating text analysis into a course

Text analysis can be woven into a course in different ways. It can be integrated as a short module just to give students a taste or it can be taught in depth as a research method. In this workshop we will look at three models for integrating text analysis into a course:

Short module in an undergraduate class: First, we will look at how this can be taught as a one week module in an undergraduate class with a hands-on tutorial. This is the example that we will walk through with the participants as if they were students.

One day workshop: Second, we will look at how text analysis can be taught in a one day workshop for students and colleagues. We will discuss how instructors can set tasks for students to tackle on their own and how one can alternate hands-on instruction with discussion.

Research methods for graduate students: Lastly we will cover how one might teach text analysis to graduate students who are going to use it as a research method. In this context we will discuss readings that can be assigned and projects one can assign. We will talk about other tools and resources that graduate students may need for research projects. We will discuss how data can be prepared for Voyant and how data can be extracted from Voyant for other tools. Voyant can be part of a larger suite of tools graduate students use in real research.

In all of these cases there are some common types of materials we will have for participants:

Scripts: We will walk through online scripts that can be used for teaching Voyant. These scripts with links to all the resources needed are based on our working scripts that have been tested teaching Voyant around the world. (See http://hermeneuti.ca/workshops for example scripts).

Readings: We will share an annotated bibliography of readings about text analysis that can be used with students.

Examples: One of the hardest things to teach is how text analysis might be reported in a real research paper. We will share examples of research reports, papers and blog essays that show how others have used text analysis and woven it into assignments.

Other Tools: Students who are pursuing original questions almost always run up against the limitations of any particular tool. We will share a list of other tools and discuss how Voyant can be used both to analyze hybrid texts from other tools or to export data for use with other tools.

What can go wrong?

An important part of the workshop will be a final discussion of what can go wrong with Voyant in a classroom and how to deal with it. Courses that introduce computing tools need to be carefully paced and tested so that the technology does not hold back the learning. When teaching with any online tool you need to have contingencies for when the server goes down or is busy with other queries. In the case of Voyant we now have backup servers and a resolver that was set up specifically for training situations. In the workshop we will go over how to prepare for teaching Voyant, how to set up multiple versions of the indexed texts that are being used for a class, and other tactics for dealing with delays with Voyant. For those that are interested we will also discuss how they can set up Voyant on their own servers so that they have control.

VSim: A new interface for integrating real-time exploration of three-dimensional content into humanities research and pedagogy

Snyder, Lisa M.
lms@idre.ucla.edu
Urban Simulation Team, University of California-Los Angeles

Study of the built environment is central to humanities scholarship. The meanings inherent in urban plazas, simple homes, and lavish government buildings are integral to an understanding of the human condition. Even as a plethora of new tools and technologies encourage exploration of our physical world in three dimensions, integrating 3D content into academic research and pedagogy remains a challenge because of the limitations of available mass market software for educational interaction with virtual environments. VSim, a new NEH-funded prototype software interface for real-time exploration of 3D content, has the potential to significantly and positively impact humanities scholarship by fostering use of 3D content across grade levels and humanities disciplines, providing opportunities for engaging pedagogical activities, and opening up newavenues for humanities research.

VSim provides a much-needed real-time interface for academics working with 3D content. It allows users three modes of navigation, includes a mechanism for creating linear narratives through the virtual world that can be augmented by text and images (think PowerPoint or Prezi in 3D space), and provides a way for content creators to link to primary and secondary resources from within the modeled environment. Through these two features — the narratives and embedded resources — VSim provides scholars, educators, and students the opportunity to build knowledge through exploration of the virtual world as never before possible. It responds to the needs of inservice educators by supporting both teacher-centered presentations and student-centered exploration, providing the opportunity for students to actively engage with the content to build knowledge by creating a personalized virtual learning environment. And most importantly, VSim breaks down the barriers to instructional use of 3D content by providing a simple interface that easily re-purposes the crowd-sourced computer models built for Google Earth and available in them Google/Trimble 3D Warehouse.

For academics actively creating 3D content, VSim is a viable alternative to online virtual worlds and gaming platforms for the exploration, presentation, and distribution of their work. Within a project team, VSim can be used to interact with raw model files and any referenced component parts (e.g., linked texture maps and external references). Simple narratives and embedded resource files can be constructed to support that interactive experience; the annotation feature can be incorporated into the construction process as an ‘in world’ strategy for marking up the 3D content and generating ‘to do’ lists; and the metadata associated with the embedded resources or included as annotations can be used to promote a dialog about the modeled environment within the project team. When it is desirable to share the model with colleagues and peer reviewers, VSim can be used to export a single file for distribution. In creating this distribution file, content creators have the option to include information about the model and its creation, locked versions of their narratives and embedded resources, and to impose restrictions on the contents. The distribution model can be packaged with a branding overlay (e.g., a screen icon of the lab logo or text identifying the content creator), an expiration date beyond which time the model file will no longer launch, restrictions on the size of the simulation window to control performance and help curb unauthorized image capture, and restrictions on user navigation. The intent of the distribution and restriction options was to provide content creators a mechanism to protect their intellectual property and encourage them to share their work, thus facilitating secondary scholarship and educational re-use of these rich academically generated virtual worlds.

At the time of this writing, the interface is undergoing final testing before the VSim 1.0 release scheduled for Spring 2013. While every effort has been made to keep the software as simple and intuitive as possible to encourage

use — even by the technologically challenged — it is still new software that comes with a learning curve. Users have to be comfortable navigating in 3D space before they can engage with the pedagogical tools developed within the interface. It is the goal of the tutorial to provide a low-stress opportunity for academics to overcome that learning curve and become acquainted with key features of the software. (VSim, a sample model of the Pantheon, and user documentation are available for download at https://idre.ucla.edu/gisvisualization/vsim.)

Figure 1.
A screenshot of the Fine Arts building, now the Museum of Science and Industry, from the Urban Simulation Team's reconstruction of Chicago's World's Columbian Exposition of 1893, an example of a real-time model being built at UCLA and intended for educational use.

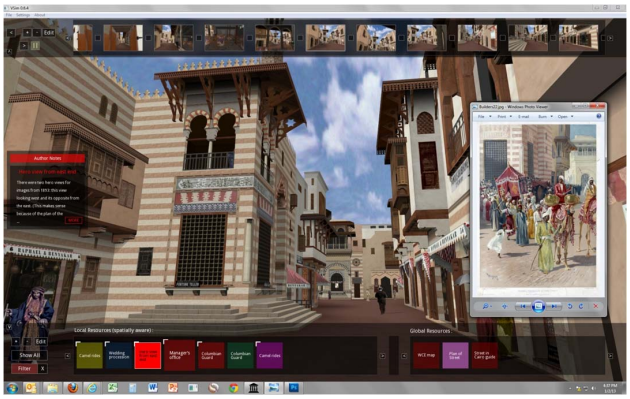


Figure 2.
In the screen snapshot above, the Urban Simulation Team's model of the Street in Cairo installation from the World's Columbian Exposition of 1893 is shown in VSim; the individual nodes of a linear narrative are visible in the bar across the top of the simulation window and the embedded resources available to the user are shown along the bottom of the screen. On the right, an illustration from Daniel Burnham's Book of the Builders has been opened for comparison with the computer model. On the left, note from the content creator discusses how the Street was photographed for 19th century viewbooks.

Tutorial Details

Instructor

Lisa M. Snyder will lead the tutorial. Snyder is a senior member of the Urban Simulation Team at UCLA and on staff with the Institute for Digital Research and Education, a division of UCLA's Office of Information Technology. She is also a principal investigator on the NEH-funded VSim project. Her research is focused on the educational use of largescale interactive computer reconstructions of historic urban environments. She is currently working on a real-time reconstruction of the World's Columbian Exposition held in Chicago in 1893. This model is regularly showcased at the Museum of Science and Industry in Chicago and appears in the documentary *Make No Little Plans: Daniel Burnham and the American City.* Snyder was also the primary modeler on the interactive computer reconstructions of the Temple Mount site that were developed jointly by the Urban Simulation Team and the Israel Antiquities Authority for the Davidson Center in the Jerusalem Archaeological Park. (For more information: http://www.ust.ucla.edu/ustweb/Projects/columbian_expo.htm and http://www.ust.ucla.edu/ustweb/Projects/israel.htm)

Target Audience

The proposed half-day tutorial will introduce VSim to interested humanities scholars. Probable participants would include academics across the humanities disciplines who are working with 3D content, supervising students on historic reconstruction projects, using available 3D content to supplement their ongoing research activities, or interested in integrating computer models into their seminars, classrooms, and conference presentations. Interest in this tutorial is difficult to predict, so has been organized to work whether capped at 20-25 to ensure individualized attention to each participant or opened up to a larger audience.(The only impact to being flexible is the size of the volunteer pool for the participant presentations scheduled towards the end of the tutorial.) The need for an advance CFP is not anticipated.

Session Outline

Introductory Presentation (30 minutes)

Snyder will begin the tutorial with an introductory presentation about the main features of VSim with a demonstration of the Urban Simulation Team at UCLA's model of the World's Columbian Exposition of 1893. This large-scale environment is an appropriate testbed

for VSim because it includes enough detailed content to allow multiple narratives, an extended suite of embedded resources, and opportunities for user-guided exploration. The introduction will also cover opportunities for humanistic research in 3D environments, suggestions for pedagogical use, strategies for using VSim in presentations and for assignments, and discussion of the metadata associated with the 3D content, narratives, and resources embedded in VSim.

Hands-On Training (2 hours and 30 minutes)

Identifying Content (15 minutes) Discussion of 3D modeling packages that can export VSim compatible files (e.g., Trimble's SketchUp or Autodesk's 3ds Max); how one might use VSim to share 3D content with colleagues; downloading 3D content from Trimble's 3D Warehouse. (From this point forward, participants will use VSim to interact with their own content, a model prepared for the tutorial, or a model downloaded from 3D Warehouse.)

Navigation Basics (15 minutes) Instructions for working with the three modes of interaction (WASD, Flight Simulation, Object Navigation); the following break an also be used to help any participants having difficulties mastering navigation

BREAK (15 minutes)

Building Narratives (30 minutes) Instructions for building narratives within the virtual environment: establishing narrative nodes; adding overlay text and images; adjusting timing on nodes and transitions; exporting narratives

Embedded Resources (15 minutes) Instructions for embedding resources within the virtual environment: metadata for resources; file types supported; adding annotations, linked files, and URLs; the auto launch setting; the auto reposition setting; exporting embedded resources; exporting .vsim packaged files

*Individual Work Session (30 minutes)*Participants are given time to finesse presentations and consider applications for the software for teaching and research

Participant Presentations (20 minutes) Three to four volunteer participants will share the projects they've created in the course of the tutorial; this will require that participants hook up to the projector and present from the front of the room.

Concluding Discussion (10 minutes)

Participant Presentations (20 minutes) Three to four volunteer participants will share the projects they've created in the course of the tutorial; this will require that participants hook up to the projector and present from the front of the room.

Concluding Discussion (10 minutes)

VSim: A summary of critical features

(Consider this an Appendix; included in case any reviewers want more information on the software.)

VSim includes key functionality for scholarly interaction with 3D content that can be broken into three categories: functions necessary to import, display, and navigate through the threedimensional content; functions that allow the content contributor and end users to augment the virtual world with multi-media content; and functions that provide the content creator with controls over their intellectual property.

Functions to import, display and navigate three-dimensional content.

a) *Cross-platform operability:* VSim is being released in both Windows and Mac versions with simple user documentation.

b) *Loading:* The software loads multiple three-dimensional file formats including COLLADA (.dae — an industry standard exchange format for 3D content), open flight (.flt — the format used by Presagis' Creator), and the native OSG formats (.osg, .ive, etc.). The software's ability to load COLLADA files is particularly important as this makes the entire catalog of 3D buildings in Trimble's 3D warehouse available to VSim users.

c) *Interaction:* VSim supports three modes of interaction: gamer-style WASD navigation, flight simulation, and Google-Earth style object rotation. This flexibility enables the user to choose how they interact with the virtual environment including a first person point-ofview with the opportunity for unlimited control in all three dimensions.

e) *Temporal changes to the environment:* The software supports two mechanisms to incorporate temporal elements in a 3D model: switches that can toggle between alternative scenarios and a time slider that allows the user to step through a sequence of options within the modeled environment organized to simulate an unlimited number of construction phases or changes over time (.flt files only).

f) *Image output:* VSim generates static images and continuous image sequences that can be assembled into

digital video files. These assets can be included in an academic paper or submitted as part of an assignment

Functions to allow the content contributor and end user to augment the virtual world.

a) *Narratives:* Through the narrative feature, either the content creator or end-user can define linear sequences through the virtual space akin to a PowerPoint or Prezi presentation in 3D. Creating a narrative involves establishing a series of key frames within the environment, augmenting those 'nodes' with text and images, and adjusting the timing for pauses on the nodes and the transitions. VSim automatically creates the movement from one node to the next. The mechanism provides an opportunity for a 'tour' mode that could be used by educators or in museum installations.

b) *Embedded resources:* Either the content creator or end-user can embed supplementary primary and secondary resources for display during exploration of the virtual world. (The image below illustrates use of embedded resources within the Experiential Technologies model of the Pantheon.) Each resource has its own parameter settings and metadata: links can be constantly available or only visible at specifics points in the model, they can be set to auto launch when the user enters their activation zone or reposition the user so that the simulation view angle matches that of the resource (e.g., to compare the real-world site with the computer model). Resources are launched using the default programs identified by the system's settings, and can be searched, filtered, and organized into categories set by the content creator. Resources can be embedded within the model to support an academic argument and/or populated for use during free navigation. The combination of the narrative and embedded resources establishes the ground work for the construction of lesson plans, arguments, and narratives by either the content contributor (as subject expert) or a student user (a constructivist learning exercise).

c) *Annotations* Annotations are a specific type of embedded resources that are the equivalent of comments or sticky notes that can be added within the virtual world. This feature can be used by content creators and/or end users in a plethora of ways: to pose questions to collaborators or students in a learning activity, in a series of personal notes from the content creator to end users ("the colors used in this reconstructed element were based on contemporary buildings by the same architect"), or as discrete elements of a non-linear argument set to auto launch as users explore the virtual world.

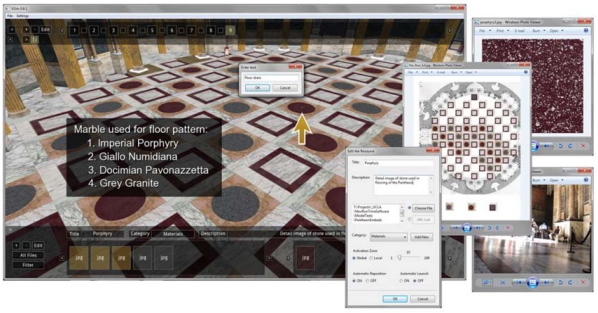


Figure 3.

Functions that provide the content creator with controls over their intellectual property.

a) *Binary file format for distribution:* The aptly titled "To Share or Not to Share?" report describes the hesitation that many scholars feel about sharing their data before they've had a chance to wring out the last bits of publishable results. Three-dimensional content creators are no different, and are particularly protective of raw modeling files. To encourage broad distribution of content, VSim allows content creators to load raw files and export an aggregated model file for distribution. This file would require considerable computing skill to 'reverse engineer,' thereby offering protection against undesirable use. VSim also allows an expiration date be added to the .vsim file to ensure control over use, appropriate updating for works in progress, and time limits that may come into play when subscription service enacted.

b) *Branding:* An overlay that can be added to the simulation viewport and is visible during any user interactions. The branding overlay can include both image files (e.g., lab logos) and text (e.g. "For Educational Use Only"). The user has the option to lock the content creator settings — and therefore the branding overlay on or off — at the point of export.

c) *Locks on model, narrative, and embedded resource files:* The export feature for narratives and embedded resources includes the opportunity for the creator to the lock the resultant file (essentially creating a READ ONLY file). The intent of this feature was to encourage content creators and/or end users to share files generated for VSim models –. For narratives, all nodes and node overlay information are packaged into a single file for distribution that can be shared and replayed by any user with a copy of the 3D model for which it was created. If locked, no changes can be made to the narrative. Exported embedded resource files can also be locked,

but only the annotations and links to websites will be preserved. (This decision was made to control the size of the distribution file; hundreds of resources for a large-scale environment might easily exceed 10GB of data.)

Taking modeling seriously: A hands-on approach to Alloy

Sperberg-McQueen,C. M.
cmsmcq@blackmesatech.com
Information-technology consultant, Black Mesa Technologies, LLC

Description

The modeling of humanities data is a core activity (some say *the* core activity) of the digital humanities. The activity so described may take a wide variety of forms; often the term is used for any compact description of a domain, whether in prose or in user-interface metaphors. Machine-processable descriptions are probably more common, but these, too, vary: the definition of an XML vocabulary, the table declarations for a SQL database, the data structures or even the executable code of a program may all be described informally as offering a 'model' of some domain or other.

The term *model*, however, is here applied more narrowly to expressions in some well defined formalization. Models are most useful when formalized in a declarative not a procedural notation and when their logical import is clear. Formulating precise models can be difficult. Inconsistencies and unforeseen interferences between parts of the model can easily creep in. With informal definitions, such shortcomings can remain undetected for long periods, even until after the model has been put to use. Formally defined models, on the other hand, can be tested systematically for logical consistency; their consequences can be established systematically. Such testing can help uncover shortcomings in a timely manner.

Alloy is a tool for "lightweight formal methods", which makes it easier to test the implications of models and to check assumptions for plausibility, consistency, and completeness. Its usual application area is the testing of software designs but the variant of first-order logic provided by Alloy is by no means limited to the description of software or electronic objects. It has been successfully used to formalize notions far removed from any software, including the nature of transcription, an application of the type/token distinction to document structure, and fragments of Goodman and Nelson's mereology and of Hilbert's formulation of Euclidean geometry. Alloy's logic is powerful enough to formulate interesting concepts, while remaining weak enough to be tractable for machine processing. Using Alloy's syntax, a modeler can formulate the axioms of a model and augment them by asserting that certain properties hold for all instances of the model, or by defining predicates which characterize particular instances of the model. The Alloy Analyzer can test the assertions and illustrate the predicates, by seeking counter-examples to the assertion or instances of the predicate.

This one-day tutorial introduces digital humanists to the use of Alloy for modeling. Topics include:

- introduction to Alloy's logic
- compressed summary of Alloy syntax
- use of Alloy for formulating assertions and predicates
- describing individual test cases for Alloy
- Alloy's place in the larger context and Alloy's relation to light-weight formal methods, to other formal methods (e.g. Z), and to theorem-provers
- limits on Alloy's logic, from a theoretical point of view (how Alloy and other tools deal with Goedel's incompleteness result and Turing's halting problem), and from a practical point of view (modeling recursion using transitive closure)

Examples will be drawn from domains discussed at recent DH conferences.

Prerequisites: some prior exposure to symbolic logic and/or programming is probably desirable; failing that, highly motivated participants may be able to benefit from the workshop if they have sufficiently high tolerance for exposure to new material.

Participants should bring a laptop computer with a current installation of Java; they may optionally preinstall Alloy 4.2 or they may install it during the workshop.

Target audience and expected number of participants

Short answer: not a large target audience (but a choice one!); estimated attendance perhaps 5-10 (no evidence).

The target audience consists of digital humanists interested in techniques for formalizing important concepts and tools for working with such formalizations. The tutorial deals with high level data modeling concepts. Some prior exposure to symbolic logic and/or programming is desirable; failing that, highly motivated participants may be able to benefit from the workshop if they have sufficiently high tolerance for exposure to new material. Participants

should bring a laptop computer with a current installation of Java; they may optionally pre- install Alloy 4.2 or they may install it during the workshop.

Outline

Full-day outline

I'd prefer to teach this as a full-day tutorial; that allows time for a mixture of lecture-style presentation of information and hands-on exercises. A tentative full-day schedule is:

9:00-10:30 Introduction to the course

- Modeling, formal logic, formal methods. Lightweight formal methods; Alloy.
- Demonstration: Alloy model of a Web interface (capabilities, security issues, user information).
- Demonstration: Using Alloy to generate test cases.
- The small-scope hypothesis; how Alloy manages to be useful despite Goedel's Theorem.
- Hands-on exercise: Using the Alloy Analyzer.

10:30-11:00 Break

11:00-12:30 Alloy's first-order logic

- Atoms, relations, tuples, sets. Basics of syntax: signatures, relations, multiplicities.
- Hands-on exercise(s) (logic puzzles, simple proofs from logic textbooks).
- Styles of expression: predicate-calculus style, navigational style, relational style. More syntax: assertions, predicates, quantification, let-expressions.
- Using Alloy to model concepts: FRBR entities, metadata records, XML and non-XML document structures.
- More exercises(s).

12:30-2:00 Lunch

2:00-3:30 Alloy as a tool for software design

- Examples: using Alloy to model an interactive concordance system, a query interface, a database system.
- Hands-on exercises.
- Idioms for modeling state, change, and dynamic systems in Alloy.
- Idioms for testing specific instances with Alloy.

3:30-4:00 Break

4:00-5:30 Recursion, Conclusion

- Using transitive closure to model recursion.
- Hands-on exercises.
- Review, questions, clarifications.
- Where to go from here? Further Alloy resources, other tools for formal methods and theorem proving.

Panels

The Design of New Knowledge Environments

Blandford, Ann
a.blandford@ucl.ac.uk
University College London

Brown, Susan
sbrown@uoguelph.ca
University of Guelph

Dobson, Teresa
teresa.dobson@ubc.ca
University of British Columbia

Faisal, Sarah
s.faisal@cs.ucl.ac.uk
University College London

Fiorentino, Carlos
carlosf@ualberta.ca
University of Alberta

Frizzera, Luciano
dosreisf@ualberta.ca
University of Alberta

Giacometti, Alejandro
alejandro.giacometti.09@ucl.ac.uk
University College London

Heller, Brooke
brooke.heller@gmail.com
University of British Columbia

Roeder, Geoff
geoff.roeder@gmail.com
University of British Columbia

Peña, Ernesto
ernesto.pena@gmail.com
University of British Columbia

Ilovan, Mihaela
ilovan@ualberta.ca
University of Alberta

Michura, Piotr
pmichura@asp.krakow.pl
Academy of Fine Arts in Krakow

Nelson, Brent
brent.nelson@usask.ca
University of Saskatchewan

Mohseni, Atefeh
amohseni@ualberta.ca
University of Alberta

Radzikowska, Milena
mradzikowska@gmail.com
Mount Royal University

Rockwell, Geoffrey
grockwel@ualberta.ca
University of Alberta

Ruecker, Stan
sruecker@id.iit.edu
IIT Institute of Design

Sinclair, Stéfan
stefan.sinclair@mcgill.ca
McMaster University

Sondheim, Daniel
sondheim@uvic.ca
University of Victoria

Vela, Sarah
svela@ualberta.ca
University of Alberta

Windsor, Jennifer
jwindsor@ualberta.ca
University of Alberta

Yi, Tian
tianyi2@ualberta.ca
University of Alberta

Dergacheva, Elena
dergache@ualberta.ca
University of Alberta

1. An Introduction to the Design of New Knowledge Environments

Ruecker, Stan | Rockwell, Geoffrey | INKE Research Group

In this panel, we report on our year 4 work in the Interface Design (ID) research team of the Implementing New Knowledge Environments (INKE) project. INKE is a 7-year major collaborative research initiative (MCRI) project funded in Canada by the Social Sciences and Humanities Research Council (SSHRC). INKE is led by Ray Siemens at the University of Victoria, and includes dozens of researchers worldwide. Each year, we have had a different focus for our work. The schedule is as follows:

Year 1: interdisciplinary citations
Year 2: corpora
Years 3 and 4: the scholarly edition
Year 5: the monograph and journal
Year 6: born-digital literature
Year 7: wrap-up and dissemination

For the first three years of the project, our focus has been on a wide range of experimental prototypes. Beginning in year 4, we attempted to aggregate these prototypes into a smaller set of "new knowledge environments", where the goal is no longer to have a single piece of functionality that we can experiment with, but instead to begin imagining how the various pieces of functionality, as represented by our prototypes, can work in concert to produce an environment for people working with electronic books.

What makes these environments "new" is that they are comprised of a set of experiments into working with digital text. More often than not, these experiments involve the design and programming of a prototype, which may exist at any one of a range of levels of fidelity. Ideally, the lowest level of fidelity is developed that is necessary in order to come to grips with the central idea. To develop further is sometimes required in order to achieve an adequate user experience study, but in any case it is important to keep in mind that the end goal is the extension of our understanding rather than the production of a piece of stable software.

So the process is to conceive of a concept and produce a prototype to help us better understand the concept. Typically a prototype will generate some new knowledge itself, and that knowledge can contribute to the next iteration of the idea. In some cases, the prototype teaches us enough that there is no need for a further prototype. If what we've learned is of potential use, then we can consider a more robust implementation of the idea. If what we've learned is that the line of thought we've been pursuing may not be fruitful after all, then we can at that point abandon the trajectory and move to another idea.

In the case of the year 4 projects in INKE, we have selected some prototypes that we think deserve to be aggregated into an environment. We are at the same time trying to learn what we can about the theoretical and technical issues involved in this kind of redesign and reprogramming for the purposes of aggregation. These are environments rather than simpler tools in the sense that they afford more than a single task or set of tasks.

In the papers that follow, we describe each of the environments, examine the state of the art in scholarly editions for tablets, and discuss the user experience study of two of the prototypes that have gone on to inform one of our new knowledge environments.

2. Reading Skins: Voyant and Tool Aggregation

Rockwell, Geoffrey | Sinclair, Stéfan | INKE Research Group

Tools are not just ways to ask questions - they are also reading skins. Interface designers have known this for some time. They design interfaces to present affordances and views that facilitate different types of reading. A dictionary is designed for consultation reading (Blair 2010), a manual for training. Likewise an e-book minimizes distractions, presenting "just the text", while text analysis environments embed the texts in alternative ways of reading, from visualizations to interactive controls. These design choices are based on a model of the reader and the tasks they are engaged in. In this paper we will look at the types of interfaces or "skins" presented by text analysis environments and end by discussing the decisions taken in the design of Voyant. We will do that in the following ways:

1 First, we will look back at one of the first computing humanists to consider visualization and the interface to text analysis tools, John Smith.
2 Second, we will survey different types of text analysis interfaces.
3 Third, we will close by discussing how the architecture of Voyant is designed to allow for different reading skins that aggregate different tools to suit different uses.

1. First Thoughts on the Interface

One of the first computing humanists to think about the interface to text analysis tools was John Smith. John Smith in “Computer Criticism” and other articles proposed a way that analytical tools like his ARRAS could fit into interpretative research practices. He saw the computer as a tool to help identify and then trace structures through a text and gave an example of how this can help rereading a text in his article “Image and Imagery in Joyce's Portrait.” In “A New Environment For Literary Analysis” he explicitly discussed how the analytical tool ARRAS was not meant to replace the inquirer but to amplify them, how it could provide what we today call visualizations, and how it should be thought of not as a program, but as an environment for work where one could switch from text analysis to editing the text or sending a message. In the presentation we will go into more detail about how Smith articulated the relationship between tool design and interpretative practices.

2. Survey of Type of Interface

In the second part of the paper we will survey various interface paradigms for text analysis tools including:

- ARRAS: We will start with the interface John Smith developed for ARRAS, and discuss his ideas for a humanities accessible command line language.
- TACT and TACTweb: We will then discuss TACT, which was released in 1989 and designed by John Bradley and others (Lancashire 1996). TACT, while running in MS DOS, was influenced by the then new idea of a Graphical User Interface (GUI). It had a primitive windowing model that let you split the screen to see multiple displays and use one to drive the other(s). TACTweb, which came later, brought TACT functionality to the web, and illustrated for the first time how the web separated interface from text database so that multiple interfaces could be built.
- HyperPo: While HyperPo (hyperpo.org) wasn’t the first text analysis environment on the web, it was one of the first to fully exploit the web. It let you upload a text and it provided a number of innovative features including making displays themselves affordances for further interaction. It was also explicitly designed as a reading environment.
- TextArc: One of the most beautiful interfaces to text analysis is the TextArc (textarc.org) visual concordance designed by W. Bradford Paley. This work pushes the idea of interface in interesting directions as it can be considered a work of art or design meant to be appreciated in and of itself rather than as a window onto something else.

Other interfaces could be mentioned like the visual programming interface idea of Eye-ConTACT that is also available in SEASR or the library interface of the MONK project, but these are more management interfaces than reading ones.

3. Voyant and Skins

In the last part of the presentation we will discuss the layered architecture of Voyant that allows one to create different combinations of tools into “skins” that aggregate different tools. A Skin Builder tool for creating your own skins will be demonstrated and different examples of skins for different reading purposes will be shown. Different uses call for different combinations of tools.

All of these text analysis interfaces are presented from the perspective that they are views on a static object that is studied, but what if the object of study is changing? What if we think of text analysis tools performing the text rather than skin it? The presentation will end with an alternative prototype interface designed not for reading, but for animating the text as if it were a performance.

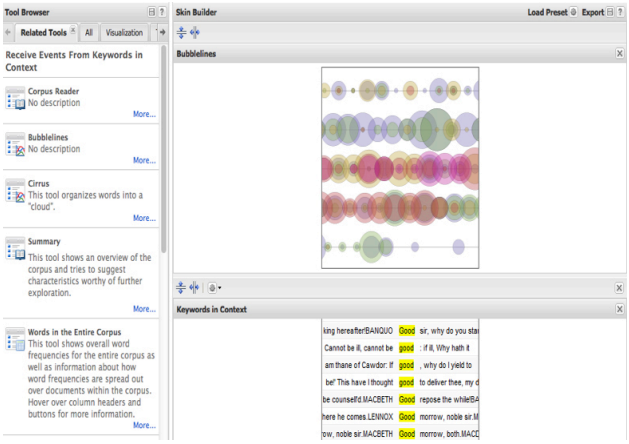


Figure 1 -
Voyant Skin Builder

References

Blair, A. (2010). *Too Much to Know: Managing Scholarly Information Before the Modern Age*. New Haven: Yale University Press.

Galey, A., and S. Ruecker (2010). How a Prototype Argues. *Literary and Linguistic Computing* 25(4): 405–424.

Lancashire, I., et al. (1996). *Using TACT with Electronic Texts: a Guide to Text-analysis Computing Tools, Version 2.1 for MS-DOS and PC DOS*. Modern Language Association of America.

Parunak, H. V. D. (1981). Prolegomena to Pictorial Concordances. *Computers and the Humanities* 15(1): 15–36.

Rockwell, G., S. Sinclair, et al. (2010). Ubiquitous Text Analysis. *The Journal of the Initiative for Digital Humanities, Media, and Culture* 2(1).

Sinclair, S. (2003). Computer-Assisted Reading: Reconceiving Text Analysis. *Literary and Linguistic Computing* 18(2): 175–184.

Sinclair, S., and G. Rockwell (2009). Between Language and Literature: Digital Text Exploration. *Teaching Literature and Language Online*. In Lancashire, I. (ed. and introd.). vii, 460 pp. New York, NY: Modern Language Association of America. 104–117. Options for Teaching (OfT): 26.

Smith, J. (1984). A New Environment for Literary Analysis. *Perspectives in Computing: Applications in the Academic and Scientific Community* 4(2): 20–31.

Smith, J. (1989). Computer Criticism. *Literary Computing and Literary Criticism: Theoretical and Practical Essays on Theme and Rhetoric*. University of Pennsylvania Press. 326–356.

Smith, J. (1973). Image and Imagery in Joyce's Portrait: A Computer-Assisted Analysis. *Directions in Literary Criticism: Contemporary Approaches to Literature.* Festschrift for Henry W. Sams, (ed.) Stanley Weintraub and Philip Young (University Park and London: Pennsylvania State University Press): 220–227.

Smith, J. (1985). Arras User's Manual: TR85-036. Chapel Hill, NC, The University of North Carolina at Chapel Hill.

3. Designing the interface within the interface: legibility and readability in the Dynamic Table of Contexts

Windsor, Jennifer | Brown, Susan | Nelson, Brent | Radzikowska, Milena | Sinclair, Stéfan | INKE Research Group

> "Typography is to literature as musical performance is to composition: an essential act of interpretation, full of endless opportunities for insight or obtuseness."
>
> — Robert Bringhurst, The Elements of Typographic Style

As humanities scholars transition from reading traditional print texts to reading on computer screens, ebooks and tablets, we are discovering that the visible word has taken a step backwards in quality. Typographic considerations are often the most challenging aspects of user experience development of digital reading environments. Ereaders are still in their infancy, and thus far little attention has been paid to textual design beyond very basic choices of typeface and font size.

We have developed the Dynamic Table of Contexts, a text analysis environment that combines the traditional concepts of the table of contents and index to create new methods for Humanities scholars to read and interact with digital text. The main interface consists of four interactive panes that perform dynamically with each other: the table of contents, the index, the xml tag list and a large pane to read the text itself (see Figure 1). In addition, there is a pane for reader notes. The Dynamic Table of Contexts interface is predominantly textual in nature, but the text at the centre of it is itself an interface – the point of interaction between the reader and that which is read. This study examines the appearance and arrangement of the textual interface within the larger interface. We asked: how can we typographically optimize the ereading experience?

The guiding fundamentals of typography are legibility and readability. Legibility concerns the ease with which a letterform can be recognized and readability refers to the ease with which text can be understood (Lupton 2004). Many problems with legibility have arisen in the transition from print to screen. For example, rather than black ink on a paper page, black text on a screen is an absence in the glowing pixels that surround it, and the result is perceived as blurry regardless of screen resolution. Anti-aliasing is used to compensate for low resolution on most computer monitors but can't be applied consistently from character to character because of where the letter may land on the physical pixel grid. In an effort to offset this problem font size is often increased, which in itself can become a reading irritant and navigationally awkward as larger type creates shorter lines of text which in turn requires more left-right eye movements from the end of one line to the beginning of the next. Fewer lines of text also necessitate more scrolling which is visually uncomfortable and requires additional time and visual energy for the reader to relocate himself in the text after each movement. While we wait for technology to catch up in these areas, this study outlines the methods we used to minimize legibility issues and discusses which existing typefaces (most of which, it must be remembered, were never intended for the screen) best counteract blurring and movement. We identify optimal type size and line length combinations for comfortable extended on-screen reading in the Dynamic Table of Contexts.

Compared to legibility, issues of readability are less often addressed in discussions of digital reading

environments. Readability is comprised not only of the arrangement of type on a page or screen, but also of attention to the entire visual entity and all the complex relationships between levels of type, symbols and images (Berryman 1984). Typography imparts semantic meaning as it interprets content and influences meaning by creating hierarchies, assigning values, and manipulating emphasis (Bachfischer, Robertson and Zmijewska 2007). Our investigations into optimal readability design for the Dynamic Table of Contexts raised interesting questions about the representation of text and the process of reading on-screen. In the remediation from print to screen, do we still read the same way? Do the rules that apply to the readability of print typography apply to on-screen typography? Do conventions that govern readability of websites also apply to extended reading of scholarly materials?

In this study we have also examined other components of the textual interface that affect readability but that are not necessarily typographic in nature. Often on-screen paging representations (such as a simulated left and right sides of a double page spread or an animated page turning) give the illusion of a traditional print book. We examine theories of whether these provide valuable perceptual cues to the reader or are simply a vestigial device that has lost its significance in a changing environment.

This research represents a move from an adequate to an optimized reading environment.

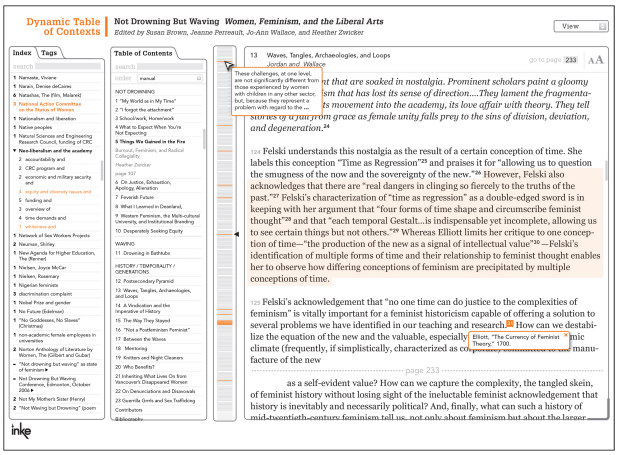


Figure 2 -
Dynamic Table of Contexts

References

Bachfischer, G., T. Robertson, and A. Zmijewska (2007). "Understanding Influences of the Typographic Quality of Text." *Journal of Internet Commerce* 6(2): 97–122.

Berryman, G. (1984). *Notes on Graphic Design and Visual Communication*. Los Altos, California: W. Kaufmann.

Bringhurst, R. (2004). *The Elements of Typographic Style*. 3rd ed. Hartley and Marks Publishers.

Larson, K. (2004). "The Science of Word Recognition: Or How I Learned to Stop Worrying and Love the Bouma." July. *Microsoft Typography site.* Retrieved August 20, 2006. http://www.microsoft.com/typography/ctfonts/WordRecognition.aspx

Lupton, E. (2010). *Thinking with Type: A Critical Guide for Designers, Writers, Editors, & Students*. New York. Princeton Architectural Press.

Mangen, A. (2008). "Hypertext Fiction Reading: Haptics and Immersion." *Journal of Research in Reading* 31(4): 404–419.

Weinzettelova, S. (2012). "Traditional Type in the Digital Era." *Bulletin - Prague College Centre for Research and Interdisciplinary Studies* 2012(2): 5–24.

4. The Tablet as a New Medium for Scholarly Editions

Mohseni, Atefeh | Sondheim, Daniel | Frizzera, Luciano | Rockwell, Geoffrey | Ruecker, Stan | INKE Research Group

How have scholarly editions been implemented on tablets? In the past few years, a significant number of scholarly editions have seen deployment on the Web, and the differences between printed editions and Web-based ones has become a matter of attention for many scholars. In this paper, we will investigate scholarly editions as they appear in tablets and other mobile devices.

Web-based scholarly editions offer many new features that do not and cannot exist in printed ones. Such features include opportunities for collaboration with other users or with the editors of the work, innovative methods to search and browse, specialized tools and visualizations for text analysis, means to customize the layout of the interface and the content, options to change the language of the interface or to see translations of the text, high resolution zoomable images, multimedia elements such as video and audio, and a wealth of material that would be too costly or cumbersome to include in a printed edition.

Despite these advantages, some users still prefer their scholarly editions to be in paper form. Reasons may include a lack of physical distance between reader and book, and the ability to interact directly with the material. Tablets have solved these problems to some extent, emulating the physicality of paper books, and allowing users an opportunity to touch and feel them. As Elena Pierazzo notes,

"Usability studies have demonstrated that reading on tablets is more enjoyable than reading on the screen of computers and, in some cases, more than reading print" (Pierazzo 2011). Additionally, the fact that an app is an independent program housed on a particular machine results in increased speed and stability over Web-based editions, which also serves to reproduce some of the positive features of paper-based editions (McDayter 2012).

Tablets represent a return to printed editions in some more negative respects as well. For instance, users cannot create shared annotations or collaborate with each other in other ways; the app on the tablet is isolated from references and related material available on the Web; tools for text analysis are absent; and no options are provided to show the interface in different languages. Part of the reason for these omissions in tablets may be due to the fact that editions that are currently available on tablets have generally been produced more for a popular audience than a scholarly one. This is reflected not only in terms of functionalities of the interface, but with respect to the content as well. Bibliographies, glossaries, and textual apparatuses are typically missing, variants tend not to be included, and notes and annotations have been recycled from older print-based editions, rather than being the result of new scholarship. Tablet-based editions are often more focused on making an edition attractive and amusing than on making it scholarly, and may include games or quizzes, methods of sharing passages or images via social media sites, and other innovative but potentially distracting features likely to be of interest to keen fans of the material rather than to scholars.

So, tablets are in a way mediating between print and the Web, sharing some of the advantages and disadvantages of both. But what effects will these advantages and disadvantages have on scholarly editions? Do scholarly editions have a future in tablets? These are questions that we will discuss in this paper.

At this point, very few tablet-based scholarly editions have been released, and as discussed above, those that have are decidedly unscholarly in some respects. Touch Press has produced some notable examples, including editions of T.S Eliot's *The Waste Land*, Leonardo da Vinci's notebooks, and Shakespeare's sonnets. *The Waste Land* is the first edition produced by Touch Press, and continues to stand up as a nice example of the genre. This edition allows the poem to be viewed as plain text, and includes annotations, audio recordings, facsimile images of the original typescript, filmed performance of the poem, interviews about the poem and some image galleries.

Many of the features provided in *The Waste Land* would make using scholarly editions easier and more pleasurable. As is the case with most works of philosophy, art, and literature, *The Waste Land* is multi-layered, and multimedia is of great help in exploring, finding and analyzing the different existing layers. Study is also eased and enhanced by being able to quickly and easily reveal or hide information, switch between viewing multiple encodings of the text, and simply by having the ability to manually scroll through the text, rather than having to use a device.

With what has been discussed, some questions require further exploration. Will the advantages of tablet-based editions win out over their disadvantages? Will the tablet become accepted as an appropriate medium in which to produce and study scholarly editions? In this paper, we will consider whether scholarly editions have a future on tablets and how such a future might look.

References

Eliot, T. S. (2011). *The Waste Land* for iPad, ed. Justin Badger and Charles Chabot, Touch Press LLP, and Faber and Faber. http://touchpress.com/titles/thewasteland/

Pierazzo, E. (2011). Tablets Apps, or the future of the Scholarly Editions? *Random Thoughts of a Digital Humanist with a Passion for Cookery*, 27 November, 2011. http://epierazzo.blogspot.ca/2011/11/tablets-apps-or-future-of-scholarly.html

McDayter, M. (2012). Are We There Yet? Touch Press's *The Waste Land* for iPad. February 25, 2012. http://clickherefordigitalhumanities.wordpress.com/2012/02/25/are-we-there-yet-touch-presss-the-waste-land-for-ipad/

5. Designing for Multi-Touch Surfaces as Social Reading Environments

Frizzera, Luciano | Vela, Sarah | Ilovan, Mihaela | Michura, Piotr | Sondheim, Daniel | Rockwell, Geoffrey | Ruecker, Stan | INKE Research Group

The INKE Interface Design group has been exploring the application of multi-touch table technology for improving the comprehension, manipulation, and analysis of variorum editions beyond what has been accomplished in previous digital variorums. These volumes consist of three general components: the text of the work itself as selected by an editor; a list of variations between this 'base-text' and other manuscript or printed versions; and a comprehensive anthology of previous scholarly annotations on a particular line, passage, or the entire work. For many texts the existing notes are so extensive that it is difficult in a print medium to present "…simultaneous access to text and relevant

commentary", leading to an effort since the mid-nineteen-nineties to produce digital variorum editions (Werstine 2011).

Scholarly editions, especially variorums, raise interesting questions about the representation of elements and the act of reading. The richness of this type of edition creates dilemmas related to the organization of different pieces of information, as well as interacting with the text. A print version presents physical space constraints, such as the two-page spread and the necessity of linear presentation. This constraint become less of an issue in a digital environment, where space, time, and dimensionality in general are more fluid.

Although current iterations of digital variora (for example, the Online Chopin Variorum Edition, and the Electronic Variorum Edition of Don Quixote) attempt to take advantage of this flexibility of the medium, they present themselves as flat webpages, losing physical engagement with the materiality of the book. In order to bring back fuller physicality, we have used multi-touch surface technologies to simulate the "real space", and to return to full gestures as opposed to clicks. A few projects focusing on eBooks and tablets have begun to emerge with the same idea (for example, *The Wasteland*, *Shakespeare's Sonnets* and Kerouac's *On the Road*), but this technology has not yet been applied to variorum editions.

Our project, *The Comedy of Errors Tangible Variorum*, involves creating a tangible user interface representing the Modern Language Association's variorum edition of Shakespeare's *Comedy of Errors* in order to explore the affordances of this technology to increase collaborative scholarly research and interpretation of the material.

Existing digital variorums have attempted to encourage collaboration and interaction by incorporating features into their web page interfaces. In the Electronic Variorum Edition of Don Quixote (EVE-DQ) from Texas A&M's Cervantes project, for example, users can add their own commentary and annotations which can then be accessed by future users. Similarly the Online Chopin Variorum Edition (OCVE) in progress at King's College London allows the attachment of notes to the base text with an option to share them publicly. For both projects, however, adding comments is an individual activity, with a single user at a workstation inputting notes with a mouse and keyboard. There is little room for group interaction, and any that exists must be sequential rather than communal. Most previous electronic New Variorum Shakespeare (eNVS) projects, particularly the recent *Winter's Tale* version, have no digital annotation function at all, limiting them to use as a reference tool, rather than an interactive platform.

Furthermore, it can be exceptionally difficult to allow simultaneous visual comparison of the numerous features of a variorum on these web page formats given the small size of most screens. A study by Wästlund, Norlander, and Archer on the effect of page layout on mental workload shows that "manipulating an onscreen text document via scrolling necessitates a shift of focus from the text to the action of controlling the page movement," (Wästlund, Norlander, and Archer, 1243) leading to decreased performance on reading comprehension tasks. Vandendorpe (2009) writes that "navigation by means of a mouse tends to give rise to chaotic, extremely rapid movement that is not very favorable to reading" (133).

By using a multi-touch table as a display device this project attempts to solve the problems of poor collaboration and broken comprehension. There has been significant research supporting the use of touch screens in improving reader focus, for example a study by Eva Siegenthaler et al. (2012) found that subjects performed better at various tasks involved in reading when using devices with tangible inputs, concluding that "a touch screen allows for an easier and more intuitive interaction" than a non-touch screen (Eva Siegenthaler et al. 2012, 94). The sizes of both the screen and the table perimeter of a multi-touch device, meanwhile, are conducive to multiple users working together. We thus believe that the application of multi-touch technology will have two effects. First, it represents another stage in the remediation of variorums, one that will better allow us to implement Unsworth's (2000) scholarly primitives: to sample, compare, discover, represent, annotate and reference different versions of the base text. Simultaneously, it encourages these tasks to be performed socially, deepening understanding by incorporating multiple viewpoints.

The collaborative uses of tangible devices in research situations is one of the main goals of this research. In driving towards this end, however, our group has faced issues in two camps: the ability of users to adapt to multiple people concurrently using touch controls; and, less expectedly, the ability of a designer to structure elements on an unconventional screen.

From the perspective of users, much of the challenge is about breaking habits. Since the introduction of personal computers, users have learned how to interact with the machine in individual work spaces. The adoption of this concept is so intense that one is liable to think that group work is less effective than work alone with the computer. However, it is noticeable that the technology has determined this situation: people cannot work together because the machine allows just one input at a time. The questions that we raise are: what happens when more than one person can interact with the machine? What kind of operations and collaborations can a group perform when all of them are engaged with the same machine in the same environment?

For designers, learning to think beyond the confines of a screen and mouse proved to be a major obstacle. The original design conceived of the mutli-touch screen as

being suspended on the wall, and the elements were placed accordingly (Figure 1). As we built and tested the interface it became clear that the layout would be effective for a single user, but was not ideal for the collaboration we were trying to achieve. While flipping the screen to a table was a logical choice to allow more users to work simultaneously, reassessing the design from that perspective resulted in a number of questions: where do you place touchstone elements when there is no clear top or bottom? How can items being used by different people overlap without disrupting anyone's work?

This paper explores the problems faced when building a social reading environment, both for users and for designers. The technology to allow such interaction exists: a multi-touch table has a size that allows for the display of multiple documents side-by-side, and its status as a touch screen enables easy and intuitive operation, lessening the mental workload required to operate it and permitting users to focus on the content rather than the interface. The system accommodates familiar gestures such as touch, pinch and flick to let the user move, select, grab and scroll through information on the screen, and since more than one point of interaction is possible, multiple people are able to work at the same time in the examination of the material, improving collaborative work. Designing for these features, however, is a mental challenge that requires an upset of standards in the minds of all those involved, users and builders, and facing these problems is a necessary step in moving towards the future of collaboration.

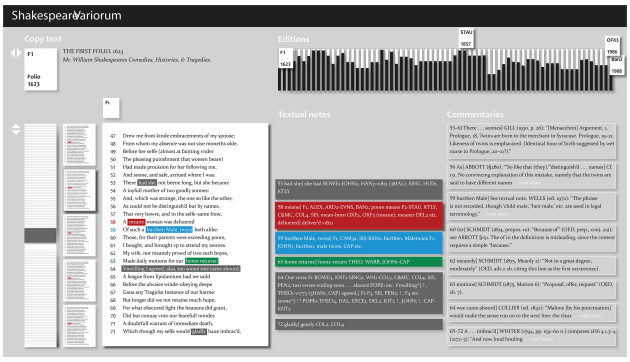


Figure 1 -
Initial view showing chosen copy text in a full view and a reading panel. All variants and commentaries on the chosen page are marked. More to the right there are views of all textual notes and commentaries connected to the chosen part of the copy text. At the top there is a representation of all editions in chronological order, in which the dark bars shows the overall difference of a particular edition to the chosen copy text.

References

Bradley, J., and P. Vetch (2006). Supporting Annotation as a Scholarly Tool–Experiences From the Online Chopin Variorum Edition. *Literary and Linguistic Computing* 22(2): 225–241.

Siegenthaler, E., et al. (2012). The Effects of Touch Screen Technology on the Usability of E-Reading Devices. *Journal of Usability Studies* 7(3): 94–104.

Furuta, R., et al. (2001). Towards an Electronic Variorum Dition of Don Quixote. *Proceedings of the 1st ACM/IEEE-CS Joint Conference on Digital Libraries*. 444–445.

Galey, A. (2005). 'Alms for Oblivion': Bringing an Electronic New Variorum Shakespeare to the Screen. *Shakespeare Association of America, Bermuda* http://pear.hcmc.uvic.ca:8080/ach/site/xhtml.xq?id=98

Hinckley, K., et al. (1997). Cooperative Bimanual Action. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 27–34.

Unsworth, J. (2000). Scholarly Primitives: What Methods Do Humanities Researchers Have in Common, and How Might Our Tools Reflect This? *Formal methods, experimental practice*. King's College, London. http://people.lis.illinois.edu/~unsworth/Kings.5-00/primitives.html

Vandendorpe, C. (2009).*From Papyrus to Hypertext: Toward the Universal Digital Library*. Aronoff, P. and S. Howard (trans.). Urbana: University of Illinois Press.

Wästlund, E., T. Norlander, and T. Archer (2008). The Effect of Page Layout on Mental Workload: A Dual-task Experiment. *Comput. Hum. Behav.* 24(3) : 1229–1245.

Werstine, P. (2011). Variorum Commentary. ArchBook: Architectures of the Book. http://inke.ischool.utoronto.ca/archbook/variorumcommentary.php

Wolff, M. (2007). Reading Potential: The Oulipo and the Meaning of Algorithms. 1.1 http://www.digitalhumanities.org/dhq/vol/1/1/000005/000005.html

6. Managing the Editorial Process: A Study of the Structured Surface Workflow Prototypes

Dobson, Teresa | Roeder, Geoff | Peña, Ernesto | Dergacheva, Elena | Brown, Susan | Heller, Brooke | INKE Research Group

Introduction

The INKE UX researchers have recently completed a user experience study of the Structured Surface Workflow prototype (see Frizzera et al. 2012). Although data analysis is ongoing, many interesting results are emerging. The majority of users, for example, imagined unanticipated ways of employing virtual workflow spaces for greater collaboration, openness, and re-humanization of the sometimes overly impersonal, email-based editorial process. This paper will discuss the findings of the study, and also suggest new ways to approach user experience studies within the digital humanities in order to maximize the use of time and participant resources. The results of and experience from the Workflow study findings can inform not only the development of future editorial prototypes but also the design and testing of prototypes in the Digital Humanities more generally by importing other well-established UX testing and research practices at earlier phases of design.

Study Background

Two closely related prototypes were tested simultaneously in the user experience study. The first, *Workflow Editorial Edition*, was designed to facilitate the digital document editing and publication process. Describing the goals of the program, Frizzera et al. (2012) remark that the design metaphor is:

> derived from the flowchart of activities that an editor can use to manage the movement of a submitted article or other item of text (such as a book review) through the stages from its initial appearance in the editor's inbox to the final step where it is ready to send to the printer. (n.pag.)

Indeed, the prototype has a flowchart-like appearance, reminiscent of index cards affixed to a corkboard (see Figure 1).

On this structured surface users position and move small circles, representing locating pins within the context of the flowchart metaphor and representing articles within the actual journal editing process. The use of familiar business and office metaphors is a well-established technique in interface design (e.g., Nadin 1988).

The second prototype employs the same "structured surface" metaphor but expands the possibilities for tracking and managing the progress of an article through the interface. *Orlando Workflow* is a customization of *Workflow Editorial Edition* for the document writing and markup process in Orlando: Women's Writing in the British Isles from the Beginnings to the Present (see Figure 2).

Integrating expanding pins with details, state, and history-tracking functions, the design of the Orlando Workflow relies less heavily on position-based information, privileging instead the information associated with each pin.

Procedures and Instruments

Participants were asked to: 1) complete a number of tasks in both the Workflow Editorial edition and Workflow Orlando edition prototypes; 2) provide "think-aloud" feedback while completing the task list items; and 3) complete a survey about the experience, one for each edition of the prototype. The study instruments included: a web-browser-based help page; a separate Likert-scaled experiential survey for each prototype; a demographics survey; and a task-list to establish common points of contact with the prototype across the participant population.

The need for more participants who had experience with the Orlando project was identified early on in the study. Without such experience, some participants were disoriented by the many acronyms and Orlando-specific terminology. The Vancouver research team developed an extension of the study to be run independently at University of Alberta (Edmonton) so that Orlando team members could be recruited as participants. With the support of the Orlando project, the Workflow user experience study successfully implemented the study extension using digital tools and communication only. The advantages and challenges of this collaborative user experience study will be briefly treated below.

Contributions to the Field of Knowledge

The preliminary findings suggest three effective directions for future designing and testing of editorial team management software. The first relates to the central focus of the Workflow prototype. Tracking the communication and human relationship dimensions of the editorial process was reported to be the greatest source of anxiety for editorial team leaders (viz., the prototypical user of the editorial Workflow). As one participant described it (narrating his thought process whenever he works on an editorial team):

> I know we talked about this with five different people but who are the other three? So I'm busy there hunting my email to track that stuff down. And I'm constantly in a small state of anxiety because I'm going, oh god, if I forget somebody - I feel I'm being incompetent or something. (Participant 6044)

The second is the actionable behaviors of the program: many users want to automate the actions of the prototype so that the possibility of human error is reduced. Indeed, participant 6044 (cited above) goes on to recommend that the interface be redesigned as a mediator among the

different communication and scheduling features already available on a computer, in order to help a user manage these small but anxiety-causing features of an editorial project.

Finally, certain common problems that many users shared with the prototype such as the size and arrangement of many of the elements suggests a way to improve present usability testing practices.

A Call for Earlier User Experience Testing

It is clear from our results that digital prototyping in early phases of development is an effective practice for interface design in the digital humanities. It is equally clear that other established practices in interface design could contribute to the resource-efficient generation and testing of a prototype at even earlier phases. For example, an earlier evaluation of the process would have been valuable in identifying some of the distracting issues before moving to a digital prototype. This would come after the phase of conceptual design but before the production of a digital prototype. Extended testing in initial phases with different low- and medium-fidelity prototypes in paper (Sefelin et al. 2003) or pdfs is known to exhibit potential and detectable usability issues. Comments from the users from the study indicate some of the issues they found distracting:

> Also those many, many, many bubbles are very confusing. If I keep having to click on them, that's going to irritate me. (Participant 6703)

> If I—let's say I was just taking a phone call and I come back and that bubble is sitting still in "Notify Author," does that mean it has or hasn't notified the author? (Participant 2751)

Low-fidelity prototype testing with potential users who have the right experiential background would identify these issues earlier in the design process. Those same users could be recruited for a full--scale user experience studies later, with the digital prototype in hand.

Another known effective path is to evaluate a very early stage of design based on heuristic principles alone (e.g., collecting expert opinions on the design principles; see Nielsen & Molich 1990; Nielsen 1994), particularly if it is not possible to count on real potential users experience at the beginning. Integrating both heuristic principles and low-fidelity prototyping could participate in early stages of design testing. Neither method requires specialized software development expertise. The implementation of this and other similar methods would contribute well to optimizing the cycle of prototype testing and improvement in the digital humanities.

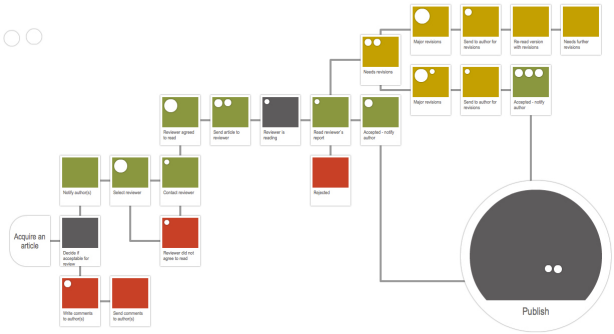


Figure 1 -
Workflow editorial edition.

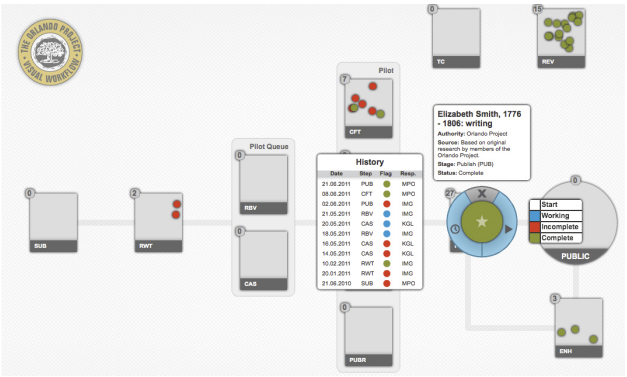


Figure 2 -
Workflow Orlando edition.

Hargreaves, D. M. G., and T. Robertson (2012). "Remote Participatory Prototyping Enabled by Emerging Social Technologies." *Proceedings of the 12th Participatory Design Conference: Exploratory Papers, Workshop Descriptions, Industry Cases - Volume 2*. New York, NY, USA: ACM. 25–28.

Nielsen, J. (1994). "Usability Inspection Methods." *Conference Companion on Human Factors in Computing Systems*. New York, NY, USA: ACM. 413–414.

Nielsen, J, and R. Molich (1990). "Heuristic Evaluation of User Interfaces." *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. New York, NY, USA: ACM. 249–256. 3 Nov. 2012. CHI '90.

Sefelin, R., M. Tscheligi, and V. Giller (2003). "Paper Prototyping - What Is It Good for?: a Comparison of Paper- and Computer-based Low-fidelity Prototyping." *CHI '03 Extended Abstracts on Human Factors in Computing Systems*. New York, NY, USA: ACM. 778–779. 3 Nov. 2012. CHI EA '03.

Circular Development: Neatline and the User/Developer Feedback Loop

Boggs, Jeremy
jeremy@clioweb.org
The University of Virginia

Earhart, Amy
aearhart@tamu.edu
Texas A&M University, United States of America

Graham, Wayne
wsg4w@virginia.edu
The University of Virginia

Kelly, T. Mills
tkelly7@gmail.com
George Mason University

McClure, David
dm4n@eservices.virginia.edu
The University of Virginia

Moore, Shawn
shawnw.moore@gmail.com
Texas A&M University, United States of America

Rochester, Eric
erochest@virginia.edu
The University of Virginia

Circular Development: Neatline and the User/Developer Feedback Loop

> Any tool should be useful in the expected way, but a truly great tool lends itself to uses you never expected.
>
> Eric Raymond, *The Cathedral and the Bazaar*

Eric Raymond's *Cathedral and the Bazaar* describes a series of lessons about the importance of sharing code with users learned during the development of Linux. Raymond emphasizes that non-technical users and third-party developers are capable of doing far more than just finding and fixing bugs—freely distributing code encourages users to take the software and develop new and unexpected things with it. "The next best thing to having good ideas," argues Raymond, "is recognizing good ideas from your users. Sometimes the latter is better." (2000) This kind of engagement with users tends to be the exception, not the rule, though, and even when developers are interested in establishing a cycle of feedback with users, that cycle has to be nurtured and maintained. Software development usually does not take place in a vacuum; software is developed for particular users and use cases. How people use software, and the ways in which they can share those uses, can be myriad. Developers are interested in learning how to recognize and nurture those uses, but this often proves difficult. Our panel will examine this complicated issue.

"Building" as a hermeneutic has gained increased attention and scrutiny among the Digital Humanities community. Ramsay (2011) argues that "the Digital Humanities is about building things" and is central to its "methodologization." Sample (2012) emphasizes the importance of building as work. In particular, Sample espouses *collaborative construction* as a group effort where each contribution takes place in dialogue with other contributions, and *creative analysis* as a way to learn through creation. An emphasis on building necessitates an equal emphasis on *builder*, and as Gina Trampani (2011) argues, nurturing a beneficial user-contributor community that allows a variety of users, regardless of existing skills, to benefit from a hermeneutic of building. Accordingly, we are interested in modeling the communal approaches to building that bridge developer, researcher, and student.

This panel will bring together developers and users to explore the symbiotic relationships built during the life cycle of a software project, to discuss the ways in which open-source Digital Humanities projects should work to build both tools *and* user/developer communities. The project that we are using as a testbed for this examination is Neatline, a set of a geo-temporal tools built by the Scholars' Lab at the University of Virginia for use in the Omeka content management system. During the months leading up to the conference, panel participants will work closely together to build and document their working relationships, all the while working to improve Neatline and implement it in productive ways. The panel will elaborate on problems and solutions for collaboration among developers and users they encountered, and suggest ways to turn users into contributors while better attuning a software development team to the needs of its users. Of particular attention to the panel will be the way in which the tool is used for multiple

purposes, including research and teaching, and how such uses impact the feedback loop.

Panel Organization & Participants

We propose to conduct a panel featuring users and developers of the Neatline suite. Each participant will open the panel with a 5 minute statement describing their particular experience over the course of their collaboration, followed by a group discussion that addresses several questions. All participants are excited to participate in this panel.

Participants

Jeremy Boggs is Design Architect for Digital Research and Scholarship at the University of Virginia Library. Boggs will discuss methods for getting outside users more easily involved in the development process for Neatline. He will focus on the tools used and documentation developed during the group's collaboration effort.

Amy Earhart is Assistant Professor of English at Texas A&M University. Earhart will discuss how her undergraduate students used Neatline to map Malcolm X's New York, pointing to particular areas of tension between pedagogical models of digital humanities tools and the feedback loop. She will offer potential ways to eliminate such issues.

Wayne Graham is Head of Research and Development for Digital Research and Scholarship at the University of Virginia Library. Graham will discuss the day-to-day management of Neatline development, and in particular his strategies for balancing user needs and contributions with the priorities of the core Neatline development team.

T. Mills Kelly is Associate Professor of History at George Mason University and Associate Director of the Roy Rosenzweig Center for History and New Media. Kelly will discuss the use of Neatline in his historical methods course, "Dead in Virginia." He will focus on the aspects of the user experience that seemed to influence student learning in the course.

David McClure is Web Applications Developer for Digital Research and Scholarship at the University of Virginia Library, and is lead developer on Neatline. McClure will talk about his perspective as a lead developer on Neatline.

Shawn Moore is a doctoral student at Texas A&M University and is a fellow for the Initiative for Digital Humanities, Media, and Culture (IDHMC). Moore will talk about the process of transitioning from a user of Neatline to a contributing developer during his ongoing dissertation project on Margaret Cavendish, Duchess of Newcastle (1623-1673).

Eric Rochester is Senior Developer for Digital Research and Scholarship at the University of Virginia Library. Rochester will discuss the tenuous balance between choosing the best tools and languages for a project with getting and encouraging outside contributions to a project.

Questions

- What benefits will a software feedback loop provide to both user and developer?
- What tools and methods did the group find most helpful during the process?
- Discuss the impact of non-specialist users, such as students, on the feedback loop.
- How can open-source projects create inclusive communities that invite contributions from people with skill-sets and backgrounds that are underrepresented in the open-source community?
- In what ways does nurturing an outside user/developer community contribute to the use and sustainability of a Digital Humanities project?
- What were the most challenging aspects of this collaboration?
- Discuss future models of the feedback loop based on what you have learned in this model.

References

Bryant, T. (2006). Social software in academia. *Educause Quarterly* 29(2): 61.

Clement, T., and D. Reside (2011). Off the Tracks: Laying New Lines for Digital Humanities Scholars. Results of an NEH Workshop, Maryland Institute for Technology in the Humanities. http://mith.umd.edu/offthetracks/ .

Cohen, D. J. (2008). Creating Scholarly Tools and Resources for the Digital Ecosystem: Building Connections in the Zotero Project. *First Monday* 13(8) http://firstmonday.org/htbin/cgiwrap/bin/ojs/index.php/fm/article/view/2233/2017 .

Easley, D., and J. Kleinberg (2010). *Networks, Crowds, and Markets: Reasoning about a Highly Connected World*. http://www.cs.cornell.edu/home/kleinber/networks-book/networks-book

Fogel, K. (2005). *Producing Open Source Software: How to Run a Successful Free Software Project*. Sebastopol, CA: O'Reilly.

Klein, L. F. (2012). A Report Has Come Here. http://lmc.gatech.edu/~lklein7/?p=86 .

McPherson, T. (2010). Scaling Vectors: Thoughts on the Future of Scholarly Communication. *Journal of Electronic Publishing* 13(2) http://hdl.handle.net/2027/spo.3336451.0013.208 .

Neatline. http://neatline.org

Nowviskie, B. (ed.) (2011). *#Alt-Academy: Alternative Academic Careers for Humanities Scholars*. MediaCommons Press. http://mediacommons.futureofthebook.org/alt-ac/

Omeka . http://omeka.org.

Perspectives on Free and Open Source Software . (2010). Cambridge: The MIT Press.

Ramsay, S. (2011). On Building. http://stephenramsay.us/text/2011/01/11/on-building.html (accessed 11 January 2011).

Ramsay, S.. (2011). Who's In and Who's Out? http://stephenramsay.us/text/2011/01/08/whos-in-and-whos-out.html (accessed 8 January 2011).

Ramsay, S.. (2012). Programming with Humanists: Reflections on Raising an Army of Hacker-Scholars in the Digital Humanities. In Hirsch, B. D. (ed.), *Teaching Digital Humanities: Principles, Practices, Politics*. Ann Arbor: University of Michigan Press.

Ramsay, S., and G. Rockwell (2012). Developing Things: Notes Toward an Epistemology of Building in the Digital Humanities In Gold, M. (ed.), *Debates in the Digital Humanities*. Minneapolis: University of Minnesota Press. 75-84.

Raymond, E. S. (2000). *The Cathedral and the Bazaar*. Sebastopol, CA: O'Reilly.

Sample, M. (2012). Building and Sharing (When You're Supposed to be Teaching). *Journal of Digital Humanities*. 1(1) http://journalofdigitalhumanities.org/1-1/building-and-sharing-when-youre-supposed-to-be-teaching-by-mark-sample/ .

Trapani, G. Designers, Women, and Hostility in Open Source. http://smarterware.org/7550/designers-women-and-hostility-in-open-source (accessed 23 March 2011).

The Future of Undergraduate Digital Humanities

Croxall, Brian
brian.croxall@emory.edu
Emory University, United States of America

Singer, Kate
ksinger@mtholyoke.edu
Mount Holyoke College, United States of America

Ball, Cheryl E.
cball@ilstu.edu
Illinois State University, United States of America

Cordell, Ryan
rccordell@gmail.com
Northeastern University, United States of America

Davis, Rebecca Frost
rdavis@nitle.org
National Institute for Technology in Liberal Education (NITLE)

McDonald, Jarom
jarom_mcdonald@byu.edu
Brigham Young University, United States of America

Posner, Miriam
miriam.posner@gmail.com
University of California, Los Angeles, United States of America

Theibault, John
John.Theibault@stockton.edu
Richard Stockton College, United States of America

The Future of Undergraduate Digital Humanities

Alongside the increasing number of digital humanities job listings, postdoctoral fellowships, and graduate programs, we have begun to see a number of introductory digital humanities courses and the creation of several programs at a wide range of undergraduate institutions — everything from small liberal arts colleges to state universities to research-intensive private institutions. Consequently, it seems opportune to think closely about how the digital humanities will shape undergraduate education — and vice versa. New jobs and fellowships presuppose undergraduates who have been and will be introduced to conversations of the digital humanities as well as humanities faculty who will teach them. While the late 1990s sparked discussions among Canadian and US institutions about the creation of undergraduate and master's programs (Rockwell 1999; Unsworth 2001; Sinclair and Gouglas 2002), new technologies and institutional interest

have renewed the conversation (Spiro 2010; Fitzpatrick 2010; Brier 2012; Reid 2012; Davis and Alexander 2012). Because such a large number and range of institutions are now considering implementing some training in the digital humanities, it now seems timely to contemplate the future of undergraduate digital humanities.

This panel considers how we might recalibrate the digital's role in the humanities by making undergraduate education—and not simply digital pedagogy — a more central preoccupation. Building on recent, compelling discussions of infrastructure and curriculum for digital humanities graduate programs (Clement 2010; Thaller et al. 2012; Boggs et al. 2012) as well as roundtables on alternative careers (Nowviskie et al. 2011), dynamic constellations for undergraduate education are emerging from the interactions among new computational methods, hybrid classroom spaces, reimagined curricula, and alternative career paths for college graduates. This panel gathers several initiators of such digital humanities programs for undergraduates to discuss their past and future.

More than simply creating students to enroll in new graduate programs, introducing the methods of the digital humanities to undergraduates provides opportunities for them to do something traditionally reserved for students in the sciences: original, collaborative research (Blackwell and Martin 2009; Norcia 2008). Moreover, digital humanities has arguably brought renewed attention to discussions of praxis and pedagogy, with online journals such as *Hybrid Pedagogy* and *The Journal of Interactive Pedagogy and Technology*; Brett D. Hirsch's recent *Digital Humanities Pedagogy: Practices, Principles, Politics* (2012); multiple panels on digital pedagogy at the 2012 MLA (Harris 2012; Berens and Croxall 2012) and a digital pedagogy unconference at the 2013 MLA (Croxall and Koh 2013); Brown University's "Teaching with TEI" seminar (2012); a dedicated track at recent Digital Humanities Summer Institutes (Harris, Sayers, and Jakacki 2012; Jakacki 2013); and several poster presentations at recent Digital Humanities Conferences (Bonsignore et al., 2011; Harris 2011; Singer 2012; Croxall 2012). Our hope is that a roundtable discussion, drawing on participants from different fields and representing many different types of U.S. institutions, will help, first, to identify some of the best contemporary approaches to undergraduate digital humanities curricula, infrastructure, course scaffolding, and praxis and, second, to sketch out new directions for the future of undergraduate education at a variety of undergraduate institutions.

Organization

Each speaker will talk for 7 minutes about a particular institutional praxis or curricular infrastructure. The organizers will then pose questions for the entire roundtable for 20 minutes, leaving the remainder of the conversation for discussion among panelists and the audience.

To begin with, panelists will map out the multiple and competing histories of digital humanities' recent incursion into undergraduate education, just as Matthew Kirschenbaum (2010) and others have sought to understand digital humanities by reflecting on its institutional histories. Part of this multiplicity, of course, is due to the ways in which "digital humanities" is understood differently at each institution, due to the specific interests of individual scholars, the focus of particular departments, and the demands of institutional mission. With these issues in mind, panelists will present several different models for integrating digital humanities into undergraduate coursework: from introductory seminars for first-year students who may lack technological skills through advanced courses for majors to specialized, independent research projects as capstone experiences. In doing so, we will consider both how best to structure something like an "Introduction to Digital Humanities" course and how to connect disparate projects, faculty, and upper-level courses. Such scaffolding naturally begs the larger question of whether digital humanities is best introduced to undergraduates as a separate discipline or as a crucial part of traditional humanities courses. Finally, while all panelists generally agree on a praxis-based approach to such courses, they will also discuss how best to execute and theorize praxis in the different disciplines.

Questions that panel organizers might pose during the subsequent discussion include:

- Is digital humanities a topic that should be based within particular departments? Or is it something that should be taught across all humanities undergraduate departments?
- What departmental or university infrastructure and support are necessary for a digital humanities undergraduate curriculum?
- What is necessary to prepare students for digital humanities work at the graduate level? Is adequate preparation possible without more formalized graduate programs in place?
- How do we redesign curricula to incorporate both DH courses and incursions into traditional disciplines?
- Is digital humanities a methodology or a topic of study? How can the two approaches be best integrated in the undergraduate classroom?
- What are best practices for praxis methodologies and project-based research approaches in the undergraduate classroom?
- How might we envision curricula to be redesigned in the future with digital tools and digital critical thinking in mind?

- How might national and international conceptions of undergraduate education shape digital humanities incursions differently?

Speakers

Recognizing the importance of undergraduate education in the future of the digital humanities, all six speakers have enthusiastically committed to attend and present at DH in Lincoln.

- Cheryl E. Ball, Associate Professor of English, Illinois State University
  Ball highlights how digital writing studies (a discipline in its fourth decade that integrates digital technology into its writing pedagogy and research) has always focused on issues current to discussions of "making" in the digital humanities: collaboration, openness, multimodality, and peer-review. Ball argues that a digital publishing curriculum, in which undergraduate students theorize and produce texts meant for an audience outside of the classroom (a key concept to digital writing studies pedagogy), is a model for DH in how it bridges theory and praxis across multiple disciplines in the humanities.
- Ryan Cordell, Assistant Professor of English, Northeastern University
  Cordell writes frequently about technology and teaching for the ProfHacker blog at the Chronicle of Higher Education and has taught digital humanities-inflected courses at both a liberal arts college and a research university. He will draw on those experiences in his contribution to this panel, where he will contend that undergraduates do not share their instructors' fascination with defining or theorizing digital humanities qua digital humanities. Rather than dwelling on such debates, he will suggest that DH instructors should embrace undergraduate disinterest in DH as an aid to curricular incursion, allowing digital practices to be introduced as routine aspects of scholarly practice.
- Rebecca Frost Davis, Program Officer for the Humanities, National Institute for Technology in Liberal Education (NITLE)
  Davis will discuss results of NITLE's 2012 Survey of Digital Humanities at Liberal Arts Colleges, institutions that largely integrate digital methodologies via disciplinary coursework and student scholarship, rather than as a separate academic program. Her research explores the motivations and mechanisms for creating, integrating, and sustaining digital humanities within and across the undergraduate curriculum.
- Jarom McDonald, Associate Research Professor and Director, Office of Digital Humanities, Brigham Young University
  McDonald will address the topic, "Considering a Moneyball approach to Digital Humanities Education." BYU has just finished a multi-year assessment project of their long-running Computers and the Humanities minor, gathering empirical data through surveys, collaborative faculty input sessions, student tracking, and external review. He will discuss how he and colleagues are now working to understand how to best implement the wealth of evidence they've collected to help their program evolve for current and future students' needs.
- Miriam Posner, Digital Humanities Coordinator and Research Technology Consultant, UCLA
  Posner, who both coordinates and teaches in UCLA's Digital Humanities program, is helping build a new interdisciplinary minor in digital humanities. She will speak on "knowledge design," a pedagogical approach she and her colleagues have adopted that emphasizes an environment of project-based collaboration. Drawing on theories advanced by Johanna Drucker and Jerome McGann, she will describe the program's studio model in which students are assigned a novel problem and asked to work across disciplines and hierarchies to solve it together.
- John Theibault, Director, South Jersey Center for Digital Humanities, Richard Stockton College of New Jersey
  Theibault began his academic career as a historian of early modern Europe and is currently Director of the South Jersey Center for Digital Humanities at Stockton College. He started teaching "Introduction to Digital Humanities" to undergraduates in the General Studies program at Stockton in 2011, which prompted his reflections about where such a course fits within a broader digital humanities curriculum for undergraduates, the topic of his presentation.

Organizers

- Dr. Brian Croxall, Digital Humanities Strategist & Lecturer of English, Emory University
- Dr. Kate Singer, Assistant Professor of English, Mount Holyoke College

References

Alexander, B., and R. F. Davis (2012). *Should Liberal Arts Campuses Do Digital Humanities? Process and*

Products in the Small College World. Debates in the Digital Humanities. Minnesota University Press. 368-389.

Berens, K. I., and B. Croxall (2011). *Session Proposal. Building Digital Humanities in the Undergraduate Classroom.* 28 Nov. 2011. Web. http://www.briancroxall.net/buildingDH/2011/11/28/session-proposal/ . 17 Oct. 2012.

Blackwell, C., and T. R. Martin (2009). *Technology, Collaboration, and Undergraduate Research. Digital Humanities Quarterly*. 3(1). http://www.digitalhumanities.org/dhq/vol/3/1/index.html . 17 Oct 2012.

Boggs, J., et al. (2012). *Realigning Digital Humanities Training: The Praxis Program at the Scholars' Lab. Digital Humanities 2012.* University of Hamburg. 18 July 2012. Poster presentation.

Bonsignore, B., et al. (2011). *The Arcane Gallery of Gadgetry: A Design Case Study of an Alternate Reality Game. Digital Humanities 2011.* Stanford University. 21 June 2011. Poster presentation.

Brier, S. (2012). *Where's the Pedagogy? The Role of Teaching and Learning in the Digital Humanities.* **Debates in Digital Humanities**. Minnesota UP: 350-367.

Clement, T. (2010). *An Undergraduate Perspective. Digital Literacy for the Dumbest Generation. Digital Humanities 2010.* King's College London. 8 July 2010.

Croxall, B. (2012). *Courting 'The World's Wife': Original Digital Humanities Research in the Undergraduate Classroom. Digital Humanities 2012.* University of Hamburg. 18 July 2012.

Croxall, B., and A. Koh. (2013). *A Digital Pedagogy Unconference*. Modern Language Association Convention. Boston. 3 January 2013. http://www.briancroxall.net/digitalpedagogy/ . 17 Oct. 2012.

Fitzpatrick, K. (2010). Undergrads Reimagine the Humanities. *Planned Obsolescence*. 11 December 2010. http://www.plannedobsolescence.net/blog/undergrads-reimagine-the- humanities/ . 25 February 2013.

Harris, K. D. (2011). Pedagogy & Play: Revising Learning through Digital Humanities. Digital Humanities 2011. Stanford University. 21 June 2011. Poster presentation.

Harris, K. D. (2011). Acceptance of Pedagogy & DH MLA 2012. *triproftri*. 14 May 2011. http://triproftri.wordpress.com/2011/05/14/acceptance-of-pedagogy-dh-mla-2012/ . Web. 17 Oct. 2012.

Harris, K. D., J. Sayers, and D. Jakacki. (2011) Digital Pedagogy in the Humanities. *Digital Humanities Summer Institute*. University of Victoria. June 2011 and June 2012.

Hirsch, B. D. (2012). *Digital Humanities Pedagogy: Practices, Principles and Politics*. Open Book Publishers, 2012.

Jacacki, D. (2013). Digital Pedagogy in the Humanities. *Digital Humanities Summer Institute*. University of Victoria. June 2013.

Kirschenbaum, M. G. (2010). What is Digital Humanities and What's it Doing in English Departments? *ADE Bulletin* 150: 1-7.

Norcia, M. (2008). Out of the Ivory Tower Endlessly Rocking: Collaborating across Disciplines and Professions to Promote Student Learning in the Digital Archive. *Pedagogy* 8(1): 91-114.

Nowviskie, B., et al. (2011). The "#alt-ac" Track: Digital Humanists off the Straight and Narrow Path to Tenure. *Digital Humanities 2011*. Stanford University. 22 June 2011.

Reid, A. (2012). Graduate Education and the Ethics of the Digital Humanities. *Debates in Digital Humanities*. Minnesota UP, 390-401.

Rockwell, G. (1999). *Is Humanities Computing an Academic Discipline?* http://www.iath.virginia.edu/hcs/rockwell.html . 25 February 2013.

Sinclair, S., and S. W. Gouglas (2002). Theory into Practice: A Case Study of the Humanities Computing Master of Arts Programme at the University of Alberta. *Arts and Humanities in Higher Education* 1.2: 167-183.

Singer, K. (2012). The Melesina Trench Project: Markup Vocabularies, Poetics, and Undergraduate Pedagogy. Digital Humanities 2012. University of Hamburg. 18 July 2012.

Spiro, L. (2010). Opening Up Digital Humanities Education. *Digital Scholarship in the Humanities*. 8 September 2010. http://digitalscholarship.wordpress.com/2010/09/08/opening-up-digital-humanities-education/ . 25 February 2013.

Thaller, M. et al. (2012). Digital Humanities as a University Degree: The Status Quo and Beyond. Digital Humanities 2012. University of Hamburg. 18 July 2012.

Unsworth, J. (2001). *A Master's Degree in Digital Humanities: Part of the Media Studies Program at the University of Virginia*. 25 May 2001. http://people.lis.illinois.edu/~unsworth/laval.html . 25 February 2013.

Women Writers Project. (2012). Taking TEI Further: Teaching with TEI. Brown University Seminar. 20-22 August 2012.

Issues in Spatio-Temporal Technologies for the Humanities and Arts

Eide, Øyvind
oyvind.eide@edd.uio.no
University of Oslo, Norway

Grossner, Karl
karlg@stanford.edu
Stanford University, USA

Berman, Merrick Lex
mberman@fas.harvard.edu
Harvard University, USA

Ore, Christian-Emil
c.e.s.ore@iln.uio.no
University of Oslo, Norway

Issues in Spatio-Temporal Technologies for the Humanities and Arts

Introduction

Spatio-temporal concepts are so ubiquitous that it is easy for us to forget they are essential to everything we do. All expressions of human culture are related to the dimensions of space and time in the manner of their production and consumption, the nature of their medium and the way in which they express these concepts themselves. The Space/Time Working Group (STWG) of the NeDiMAH[1] network held a full day workshop at the DH2012 conference on the topic of theorising methods that exploit space and time in the Digital Humanities. This session proposal is intended to continue and contribute to that discussion.

Issues

Connectivity. Different types of media provide different affordances (Gibson 1986) for representing space and time. We need more work on connecting space and time as represented in texts, audio, and video with the representations being created in GIS systems and other spatiotemporal databases. Texts, sound, and (moving) images are not simply "media" to be spatiotemporally tagged, but may have narrative structures that represent alternative models of space and time (Jewell 2005).

Uncertainty. There has been a lot of emphasis on ambiguity and vagueness in humanist theorising. Our digital tools depend upon our ability to demarcate boundaries, but do the demarcations allow room for the recognition of vagueness, ambiguity, and uncertainty? Should we abandon borders altogether (Berman 2005)? Places and periods are vague, socially defined constructs and source data always leads to imprecise and/or inaccurate data. Can we find a way of encapsulating ambiguity and uncertainty in metadata itself? How can we model for 'vectors of intensity'—impact—thinking about what we really want to do with what time and space tell us?

Contrasts. Although space and time are closely related, there are significant differences between the two. Bakhtin's concept of the chronotope (Bakhtin 1981) implies a strong connection between time and space, but this is a connection between quite different things. Among the most important differences are the natures of their dimensionality (three dimensions vs. one), their different relationships to the static—dynamic continuum where space is static but changeable, whereas time is a "flow" but the past is unchangeable, and the different methods we use to make the communicative leap across spatial and temporal distance.

Representation. Every medium, whether textual, tactile, illustrative or audible, or some combination of them, exploits space and time differently in order to convey its message. "Every medium has the capacity of mediating only certain aspects of the total reality" (Elleström, 2010, 24). The changes required to express the same concepts in different media are often driven by different spatio-temporal requirements. Authors and artists must decide how to collapse reality into the spatio-temporal limitations of a chosen medium, and the nature of those choices can be as interesting as the expression itself.

Absolute vs. Relative. How do we model for movement, trajectory, fluidity and momentum of events and ideas? How do we allow references to float in time and space? Relevant and foundational works exist, particularly for temporal models (e.g. Doerr and Yiortsou 1998; Plewe 2002, 2003; Grossner 2010), but the DH community would benefit greatly by a targeted effort for continued tool development in this area. Holmen and Ore (2010) suggest a way to handle uncertain time in a conceptual documentation system for archaeology. It is, however, not trivial to extend their model to include space as well.

Technical Literacy. In an area combining old with emerging conventions, how do humanists learn when to read a complex visualisation 'with a grain of salt' and to distinguish the 'truthiness' of something that appears on a screen from the complex process of selecting the sources that underlie it? How can humanists learn to justify and critique tool choice in the same way they justify and critique their selection of sources? The development of tools has to be based on informed cooperation, where the representatives of each discipline are allowed to work together on an equal footing.

Theory vs. Practice. The representational issues discussed above also connect to wider questions. The desire to connect digital tool building with the theoretical discourse of the humanities is often expressed, but it is not clear how to do it or what the utility of this will be. Tool building has its own theories, expressed in the form of encoding schemes, data structures, and ontologies. How can we bring the last few years' discussions in this area forward? This is more than theory-based practice; it also includes practice as a source of theory on many levels, including the experiment as a theory-testing device as well as experiments creating theoretical ideas, also in serendipitous ways.

Presentations

The four panellists will each make a ten minute introduction, discussing aspects of how spatio-temporal datatypes have been, and can be modelled in relational databases. This includes associated operators and algorithms to enable computation of probabilistic or "fuzzy" extents, in the context of specific cases faced when dealing with their own spatial / temporal data and related materials. The nature of the choices that have to be made in order to represent aspects of the spatiotemporal reality in media expressions is a key issue. The focus of each of the presentations will be:

1. The first presentation will be an introduction to the panel where the problems of media translations, especially between texts and maps, will be highlighted by pointing to specific examples. Fiction as well and non-fiction texts will be shown to include spatial description that makes a translation into a map impossible. However, if one sees each map as a possible interpretation of the text, then the complete text-map expression can be seen as a richer expression than either a text or a map alone.
2. Dealing with dates, time periods, temporal granularity, data formats and asynchronous changes is one of the main issues that needs to be addressed in space-time data models. In the second presentation, we will provide some concrete examples from Chinese history of how using specific dates, qualitative typing of dates, and named time periods can be modelled and queried for spatial objects that change over time. Examples of asynchronous changes in GIS will demonstrate the differences between databases set up for time slices, time series, and temporal networks.
3. The third presentation will discuss data modelling challenges encountered in representing temporal, spatial and thematic dimensions of the lives of ~28,000 Britons, spanning a period of several hundred years—all related by birth or marriage. Lifespans are in many cases bounded by vague or uncertain dates, and geographic associations have varying granularity. The goal is a meaningful "contemporary-of" relation joining the problematic temporal and spatial data with tags for individuals' professions, aggregated to activity spheres.
4. The concluding presentation will put the examples into the perspective of ontological modelling in culture heritage. Using time as an example, it will be demonstrated how a tool implementing modelling principles from CIDOC-CRM and Allen operators (Allen 1983). The tool can infer conflicting dating, increase precision of starts, ends and durations of events and finally display a chronological overview from a given a dataset of events, their time-spans and relations between events. The system can display all possible chronologies for the events in the set thus adding information to a combined data set.

After these four presentations, we will then invite the audience to a discussion, before we conclude by suggesting a list of important issues for further research. This list will be included in a panel report that will be written by the panellists and published digitally for further discussions, online as well as at physical meetings. Through the practical examples we hope to catalyse future workshops where new methods (such as topologies or qualitative descriptors) can be applied in practice, under the umbrella of the Nedimah Space/Time Working Group. Interested participants will be invited to take part in this continued work together with the panellists.

References

Allen, J. F. (1983). Maintaining Knowledge About Temporal Intervals. *Commun. ACM* 26(11): 832–43.

Bakhtin, M. M. (1981). Forms of Time and the Chronotope in the Novel. Notes toward a Historical Poetics. In Holquist, M. (ed.), *The dialogic imagination: four essays*. 84–258. Austin: University of Texas Press.

Berman, M. L. (2005). Boundaries or Networks in Historical GIS: Concepts of Measuring Space and Administrative Geography in Chinese History. *Historical Geography* 33: 118–133.

Doerr, M., and A. Yiortsou (1998). Implementing a Temporal Datatype, *Technical Report ICS-FORTH/TR-236*. url: www.ics.forth.gr/isl/publications/paperlink/implementing_a_temporal_datatype.ps.gz (checked 2009-05-27)

Elleström, L. (2010). The Modalities of Media: A Model for Understanding Intermedial Relations. In L. Elleström (Ed.), *Media borders, multimodality and intermediality*. 11–48. Basingstoke: Palgrave McMillan.

Gibson, J. J. (1986). *The ecological approach to visual perception*. Hillsdale, N.J.: Lawrence Erlbaum. 1st ed.: 1979.

Grossner, K. (2010). Representing Historical Knowledge in Geographic Information Systems. Ph.D. dissertation, University of California, Santa Barbara, United States -- California. URL: http://www.kgeographer.org/assets/Grossner_dissertation_2010.pdf (checked 2012-11-03)

Holmen, J., and C.-E. Ore (2010). Deducing event chronology in a cultural heritage documentation system. In J. W. Crawford and D. Koller (Eds.), Making History Interactive. *Computer Applications and Quantitative Methods in Archaeology (CAA). Proceedings of the 37th International Conference*. Oxford: Archaeopress.

Jewell, M. O., K. F. Lawrence, and M. M. Tuffield (2005). OntoMedia: An Ontology for the Representation of Heterogeneous Media. In, *Multimedia Information Retrieval Workshop (MMIR 2005) SIGIR*, ACM SIGIR.

Plewe, B. S. (2002). The nature of uncertainty in historical geographic information. *Transactions in GIS*, 6(4): 431-456.

Plewe, B. S. (2003). Representing datum-level uncertainty in historical GIS. *Cartography and Geographic Information Science*, 30(4): 319-334.

Notes

1. Network for Digital Methods in the Arts and Humanities (http://nedimah.eu). NeDiMAH is funded by the European Science Foundation.

Excavating Feminisms: Digital Humanities and Feminist Scholarship

Harris, Katherine D.
katherine.harris@sjsu.edu
Department of English & Comparative Literature, San Jose State University

Wernimont, Jacqueline
jwernimo@scrippscollege.edu
Department of English, Scripps College

Inman Berens, Kathi
kathiberens@gmail.com
Annenberg School for Communication, University of Southern California

Grigar, Dene
dgrigar@mac.com
Creative Media & Digital Culture Program, Washington State University Vancouver

Session Abstract

In a field that sometimes organizes itself around the credo “more hack, less yack,” the role of theory and critical reflection upon race, sexuality, gender and and other discursive identity categories might seem a subordinate concern. But recent calls by Alan Liu , Laura Mandell, Tara McPherson, Adeline Koh, and Jamie Skye Bianco, among others, prompt this panel’s speakers to take up Liu’s challenge to “extend the issues involved” in building and making “into the register of society”:

> We digital humanists develop tools, data, metadata, and archives critically; and we have also developed critical positions on the nature of such resources.... But rarely do we extend the issues involved into the register of society, economics, politics, or culture in the vintage manner, for instance, of the Computer Professionals for Social Responsibility (CPSR). How the digital humanities advance, channel, or resist the great postindustrial, neoliberal, corporatist, and globalist flows of information-cum-capital, for instance, is a question rarely heard in the digital humanities associations, conferences, journals, and projects with which I am familiar.

Liu *et al.* have helped us raise a set of difficult questions: Has feminism slowly been leached out of some DH projects as those projects have become successful? Is a “revolutionary” pedagogy possible if the objects we train our students to make are welcomed by administrators and other higher ed stakeholders as precisely the “information-cum-capital” that Liu warns us against? How does one woman artist’s elision from a field-defining reception history reveal the apparati by which careers are made or broken? Can structuring tools like XML or xHTML or PHP be deployed as a feminist intervention, and how would that differ from other deployments? Do the institutional and funding structures of the field constrain our ability to make ethical interventions and encourage work that turns away from the kinds of social engagement that Liu describes?

The 1970s feminist slogan “the personal is political” taught people to break down systematic sexism into molecules that could at any moment cluster into atoms and catalyze real power. While recognition of the politics of

everyday life and the value of quotidian genres of cultural production has clearly influenced certain kinds of recovery projects, moves within Digital Humanities to argue the necessity of advanced programming and development skills fails to recognize the ways in which structural elements of American education, family life, and work remain deeply gendered. Further, advocacy of a "bootstrap" or "DIY" ethos risks perpetuating a dangerous suggestion that historically disenfranchised groups should make "do with what you can scrounge" in order to earn their way into spaces of greater privilege, and fails to recognize the important role that networks of support, mentoring, and technology access play in DH work. While exhortations to "do it yourself" are often made in the spirit of rebellious assertions of power and independence, access to the practices behind these memes is culturally and economically bounded.

These issues have surfaced with striking force in the last couple of years. Miriam Posner's provocative post from March 2012, " Some things to think about before you exhort everybody to code ," prompted 59 deeply engaged and passionate responses. Her thesis:

> The point is, women aren't [learning to code]. And neither, for that matter, are people of color. And unless you believe (and you don't, do you?) that some biological explanation prevents us from excelling at programming, then you must see that there is a structural problem.

DH practitioners have gathered into a community founded upon sharing and collaboration, a workflow ethos traceable to second-wave feminism. Several of the major Digital Humanities projects that are now at the forefront of the field and were driven originally by feminist imperatives (WWP, Orlando, Dickenson E Archive); but as these projects developed, their relationship to the communities they were built to serve attenuated at the same time that new types of DH practice, like "big" data mining, have sought to address different agendas. These shifts have left some despairing of the ethics and politics of DH. "I'm a teacher," says Stephen Ramsay, in a comment on Posner's "Code" post:

> I care about students who want to learn, learning. I'm not so naive as to think that we can reform that culture from without, but honestly, if we just re-duplicate that culture [of sexism] in DH, then we have failed. And we might as well go back to whatever we were doing before.

DHers would not naively posit themselves as uninfluenced by such hegemonic incursions; some of us are particularly mindful of them because we build projects that critique and resist them. But speakers on this panel submit that a tolerance for "yack"— a word that simultaneously conjures ladies chattering, not working, but that also signals attention to discursive and social relations— might be an initial intervention in the terms by which we construe the messy work of "excavating" feminist "art-&-facts" from DH's rich silt: archives (Katherine D. Harris and Jacque Wernimont), encoding methodologies (Jacque Wernimont), literary reception (Kathi Inman Berens) and pedagogy (Dene Grigar). While we each canvass a discrete topic, we see significant overlap in terms of the means, histories, and technologies of digital production and teaching. Our roundtable will thus be structured by four short position statements, as outlined below, and then will engage all participants in the room to mount a community discussion exploring how and why the histories of cyberfeminism and feminist digital production matter right now as DH becomes "The Hot Thing." Even as we "excavate," we look forward to building anew: what are the salient lessons to be gleaned from the presenters' statements? How do they, and statements and observations from others in the room, suggest new or continuing avenues of work? Two of our panelists (Grigar and Wernimont) run their own labs. Is the material investment in women leading labs and programs an essential intervention in how privilege is disbursed? Or is it just essentialist?

The wide range of reviewers' responses to our panel suggests a wish for us to address simultaneously an untenable range of feminisms: to be both grounded in material practice, but also theoretically expansive to address "Occupy" and other 21st-century feminisms. We suggest that the reviewers' collective wish for feminist critique to be both united and expansive in its approach freights it with responsibility beyond the scope of our actual claims.

We appreciate that we can do more to unite the panel and will begin with a foundational grounding that includes of the trajectory of 21st century cyberfeminism that stretches back to Donna Haraway and bell hooks, and moves forward to practical and popular appropriations of Haraway's theories twenty-five years later by Douglas Rushkoff and Howard Rheingold. We will then turn to the short papers arising from our particular practices and areas of expertise. Rather than seeing this as a weakness, we consider this diversity an important feature of our scholarship — a necessary link between our material practice, historical positions, and theoretical interventions. Additionally, each paper engages with a history of appropriation of technologies, practices, and ideas. The papers are short to encourage discussion by attendees, who will each bring their own expertise and thereby expand the scope of the session. Digital Humanities at large needs this kind of conversation in which feminist practice is deliberately applied in labs, archives, local communities, literary histories, and classrooms.

Notes

1. Liu, Alan. “Where Is Cultural Criticism in the Humanities?” http://liu.english.ucsb.edu/where-is-cultural-criticism-in-the-digital-humanities/
2. see for example: http://mayabielinski.com/?p=297 , http://www.trevorowens.org/2011/07/the-digital-humanities-as-the-diy-humanities/ ,
3. see July 24 comment at 12:38 pm
4. Posner, Miriam. “Some Things to Think About Before You Exhort Everybody to Code.” http://miriamposner.com / blog /?p=1135
5. See Stephen Ramsay’s comment posted March 3, 2012 at 9:47 am: http://miriamposner.com/blog/?p=1135&cpage=1#comments
6. “The Hot Thing” is the title of Stephen Ramsay’s talk he gave at the Debates in DH launch April 2012. http://stephenramsay.us/text/2012/04/09/hot-thing.html

Seeking feminisms in digital literary archives

Wernimont, Jacqueline

The connections between digital literary projects and certain kinds of feminist work are manifold. The Women Writer’s Project and the Orlando Project are exemplary for their commitment to the recovery of women’s textual work and lives. These projects, and other gender-based digital projects afford users the thrill and affirmation of having “women’s countless contributions to Western culture and society made visible.” There is little doubt that such projects make texts available for reading, research, criticism, and teaching in ways that the print industry is increasingly unable to do.

Nevertheless, I argue that we need to return with a critical eye to such projects in order to better understand how, if at all, we should understand the technological and methodological components of these digital literary projects as also feminist. It is clear from the histories of such projects that feminist ideologies are central to their inception and their collection practices — but is that the same as saying that there is a feminist encoding practice or a feminist approach to interface design? Can XML or xHTML or PHP be deployed as a feminist intervention, and how would that differ from other deployments? Is it enough to talk about feminist workflows? What kinds of users should be addressed or created by a feminist archive? Where, in an assessment of digital archives should we look to find the traces of feminist epistemology, practice, or politics?

In many ways, this need to grapple with a possible feminist forms and practices, as well as an underlying feminist impulse, speaks to two issues raised at the 2011 Modern Language Association annual meeting around digital humanities projects: the roles of building and theory. By asking how various projects are constructed, conceptually, materially, and in terms of encoding, I am asking if building — the creation of digital literary texts, of digital literary archives, of interfaces that engage literary texts — can be helpfully approached as itself theoretical. Can we usefully theorize the “under the hood” work that goes into digital literary archives in feminist terms? In so far as feminist literary praxis has come to be understood as entailing the twin projects of critique and construction, the answer is yes. But there are limits and boundaries to my positive answer, which may help us begin to develop a more dynamic theory of the “hermeneutics” of building suggested by Stephen Ramsay. Such theorizing is critical as an address to the critical silence, or the curious absence of cultural criticism within digital humanities work. As Alan Liu suggests, this absence threatens the vitality of digital literary scholarship by failing to cast our insights into literary texts, digital repositories, textual structures, and media translation in the context of cultural analysis — for Liu, we miss the opportunity to leverage our literary insights into cultural insights. I would argue that this seems to be particularly true for feminist theory, which seems to be relatively absent from digital humanities interventions, despite the number of literary archival project that began from a feminist impulse of one sort or another and the powerful ways in which it can help address imbalances in technological work and culture.

Notes

1. Susan Fraiman. “In Search of Our Mothers’ Gardens —With Help from a New Digital Resource for Literary Scholars,” *Modern Philology*, August 2008, 142-48.
2. See Stephen Ramsay on the issues of building that he raised in “On Building”: http://lenz.unl.edu/wordpress/?p=340.
3. See, for example, Maggie Humm’s “Feminist Literary Theory” in *Contemporary Feminist Theories*, Jackie Jones and Stevi Jackson, eds. Edinburgh University Press, 1998, Mary C. Carruth *Feminists Intervene in Early American Studies, Part 2: New Directions Early American Literature* - Volume 44, Number 3, 2009, pp. 639-640
4. Alan Liu, “Where is Cultural Criticism in the Digital Humanities?” http://liu.english.ucsb.edu/where-is-cultural-criticism-in-the-digital-humanities/

Loud Silences in Digital Archives

Harris, Katherine D.

In "The Master's Tools Will Never Dismantle the Master's House" (1984), Audre Lorde identified a schism in feminism that would include missing voices, those voices that did not align themselves with patriarchal control, voices that refused to work within the system to gain power. In Digital Humanities' interactions with literary studies, especially in the construction of databases, digital archives, and repositories, those marginalized voices exist, but they exist outside the scope of the traditional literary canon even still. Amy Earhart and Jamie Skye Bianco both notice this lack in digital representations of historical and literary materials; while Earhart focuses on the lack of diversity and the replication of the standard literary canon in "Can Information Be Unfettered?: Race and the New Digital Humanities Canon," Skye Bianco asserts something more provocative about the very infrastructure of Digital Humanities:

Boiled down blithely, the theory is in the tool, and we code tools. Clearly this position never refers to Audre Lorde's famous essays on tools nor to 'the uses of anger," but it does summon their politics. . . . Tools don't reflect upon their own making, use, or circulation or upon the constraints under which their constitution becomes legible, much less attractive to funding. They certainly cannot account for their circulations and relations, the discourses and epistemic constellations in which they resonate. They cannot take responsibility for the social relations they inflect or control. Nor do they explain why only 10 percent of today's computer science majors are women, a huge drop from 39 percent in 1984, and 87 percent of Wikipedia editors—that would be the first-tier online resource for information after a Google search—are men. Tools may track and compile data around these questions, visualize and configure it through interactive interfaces and porous databases, but what then? What do we do with the data? ("This Digital Humanities Which is Not One," *Debates in the Digital Humanities* 99)

The tools, like mark-up, by their very nature ennact a sort of politics that replicates these archival silences that were the topics inspired by Miriam Posner's blog post. By offering a "stable publication environment" and peer review to small-scale digital scholarly editions, the 2012 inaugural issue of the revised *Scholarly Editing* (http://scholarlyediting.org/), under the editors Amanda Gailey and Andrew Jewell, attempts to balance the digital offerings of cultural materials beyond canonical authors and figures. In "Googling the Victorians," Patrick Leary concludes his essay by asserting that whatever does not end up in a digital archive, represented as cyber/hypertext will not, in the future, be studied, remembered, valorized and canonized. Though this statement reflects some hysteria about the loss of the print book, it is also revealing in its recognition that digital representations have become common and widespread, regardless of professional standards. But, as Gailey and Jewell point out, digital editions and archives haven't lived up to their promise to provide access to inaccessible and non-canonical materials—most among these are works by women.

While Digital Humanities pushes ever outward toward innovation, the issue of feminist recovery projects and scholarly editions still persists on the margins. In order to attract funding, even users, these types of digital projects have to represent the stars of the literary canon. This, in effect, crushes the purposes of the *digital* archive—to provide access to an under-represented set of authors. If the traditionally marginalized authors are marginalized *now* because it's no longer innovative to digitize and mark-up those collections, then how far have we really come? The voices that are lost, those silences in the archives, represent gaps in the traditional literary canon. How can Digital Humanities return to those small feminist recovery projects, offering help, professional credit, and authority?

Debugging "The Personal Is Political": Uncle Roger's Grandmother

Inman Berens, Kathi

Prior to reading Jill Walker Rettberg's excellent Electronic Literature Seen From a Distance: The Beginnings of a Field," I'd suspected that Judy Malloy's elision from the electronic literature reception history as the *first* author of hypertext fiction was attributable to genre: her comic piece *Uncle Roger*, a romp through Silicon Valley set in then-present day 1986, didn't evince the seriousness, ambiguity, and intricate plotting that critics and other purveyors of taste associate with high art. I accepted Robert Coover's 1992 declaration of Michael Joyce as the "granddaddy of e-lit" without question, even though Judy's *Uncle Roger* pre-dates Michael's *afternoon: a story* by at least one year and possibly three, if one measures from publication date rather than *afternoon*'s introduction to the coterie of enthusiasts who exchanged stories authored on Hypercard and other systems. *Afternoon* is a magnificent work that merits its august reputation. But Rettberg traces the far-reaching implications of that reputation in her distant reading, which demonstrates that *afternoon* is—by an order of magnitude—the most cited and taught work of electronic literature. The status Coover conferred in his review became a self-fulfilling prophecy. It's such a small thing, just one sentence in the *NYTimes*; but its impact has been field-defining.

Afternoon's ISBN, and *Uncle Roger*'s lack of one, is a crucial differentiator in Malloy's and Joyce's divergent receptions. As Rettberg's analysis shows, the presence or

absence of an ISBN determined whether a work could be archived, collected or sold. The other key differentiator was the invention of the browser. Malloy's *Uncle Roger* initially excited popular attention as articles about it in *Newsweek* and *The Wall Street Journal* attest. But there was no way for people to follow up on their curiosity, because in 1986, the browser was still four years from being invented by Sir Tim Berners-Lee. Hence attention to *Uncle Roger* was ephemeral. *Afternoon*'s publication, by contrast, was effectively coincident with introduction of the browser in 1990; the ability to find the work and then to buy it hugely impacted reception.

In an interview with Judy Malloy in July 2012 near her home in Berkeley, I surmised that the though the ISBN is attached to relatively few pieces of early hypertext, it united disparate stewards (programmers/developers, librarians, academics, vendors) to collect and fortify those works against bit-rot or obsolescence. Works lacking ISBNs, such as *Uncle Roger*, were left to the authors to maintain or abandon; in point of fact, it would be much later (1997) before Malloy would author *Uncle Roger* in a browser-friendly format. By then the excitement for the novelty of hypertext had given way to interest in Flash-based works. A moment had passed and with it, the power that comes from cultural currency.

Such different fates for these early hypertextual works adumbrate the the artists' career trajectories. Joyce is an internationally acclaimed writer and tenured full professor. Malloy is an internationally acclaimed writer/programmer who is struggling—and failing—to land part-time teaching work. To compare them is to debug the "Personal Is Political": the old story about seemingly "individual" choices disclosing, in aggregate, a systemic exclusionary logic. What does it mean for this familiar paradigm to survive our shift into seemingly disembodied virtual environments? What "post-feminism"? What "post-human"?

During our own cultural moment in 2012, when we are figuring out why curation and live presence matters for e-lit works that are for the most part perpetually available online, one goal of a Malloy reception history would be to show that from the traditionally "female" work of organizing gatherings to share and exchange work — a process now, thankfully, archived in databases like ELMCIP and so "visible" — a polysemous picture of field activity can be a site of feminist intervention.

References

Rettberg, J. W. Electronic Literature Seen From a Distance: Beginnings of a Field. http://www.dichtung-digital.org/2012/41/walker-rettberg/walker-rettberg.htm

Coover, R. (1992). The End of Books. *New York Times Book Review*. June 21.

ELMCIP: Electronic Literature as a Model of Creativity and Innovation in Practice is a collaborative research project funded by Humanities in the European Research Area (HERA) JRP for Creativity and Innovation. ELMCIP involves seven European academic research partners and one non-academic partner who are investigating how creative communities of practitioners form within a transnational and transcultural context in a globalized and distributed communication environment. It focuses on electronic literature as a model of networked creativity.

hooks in the 21st Century: Feminist Pedagogy in Action

Grigar, Dene

In 1994 belle hooks wrote in *Teaching to Transgress: Education as the Practice of Freedom* that "[t]here is a serious crisis in education. Students often do not want to learn and teachers do not want to teach . . . " (12). A clear lack of excitement in the classroom, which some see as "disruptive of the atmosphere of seriousness," as well as a lack of engagement and of freedom to explore are among some of the ills she cites in her book (6-10). hooks' call to action empowered many feminist teachers to experiment with new teaching strategies, to address the needs of the whole student, and to become with their students active participants in the learning process—in short, it inspired us to embrace what she refers to as "transformative pedagogy" (39).

In 2006 I arrived at Washington State University Vancouver to build the Creative Media & Digital Culture Program, bringing with me the principles of transformative pedagogy and eager to apply them to the digital media classroom. This paper lays out the approaches and projects that I and the faculty in the program have undertaken in the last six and a half years and the positive outcomes for students, faculty, and staff for this grand experiment. One example the paper highlights is the Mobile Tech Research Initiative (MTRI) that provided all of the faculty, staff, and 10 students, during summer 2011, to learn—together—how to design for and develop mobile apps. Out of MTRI our program was able to integrate mobile media into all of our curriculum; our students were able to receive fully funded fellowships and were fast-tracked through the program and land good jobs and placement in graduate programs specializing in interactive design; and faculty have been able to continue creating mobile media for their scholarship.

The paper provides those interested in transformative pedagogy with best practices for applying this approach to

teaching in their classroom. As hooks reminds us, "[t]he academy is not paradise. But learning is a place where paradise can be created" (207).

Computational Rhetoric: Adapting Graph Theory Analytics to Big Data

Hart-Davidson, William
hartdav2@msu.edu
Michigan State University, United States of America

Rehberger, Dean
rehberge@msu.edu
Michigan State University, United States of America

Grabill, Jeffrey
grabill@msu.edu
Michigan State University, United States of America

Omizo, Ryan
omizo@msu.edu
Michigan State University, United States of America

Computational Rhetoric: Adapting Graph Theory Analytics to Big Data

This panel presents three short papers by a research group working in a new area called "computational rhetoric." As the name implies, computational methods are used to perform rhetorical analysis on large corpora of texts. In this case, our examples are drawn from online discussion forums related to science topics and hosted by informal learning institutions (science centers and museums). The purpose of the panel is to present a few highly experimental methods developed to conduct rhetorical analysis on big data sets for critique and feedback by members of the DH community who might find such methods useful.

The Rhetoric of Facilitating Learning in Social Media Environments

Speaker one will focus on specific outcomes from five years of studying how people interact via social software tools hosted by two science museums and designed to support science learning. Studies have shown that computer-based learning environments can make inquiry experiences more successful by offering participants cognitive and procedural guidance (diSessa, 1992; Kafai & Resnick, 1996; Linn & Hsi, 2000; Linn & Slotta, 2000; White & Fredericksen, 1998). Internet- based learning environments can also provide embedded reflective opportunities that capture participants' abilities to critique evidence, make arguments, and reach conclusions.

Our results suggest that social media environments can be effective informal learning spaces because of discourse characteristics that seem to support learning, such as sharing, making ideas public, and writing to learn, all facilitated in particular ways (e.g., invitations to explain and connect; connecting people and their ideas together). My focus in this paper, however, is on the specific rhetorical moves of facilitation that lead to productive outcomes and how to reliably identify them in large discourse sets. As a rhetoric scholar leading a project on science learning, it is this area of expertise—rhetorical discourse analysis—and our ability to develop analytics at scale that has been useful to the larger project and hopefully interesting and relevant to other digital humanists.

The analytical work outlined in this paper begins with understanding that most social software environments are written. That is, people interact through what they write, and so analyzing those written interactions is fundamental to understanding online activity. There are exceptions, of course, but the point is to turn to analytic techniques and expertise appropriate for the modes of interaction (writing, image, video, and so on). In our case, we turned to rhetorical theory and to discourse analysis in order to supplement approaches from the learning sciences. Discourse analysis is commonly used in communication studies to characterize and understand interactions (Fairclough, 1992; Dijk, 1997; Wood & Kroger, 2000; Schiffrin, Tannen, & Hamilton, 2001; Bazerman & Prior, 2004; Gee, 2005). This paper sets the scene for the project and devote much of my time to laying out the analytical approach and tools for the audience. My doing so will make the two other papers on this panel intelligible, but this approach to rhetorical discourse analysis also represents an innovation in rhetorical studies of online interactions (see Grabill and Pigg, 2012).

Betweenness Centrality as Rhetorical Arrangement

Speaker Two reports on experimental protocols used to transform the facilitation data discussed by Speaker 1 into

computational formats with the aim of producing an analytic program that might replicate the insights of the human coders on large-scale datasets. These protocols incorporate text tokenization methods employed in computational linguistics and natural language processing to produce contiguous bigrams of words, which, aside from being the smallest graphable unit, we believe retain the lexical and temporal contours of the original discussion board threads.

Speaker Two will then outline a graph-driven approach to these bigram units based on the measure of *betweenness centrality*. Developed in the field of social network analysis, betweenness centrality holds that for a given sequences of connected nodes numbering more than 2, the node positioned between the most adjacent nodes will serve as the prime mediator of information desseminated through the network, thus rendering it the controller node (see Borgatti, 2005; Freeman, 1979; Freeman, Borgatti, & White, 1991; Friedkin, 1991; Hanneman & Riddle, 2011). Though this type of social network measure is usually applied to discrete entities or interrelated social actors, Speaker Two will demonstrate how we can use the betweenness centrality of interrelated words within a text to understand how a text is organized according to critical terms deployed at critical moments over time. In short, we are interested in how words with high betweenness scores influence words associated with them, signalling probative indices, which we call “matters of concern.”

In marking these matters of concern on discussion board threads focused on the facilitation of science learning, we feel that we have arrived at two significant innovations that could interest the community of digital humanists. First, the ability to graph matters of concern can serve as a productive first step in the global, computational rhetorical analysis of big data. This move could lead to more granular analysis and specific discursive codings. Second, the emergence of machine-driven rhetorical analytics powered by the pattern distribution of bigrams challenges conventional notions of a how written communication creates meaning.

Aristotle, Toulmin, Erdös-Rényi & Markov: Modeling rhetorical moves and patterns of argument using graph theory concepts

Speaker Three presents a brief review of the various mathematical models used in our attempts to render internet discussion threads as computational objects (graphs). The focus of this paper is on the correspondence between these mathematical models and concepts from rhetorical theory, beginning with the way both human coders and our computer model focus on basic units of analysis, and discussing how these units are understood to form larger discursive structures such as arguments. We will also show how we understand random graph and Markov chain models to provide us with a baseline for describing and perhaps predicting the way discussion threads develop over time.

Our group is not the first to propose using text-extraction techniques from Natural Language Processing and mathematical models, some adapted from computational linguistics, to examine rhetorical patterns (see, Grasso 2002, Reed & Grasso 2007, Grasso, et. al. 2010). Our methods take a turn from a relatively strict focus on argument dialectic, strictly defined, to a more holistic view of rhetorical reasoning, Our aim is to understand what rhetorical strategies work in specific situations and what, if any, relationships can be detected between rhetorical patterns deployed and specific outcomes desired. In aim, our work is similar to Ishizakie, Kaufer & colleagues’ efforts using their DocuScope analysis software (Ishizaki & Kaufer, 2011). Where our work differs a bit from Ishizaki & Kaufer is in the use of graph theory techniques to construct textual models for analysis.

Our approach to computational rhetorics, metaphorically, involves attempts to assay complex rhetorical compounds found in nature, with rules of grammar and syntax providing specific affinities to bind markers, amplify & stabilize signals against a noisy background. We seek to isolate the discursive equivalent of what biologists refer to as chemical 'pathways' that yield certain kinds of results in the real world. Of course, we don’t call them pathways. Rather, we are trying to understand if more familiar rhetorical terms like "topoi" might function like chemical pathways. If so, we can consider certain patterns as normal (expected) or abnormal (reflective of some pathology). This view also suggests that there may methods to inhibit a pathway at some point, thus changing the outcome but in a (relatively) predictable way. To know for sure, we must first perform some ‘basic science’ to understand how useful known models that could be applied – here we focus on Erdös-Rényi random graphs and Markov Chains – can be for describing rhetorical patterns in a large text corpus.

References

Bazerman, C., and P. Prior (eds). (2004). *What writing does and how it does it: An introduction to analyzing texts and textual practices*. Mahwah, NJ: Lawrence Erlbaum.

diSessa, A. (1992). Images of Learning. In De Corte, E., M. C. Linn, H. Mandel, and L. Verschaffel (eds.), *Computer-based learning environments and problem solving*. Berlin: Springer-Verlag.

Dijk, T. A. V. (1997). Discourse as interaction in society. In Dijk, T. A. V. (ed.), *Discourse as social interaction*. London: SAGE. 1-37.

Fairclough, N. (1992). Discourse and social change. Cambridge, UK: Polity Press.

Gee, J. P. (2005). *An Introduction to Discourse Analysis: Theory and Method*. London: Routledge.

Grabill, J. T., and S. Pigg (2012). Messy Rhetoric: Identity performance as rhetorical agency in online public forums. *Rhetoric Society Quarterly*. 42: 99-119

Grasso, F. (2002). Toward Computational Rhetoric. *Informal Reasoning* 22(3): 195-229.

Grasso, F., I. Rahwan, C. Reed, and G. R. Simari (2010). Introducing Argument & Computation , Argument & Computation, 1(1): 1-5.

Ishizaki, S., and D. Kaufer (2012). Computer-Aided Rhetorical Analysis. A*pplied Natural Language Processing and Content Analysis: Identification, Investigation, and Resolution*, ed. McCarthy, P. & Boonthum, C. IDI Global. 276-296.

Kafai, Y., and M. Resnick (eds.) (1996). *Constructionism in practice: Designing, thinking, and learning in a digital world*. Mahwah: Erlbaum.

Linn, M. C., and S. Hsi (2000). *Computers, Teachers, Peers: Science Learning Partners*. Mahwah: Erlbaum.

Linn, M.C. and J.D. Slotta (2000, October) WISE science. *Educational Leadership*, 58(2): 29-32.

Reed, C., and F. Grasso. (2007). Recent advances in computational models of natural argument', *Int. J. Intell. Syst.* 22: 1–15.

Schiffrin, D., D. Tannen, and H. E. Hamilton (eds.) (2001). *Handbook of discourse analysis*. Oxford: Blackwell

White, B. Y., and J. R. Frederiksen (1998). Inquiry, modeling, and metacognition: Making science accessible to students. *Cognition and Instruction*. 16: 3-118

Wood, L. A., and R. O. Kroger (2000). *Doing discourse analysis*. Thousand Oaks, CA: Sage.

Center for Historical Information and Analysis: Big Data in History

Manning, Patrick
planeterra@comcast.net
University of Pittsburgh, United States of America

Mostern, Ruth
rmostern@ucmerced.edu
University of California - Merced, United States of America

Cao, Kai
kai.cao@pitt.edu
University of Pittsburgh, United States of America

Johnson, Ian
ian.johnson@sydney.edu.au
University of Sydney, Australia

Session Abstract

This session address the development of a significant new digital humanities resource in the form of a world-historical dataset. The session provides an overview of the project and details of two key elements in its early stages of development. The commentator will set this project in the context of other complex, multidisciplinary datasets. The Collaborative for Historical Information and Analysis (http://chia.pitt.edu) is a multi-institutional collaborative of scholars in humanities and in social, natural, and information sciences. The Collaborative, structured as a Research Collaborative, Headquarters, and a wider informal network, has recently received major support from the National Science Foundation, which has provided CHIA with an award in its Building Capacity and Community program. On the Research Collaborative side, the award is to strengthen the organizational and technical infrastructure of linking participating institutions that are collecting data on historical population, climate, and other topics with a crowdsourcing tool to demonstrate the feasibility of building a continuously growing collection of diverse historical data and metadata. On the Headquarters side, the award is to assemble and develop knowledge on repository design to develop a repository sufficient to house the incoming data and permit global and interactive analysis. The Collaborative for Historical Information and Analysis's future plans include expanding its collection and processing of historical data, broadening its community of social and natural science researchers, analyzing historical patterns of global change, and sharing its resources with researchers, policy-makers, teachers and students. CHIA is headquartered at the University of Pittsburgh with formal affiliates at the University of California – Merced and the International Institute of Social History (Amsterdam) plus two affiliates at Harvard University (Institute for Quantitative Social Sciences and Center for Geographic Analysis). Other associated groups are at Boston University and Michigan State University (participating in the NSF-funded project) as well as the University of Portsmouth,

the University of California, Irvine, and the Council for the Development of Economic and Social Research in Africa (CODESRIA).

Three papers discuss varying aspects of CHIA and its work. Patrick Manning of the University of Pittsburgh, director of the project, provides an overview of the project objectives, philosophy, structure, and its practical milestones. Ruth Mostern of the University of California - Merced, a member of the governing Executive Committee, presents on the issue of soliciting, linking, and evaluating scholarly datasets. Kai Cao, the technical lead on the project at the University of Pittsburgh, describes the work of creating the prototype archive.. The discussant, Ian Johnson, the developer of a parallel project at the University of Sydney, will comment on both generalities and specifics of the CHIA project as presented.

The Collaborative for Historical Information and Analysis: Framework for Creating a World-Historical Data Repository

Manning, Patrick

Big Data in history will provide a new, comprehensive level of documentation on the past. Currently available historical information, while immense in its overall quantity, is scattered and dispersed. Libraries and archives in great cities hold treasure troves of data on trade, politics, and religion for national and imperial centers, but each archive is separate from the others, and the totality of their records provides scanty information on people of rural areas. The idea of Big Data in history is to digitize a growing portion of existing historical documentation, to link the scattered records to each other (by place, time, and topic), and to create a comprehensive picture of changes in human society over the past four or five centuries.

The Collaborative for Historical Information and Analysis (CHIA), formed in 2011, now has five institutional members based in the U.S. and Europe, four additional, informal associated institutions, and expects to grow further through links with such organizations as the Council for the Development of Economic and Social Research in Africa (CODESRIA, based in Dakar). The Collaborative, directed by an international Executive Committee, is administered at the University of Pittsburgh. Its collaborative projects include the creation of a prototype and a full archive able to contain consistent and documented world historical data, along with systems for gathering and incorporating new data and for analyzing and visualizing the data. Data included are expected to begin with demographic, economic, social, political, health, and environmental variables, and to display their patterns and interactions. This presentation will trace the development of the overall project over the past five years and identify the main problems to be taken on in the next three years.

Tasks addressed from 2007 have included initial articulation of the objective, overall design of the research project and work on the initial steps of implementing several aspects of that design. In particular project members have emphasized recruiting and sustaining collaborating groups at several institutions. In addition, work has included collection of various sorts of historical data, design of the repository for worldwide historical data, development of an ontology describing world-historical data at various levels of detail, and systems of cleaning and documenting data.

At an empirical level, major advances have taken place in collecting, displaying, and linking U.S. data on disease, climate, and population.

The NSF-supported project includes three years of a projected ten-year project to create a global- historical dataset, with the hope that it would be taken over and sustained through the efforts of UNESCO or the World Bank. The CHIA project pledges to maintain open-source, open-access, non-proprietary standards throughout its work in constructing a world-historical archive. We have allocated the range of our activities among three basic missions.
Mission 1. Gather historical data.
Mission 2. Aggregate data up to the global level.
Mission 3. Visualize, analyze, and mine the data.

Currently funded infrastructure projects include:

- **Crowd-sourcing application** for collecting and archiving historical and social data will open the bottleneck that has so far prevented systematic study of human society at a large scale (University of Pittsburgh).
- **Prototype archive** – programming a prototype for global integration and visualization of data, relying on the model of the Dataverse Network and a selection of world-wide data for the twentieth century(University of Pittsburgh, Harvard University).
- **"Hoovering" data**, the collection of available datasets and a survey of social scientists to determine the availability of historical datasets (UC-Merced)
- **Data retrieval for South Asia and Southeast Asia** led by the Asian Studies group at Michigan State University
- **Colonialism** – integration and visualization of data collected in a previously project, CLIO, at Boston University.
- **Collaboration as infrastructure among social scientists** all over the world, through the creation and

maintaining of a global system of historical data, will bring additional sharing of data and analysis.

Additional areas of activity

- **Peer-reviewing of datasets** through the Journal of World-Historical Information will bring recognition of the scholarly value of creating datasets, and will ensure that high standards for creating historical datasets are created and maintained.
- **Archive design at Big-Data scale** – design and programming through XSEDE program in association with the Pittsburgh Supercomputer Center.
- **Labour Relations** – a program of distributed historical research to assess the structure of labor forces in numerous historical situations, supported by the Netherlands National Science Foundation (International Institute of Social History).
- **Technical and analytical skills of social scientists** will advance through the process of collecting and analyzing data, and demonstrate the parallels and the links of social sciences and natural sciences.
- **Theoretical debate.** The expanded effort to link and apply social science theories, especially in order to fill in missing historical data, will strengthen theory and analysis in social science.

To understand global social patterns as they exist today, it is increasingly clear that we need to understand how they have evolved over recent centuries. The Collaborative for Historical Information and Analysis responds to this need and takes historical analysis into the realm of Big Data. It is expected that the data resources will grow to several terabytes in size. This project will stimulate development of more efficient research collaborations, enabling systematic large-scale consolidation of diverse historical data sources. Once collected and integrated, the data repository and analytical system will allow scholars to address a wider set of questions testing hypotheses about long-term and short-term social change at the global scale and catalyzing an expansion of the evidence base in humanities and social sciences. For example, our understanding of important societal issues can advance by linking health to demography and by incorporating climate and health factors into economic studies. Disciplinary theory will advance through interaction among the various scientific fields, so that a global network of humanities and social-science researchers will emerge.

The project addresses the global dynamics in humanities issues and social-science variables over the past several centuries. Contemporary globalization and concerns about future global trends naturally raise questions about past patterns of global change. What were the interactions of population, economy, governance, and social inequality with each other and with climate and disease? Historical social science, focused at national and subnational levels, has scarcely addressed global issues. Our group expects to collect, document, and analyze historical data to permit cross-disciplinary analysis of human society over time. The overall topic is immense, but we believe we have found an orderly and productive way to work on it.

How are we to create consistent data at regional and global levels over time? Our group, rather than tunneling within a single discipline, seeks to coordinate data collection and research in multiple disciplines. We advance an explicit focus on the global and historical character of human society. The existing data sources are mostly used for regional comparative efforts; they vary widely in degree of consistency, reliability, completeness, as well as in data representation format. The task of large-scale data utilization can only be resolved via collaborative efforts within a large network of researchers.

What criteria distinguish this global strategy in humanities research from other large-scale projects? Our project is not simply to archive large quantities of data but to define and link them into a single overarching set of interacting, historical data. We require a coherent metadata framework to link data to their sources and each other. Creating this mass of new metadata—as we incorporate, integrate, and aggregate data—requires a strong ontological base and a crowdsourcing procedure to link many contributors. Our collection of base data on population worldwide is to go back four centuries and to include migration and other extra-census data—it is thus complementary to rather than competitive with the Terra Populus (University of Minnesota) collection of census data.

The Collaborative responds to an imperative of the current moment: the need to understand global social patterns not only as they exist today but as they have evolved over some four centuries. The program argues that some national resources should be put into research at this broad level, to clarify and diffuse a global strategy in social-science research. It also focuses on population and climate as key layers of global data. In research agenda, CHIA addresses human interaction with the natural world, global population change, patterns in social inequality, and local and global patterns of governance.

Soliciting, Integrating and Evaluating World-Historical Data

Mostern, Ruth

A signal characteristic of world-historical data is that much of it will need to be assembled piecemeal from datasets created by specialists. Some of these datasets, such as those which concern climate information, are quite large. However, at the global historical scale, even climate data needs to be integrated based on local and regional analyses. Ocean sediment samples, ice cores, or dendrochonologies may offer centuries of continuous information, but they concern particular locations. Census and epidemiology datasets often record information about millions of people, but they are episodic in nature and regional in spatial scale. Beyond these types of data, contributions range downward in size and upward in analytical complexity. The history of commodity exchange, for instance, requires meticulous reconstruction from bills of lading and tax documents that may be difficult to locate, trust, and make commensurable. One of the challenges of the CHIA effort concerns this work. The solution involves three tasks.

- "Hoovering" data. CHIA needs to engage in a labor-intensive process to identify the specialist holders of relevant datasets and work with them to solve issues of data structure and intellectual property that may prove to be barriers to contributing them to a CHIA archive. This paper will report about the development of a CHIA survey of historians and historical social scientists regarding their creation, preservation, integration, and use of historical datasets.
- Integrating data. CHIA needs to identify appropriate standards for the formal descrip- tions of historical datasets. Librarians and archivists stress the importance of appropriate metadata for guiding the ingest, discovery, and integration of datasets, and many domain- specific standards exist. We historians have our own challenges, since our disciplinary traditions (rich footnoting, bibliographies, and descriptive text about method) mean that we need particular standards for describing the dataset as a work, the primary and secondary sources (including other datasets) consulted in its production, and the operations conducted upon them to create the final work. Collaboratively developed datasets have primary and secondary authors as well as technical experts and publishers. This paper will discuss existing metadata standards, best practices for less formal data description, and the promise of linked data solutions.
- Evaluating data. Historical datasets lack established conventions of form, content, and genre. Authors do not have clearly recognized models to follow or oppose even as they seek to create effective, excellent, and communicative work; and reviewers, along with readers/users, have to assess the value or character of any given dataset *de novo* and *ad hoc*, rather than engaging with a given dataset as an exemplar of a familiar category. CHIA needs to develop standards for reviewing datasets and offering the imprimateur of publication in order to overcome disincentives to pursuing digital scholarship on the part of authors, and trusting it on the part of users. Until these matters are resolved, it will be difficult for historians to contribute data to the CHIA archive, for CHIA editors to evaluate data, and for the CHIA system to handle data content and data types that may be quite diverse. This paper will discuss CHIA efforts to evaluate datasets; in particular by publishing dataset reviews in the *Journal of World Historical Information*.

Creating a Prototype Archive for a World-Historical Dataset

Cao, Kai

A key element of the overall project of the Collaborative for Historical Information and Analysis is construction of an appropriate archival system. Three CHIA affiliates collaborate in the initial stage of archive construction by creating a prototype archive and retrieval system based on a "faceted search," to be created and tested by mid-2013. The collaborating groups are the World-Historical Dataverse at the University of Pittsburgh (Pitt) (with the author as lead developer), the Institute for Quantitative Social Science at Harvard (with Gustavo Durand as principal developer), and the Center for Geographic Analysis at Harvard (with Benjamin Lewis as developer). The archive will be based initially on the Dataverse Network (DVN), to enable linkage of multiple data files (which themselves include explicit spatial and temporal data) so as to develop data that can be searched so as to identify patterns at the global level and interactions among variables at various spatial and temporal levels. Multiple data files will be stored in the DVN system as "studies," and will be accessed by the system of retrieval.

The "faceted search" is the key element of the archive. It is a search portal, which enables users to define selected data by space, by time, and by topic in a text box. The search is "faceted" in the sense that when the user adjusts the range on one dimension (e.g., space), the range on the other dimensions adjusts appropriately. Once the search criteria are entered, the program identifies the studies that meet the search criteria and displays their geographic distribution within the bounding box on the map. With that, the user can then explore the studies by clicking the dots / links visualized in the mapping area.

Four categories of data have been identified for this task. The four categories of development are:

i. Global population data for the 20th century at national level or provincial level for countries exceeding 100 million to 2000, at 10-year or 5-year intervals.
ii. Climate for the 20th century for identified places and times within the same units as above.
iii. Silver flows of production and trade for the 20th century, by place, trajectory, and time (annual or quinquennial).
iv. Wars during the 20th century, identified by time, space, national or ethnic combatants, casualties.

A graduate researcher at the University of Pittsburgh has been engaged to collect these data.

In addition to the above outline, the presentation will address as many as possible of the details involved in constructing the archive, and will discuss how it is to be used in later phases of project activity. For each study, metadata are to be defined and implemented to ensure that each individual value is fully defined. As a next step, the localized files that are entered into the archive are to be aggregated in order to yield continental and global summaries of data. Besides, we will define and impose the elements of a mid-level ontology for the data, to define relations among all the topical categories, and allow users to use unstructured tags within it.

The archive and faceted search, once implemented, will enable users to envision the breadth of the world-historical analytical system, which is the ultimate objective of the work. The associated website will appear as a storytelling platform. We also expect to rely on social media to spread the word on the archive to convey the idea that its continued development can build a resource of global and interdisciplinary interest.

This archive and faceted search, after its initial development, is to be articulated with other aspects of the overall CHIA project. To begin with, based on the responses and evaluations of users, we will carry out revisions of the faceted search created in this Phase 1, advancing it to lower the barriers of entry for participants and expand the reach of the project. Further, we will encourage the collection of additional categories of data and develop models to improve the analysis of data. In perhaps the most important next stage, we will link the archive and faceted search to the crowd-sourcing data- input application under separate development at Pitt. With the linkage of these two applications, it will become possible for large quantities of data to be incorporated and integrated into the world-historical archive.

Text Theory, Digital Document, and the Practice of Digital Editions

Van Zundert, Joris Job
joris.van.zundert@huygens.knaw.nl
Huygens Institute for the History of the Netherlands-Royal Netherlands Academy of Arts and Sciences

Van den Heuvel, Charles
charles.van.den.heuvel@huygens.knaw.nl
Huygens Institute for the History of the Netherlands-Royal Netherlands Academy of Arts and Sciences

Brumfield, Ben
benwbrum@gmail.com
Independent Digital History Software Services

Van Dalen-Oskam, Karina
karina.van.dalen@huygens.knaw.nl
Huygens Institute for the History of the Netherlands-Royal Netherlands Academy of Arts and Sciences

Franzini, Greta
greta.franzini@gmail.com
UCL Centre for Digital Humanities

Sahle, Patrick
sahle@uni-koeln.de
Institute for Documentology and Scholarly Editing (IDE) / Cologne Center for eHumanities (CCeH)

Shaw, Ryan
ryan.b.shaw@gmail.com
University of North Carolina, Chapel Hill, School of Information and Library Science

Terras, Melissa
melissaterras@gmail.com
UCL Centre for Digital Humanities

Text Theory, Digital Document, and the Practice of Digital Editions

Subject

Many digital tools have been and are being developed aimed at transcribing, annotating and publishing editions of literary or historical texts making use of crowd sourcing for collaborative research. This panel discusses the question how well digital scholarly editions produced by such tools reflect the theoretical notions of digital scholarly editions, and how such may be assessed based on both empirical examination of current practices and text theory in the digital era.

The practice of preparing and producing digital editions is increasingly supported by purpose made and specialized digital tools, many of them involving crowd sourcing. Although an exhaustive survey and typology of these tools is still missing, by and large we can see that most of these tools are highly similar in functionality, text model, and editorial process. As such they express a fairly straightforward transformation from the physical book to a digital metaphor of the book, roughly along a trajectory of transcription, annotation and publication. Quite contrary text theory in the digital era seems to express a different scholarly ideal of representation of text, far more rooted in notions of instability (McGann 2001), fluid text (Bryant 2002), transclusion (Nelson 1982), text as process (Buzzetti 2002, Gabler 1987), transmedialization (Sahle 2010), and distributed editions (Zundert 2011) for instance. Also the usage patterns of emergent digital technologies and their applications such as Web2.0/3.0, crowd sourcing, cloud based services, open notebook science, data as service, multi device enabled layouts –to name but a few– seem to favor shaping the representation of digital text more in line with theory than with the practices of current scholarly digital edition tools.

If we follow Internet pioneer Cailliau (Cailliau 2012) information at our fingertips will become essentially undocumented, in the sense of not being a conventional cover-to-cover document, not even as a metaphor. Rather specific parts and facets of information will adapt to different devices and context, rather reminiscent of the concept of Nelson's envisioned docuverse (McKnight 1991). It should therefore be critically examined if it is still opportune and adequate to speak of a digital edition as a document, and if a digital editing process should necessarily lead to a single or physical publication to serve maximum scholarly expressiveness. Overall the tools in use mostly let the metaphor of the book be inferred as the de facto model for digital publication. But as users and scholars choose and adapt new technologies and new forms of engagement with information, should scholarly publishing and digital editing follow these patterns of usage? Does an audience oriented approach such as the ideas on minimal and maximal editions expressed by Vanhoutte (Vanhoutte 2011) strike a middle ground between theory and practice? What is the intellectual loss or gain if the metaphor of the book prevails over using the medium of the Internet as an expression of text as process?

It is these kinds of questions about the theoretical underpinnings of the digital scholarly edition that arise at the intersection of shaping technologies, standing scholarly practice, and changing usage. Trying to establish some practices of quality a number of comparative studies have been conducted into transcription tools and crowd sourcing tools for digital editions, notably by members of this panel. Most of these studies have been based on analysis and comparison of functional requirements, usability aspects, and user feedback. However, a text theory based aspect of evaluation of such tools and editions is mostly lacking.

This panel will explore the issue of practice *and* theory based quality assessment of digital editions, building on the results of a comparative text theory based empirical survey of tools for digital scholarly editions the design of which is the subject of our preparatory paper presented at the NeDiMAH Expert Meeting on Digital Scholarly Editions held in conjunction with the 2012 conference of the European Society for Textual Scholarship. The panel focuses on several prominent digital tools and projects for preparing digital scholarly editions with varying approaches. From this broader view specific themes and issues will be examined, such as:

- The metaphor of the book as enabler or inhibiter of new avenues for research.
- User surveys and feedback as shaping forces of tools for digital editions.
- The role of users, editors, researchers, and funders in determining quality aspects.
- The digital edition as an expression of text in flux versus the iconic object.
- Text models for distributed documents.
- Is the generic or the specific a hallmark of quality of tools for digital editions?
- Crowd sourcing and open notebook science as determining aspects of digital editions.
- Visualization of instability of text as a scholarly quality of the digital edition.
- The relationship between formalization of editorial process and the instability of text.

Organization of the panel

The methodological research program of the Huygens Institute for the History of the Netherlands, part of the Royal Netherlands Academy of Arts and Sciences (KNAW) is the initiator and organizer of this panel. Panel members include e-Humanities researchers from KNAW involved in the development of the transcription and crowd sourcing tool eLaborate (https://www.elaborate.huygens.knaw.nl/); researchers from University College London involved with amongst others the Transcribe Bentham Project (http://www.ucl.ac.uk/transcribe-bentham/); researchers and editors of digital editions (e.g. http://www.i-d-e.de/); researchers working on open science approaches (e.g. http://editorsnotes.org/ and http://ecai.org/mellon2010/); and developers and researchers of crowd sourcing software (http://manuscripttranscription.blogspot.com). The panelists are engaged in the study and development of different digital humanities tools and projects pertaining to digital scholarly editions, specializing in transcription tools and crowd sourcing projects, which grants this panel a unique opportunity to comparatively explore various strategies in building and using digital editions, and to reflect on both theoretical and practical concerns of that process. In addition, the panel will critically evaluate the themes and issues listed above.

The panel session will be organized in the following way:

- The panel chair will introduce the main topic, discussion questions, and the panelists; duration: 3 minutes;
- Each of the panelists will give a short presentation (6 minutes), followed by questions from the audience (4 minutes); duration: 60 minutes;
- The themes and questions raised in the presentations will be further discussed in an open forum between the panelists and the audience; duration: 25 minutes
- The panel chair will briefly reflect on future plans, provide contact information, and close the panel: 2 minutes.

The names and affiliations of confirmed panelists are as follows:

- Ben Brumfield, Independent Digital History Software Services, Austin Texas (US)
- Karina van Dalen-Oskam, Huygens ING (The Netherlands)
- Greta Franzini, UCL Centre for Digital Humanities (UK)
- Patrick Sahle, Institute for Documentology and Scholarly Editing (IDE) / Cologne Center for eHumanities (CCeH) (Germany)
- Ryan Shaw, University of North Carolina, Chapel Hill, School of Information and Library Science (US)
- Mellisa Terras, UCL Centre for Digital Humanities (UK)

Moderators for this session are:

- Charles van den Heuvel (panel chair)
  Huygens Institute for the History of the Netherlands (Royal Netherlands Academy of Arts and Sciences)
- Joris van Zundert
  Huygens Institute for the History of the Netherlands (Royal Netherlands Academy of Arts and Sciences)

References

Bryant, J. L. (2002). *The Fluid Text: A Theory of Revision and Editing for Book and Screen.* University of Michigan Press.

Buzzetti, D. (2002). Digital Representation and the Text Model, *New Literary History.* 33: 61–88.

Cailliau, R. (2012).'WWW: Etat des lieux, Grande conférence de Robert Cailliau'. In Cailliau, R. *Les pionniers de l'Internet en Europe*. held 22 October 2012 at Mons University. http://expositions.mundaneum.org/fr/conferences/robert-cailliau-co-inventeur-du-world-wide-web (accessed 23 October 2012).

Gabler, H. W. (1987). The Text as Process and the Problem of Intentionality. *TEXT, Transactions of the Society for Textual Scholarship* 3: 107–116.

McGann, J. (2001). Radiant Textuality: Literature after the World Wide Web. New York: Palgrave.

McKnight, C., A. Dillon, and J. Richardson (1991). Hypertext in Context. Cambridge University Press.

Nelson, T. H. (1982). Literary Machines. Minfull Press.

Sahle, P. (2010). Zwischen Mediengebundenheit und Transmedialisierung. Anmerkungen zum Verhältnis von Edition und Medien. *Editio* 24: 23-36.

Vanhoutte, E. (2011). So You Think You Can Edit? The Masterchef Edition. *The Mind Tool: Edward Vanhoutte's Blog*. http://edwardvanhoutte.blogspot.nl/2011/10/so-you-think-you-can-edit-masterchef.html (accessed 14 March 2013).

van Zundert, J., and P. Boot (2011). The Digital Edition 2.0 and the Digital Library: Services, not Resources. *Digitale Edition und Forschungsbibliothek* 44: 141–152.

Further reading

Buckland, M. (1998). What is a 'digital document'? *Document Numérique* 2(2): 221-230. http://people.ischool.berkeley.edu/~buckland/digdoc.html (accessed 21 October 2012).

Eggert, P. (2009). *Securing the Past: Conservation in Art, Architecture and Literature*. Cambridge: Cambridge University Press.

Pierazzo, E. (2011). A rationale of digital documentary editions. *Literary and Linguistic Computing* 26(4): 463–477. doi: 10.1093/llc/fqr033

Robinson, P. (2005). Current issues in making digital editions of medieval texts — or, do electronic scholarly editions have a future? *Digital Medievalist* 1 http://www.digitalmedievalist.org/journal/1.1/robinson/ (accessed 10 September 2012).

Sahle, P. (2012). Kriterien für die Besprechung digitaler Editionen. Köln: Germany. v1.0. IDE, English version "Criteria for Reviewing Scholarly Digital Editions" forthcoming. http://www.i-d-e.de/aktivitaeten/reviews/kriterien-version-1 (accessed 23 October 2012),

Shillingsburg, P. (2006). *From Gutenberg to Google: Electronic Representations of Literary Texts*. Cambridge: Cambridge University Press.

Siemens, R., and M. Elkink, et al. (2010). Underpinnings of the Social Edition? A Narrative, 2004-9, for the Renaissance English Knowledgebase (REKn) and Professional Reading Environment (PReE) Projects. In McGann, J., and A. Stauffer, et al. (eds.), *Online Humanities Scholarship: The Shape of Things to Come*. Houston: Rice University Press. http://cnx.org/content/m34335/ (accessed 10 September 2012).

Sinclair, S., and G. Rockwell (2009). Between Language and Literature: Digital Text Exploration. In Lancashire, I. (ed.), *Teaching Language and Literature Online*. New York: Modern Language Association.

Current Research & Practice in Digital Archaeology

Watrall, Ethan
watrall@msu.edu
Michigan State University, United States of America

Graham, Shawn
Shawn_Graham@carleton.ca
Carleton University, Canada

Frey, Jon M.
freyjona@msu.edu
Michigan State University, United States of America

Schopieray, Scott
schopie1@msu.edu
Michigan State University, United States of America

Adams, Brian
adamsb@msu.edu
Michigan State University, United States of America

Brock, Terry P.
brockter@msu.edu
Michigan State University, United States of America

Wells, Joshua J.
jowells@iusb.edu
Indiana University, South Bend

Anderson, David G.
dander19@utk.edu
University of Tennessee, Knoxville

Yerka, Stephen J.
syerka@utk.edu
University of Tennessee, Knoxville

Kansa, Eric C.
ekansa@ischool.berkeley.edu
University of California, Berkeley

Whitcher Kansa, Sarah
skansa@alexandriaarchive.org
Alexandria Archive Institute

Noack Myers, Kelsey
kejmyers@indiana.edu
Indiana University, Bloomington

DeMuth, R. Carl
rcdemuth@gmail.com
Indiana University, Bloomington

Pett, Daniel
DPETT@thebritishmuseum.ac.uk
The British Museum

Session Abstract

Archaeology has a long history of innovative work with information and computing technology. While there are a small number examples in the late 1950s, the most influential comes courtesy of James Deetz's seminal work on Arikara ceramics. Carried out in the early 1960s, Deetz's project used the IBM704 mainframe at the MIT Computation Laboratory to discover "stylistic coherence" on over two thousand rim sherds from central South Dakota Medicine Crow site. Deetz's work was extremely important as it suggested that computers were excellent tools for statistical, typological, chronological, or stylistic analysis of large and complex sets of data (a hallmark of archaeology).

Since these early days, digital archaeology has remained intently focused on the analysis, interoperability, and preservation of digital data. By the mid-1980s, however, the personal computer had reached a point where they became effective tools for archaeological visualization and imagery. Desktop applications such as GIS, which allowed for the visualization, analysis, and modeling of socio-spatial data, and CAD, which facilitated the production of detailed and geometrically accurate archaeological maps at various scales without time-consuming redrafting, became central in the digital archaeological ecosystem.

In recent years, along with many other disciplines in the humanities and social sciences, archaeology is entering a new age in which information, computing, and communication technology is having transformative impact on all aspects of the field. The archaeological domains and activities in which digital approaches, methods, and technologies are relevant have grown well beyond the traditional trinity of data, GIS, and CAD. All aspects of research (including field and lab methods), teaching, outreach, publication, and scholarly communication are being impacted in new and unpredictable ways by "digital." Quite frankly, gone are the days in which digital archaeological methods were siloed off from the main body of scholarly practice. In many ways, one might argue that we have entered an age in which all archaeology is digital archaeology and all archaeologists are digital archaeologists.

In is within this context that this session will highlight a series of innovative projects and practices that represent the forefront of work in digital archaeology. Special attention has been made to highlight projects which represent a variety of domains within digital archaeology including digital data, public engagement, data & topic modeling, crowdsourcing, linked-open data, and digital fieldwork records management. All papers in the session also speak to the changed and changing nature of scholarly and professional practice within archaeology, addressing new approaches to collaboration, community engagement, citizen scholarship, cyberinfrastructure, preservation & access, capacity building, and sharing. A second, but no less important, goal of this session is to challenge the rather curious separation that exists between digital archaeology and the digital humanities by clearly placing the two domains parallel to one another and recognizing the fact that they both have much to learn from one another. The ultimate goal in this regard is to foster and support fruitful discussions and collaboration between digital archaeologists and digital humanists.

Topic Modeling Time and Space: Archaeological Datasets as Discourses

Graham, Shawn

Topic modeling is very popular at the moment in the digital humanities. A recent tutorial on getting started with this tool explains them as tools for extracting topics or injecting semantic meaning into vocabularies: "Topic models represent a family of computer programs that extract topics from texts. A topic to the computer is a list of words that occur in statistically meaningful ways. A text can be an email, a blog post, a book chapter, a journal article, a diary entry – that is, any kind of unstructured text" (Graham, Weingart, and Milligan 2012). In that tutorial, 'unstructured' means that there is no encoding in the text by which a computer can model any of its semantic meaning.

Archaeological datasets are rich, largely unstructured bodies of text. While there are examples of archaeological datasets that are coded with semantic meaning through xml and Text Encoding Initiative practices, many of these are done after the fact of excavation or collection. In the field, things can be rather different, and this material can be considered to be 'largely unstructured' despite the use of databases, controlled vocabulary, and other means to maintain standardized descriptions of what is excavated, collected, and analyzed. This is because of the human factor. Not all archaeologists are equally skilled. Not all data gets recorded according to the standards. Where some see few differences in a particular clay fabric type, others might see many, and vice versa. Archaeological custom might call a particular vessel type a 'casserole', thus suggesting a particular use, only because in the 19th century when that vessel type was first encountered it reminded the archaeologist of what was in his kitchen – there is no necessary correlation between what we as archaeologists call things and what those things were originally used for. Further, once data is recorded (and the site has been destroyed through the excavation process), we tend to analyze these materials in isolation. That is, we write our analyses based on all of the examples of a particular type,

rather than considering the interrelationships amongst the data found in the same context or locus. David Mimno in 2009 turned the tools of data analysis on the databases of household materials recovered and recorded room by room at Pompeii. He considered each room as a 'document' and the artefacts therein as the 'tokens' or 'words' within that document, for the purposes of topic modeling. The resulting 'topics' of this analysis are what he calls 'vocabularies' of object types which when taken together can suggest the mixture of functions particular rooms may have had in Pompeii. He writes, 'the purpose of this tool is not to show that topic modeling is the best tool for archaeological investigation, but that it is an appropriate tool that can provide a complement to human analysis....mathematically concrete in its biases'. The ‘casseroles’ of Pompeii turn out to have nothing to do with food preparation, in Mimno’s analysis.

To date, this is the only example of topic modeling applied to archaeological data. As such, it is novel in the digital humanities for applying the tools of data mining not to texts, but to things. In this paper, I explore the use of topic models on another rich archaeological dataset, the Portable Antiquities Scheme database in the UK. The Portable Antiquities Scheme is a project "to encourage the voluntary recording of archaeological objects found by members of the public in England and Wales". To date, there are over half a million unique records in the Scheme's database. I use topic modeling of this database to tease out archaeological patterns -the discourses of topic modeling, to use Ted Underwood's phrasing - over both time and space. In order to visualize these discourses, I map them both in geographic and relational space, using the network analysis program Gephi. The constellation of ideas (the resultant ‘topics’) that make up the various discourses in the data can be represented as nodes while the strengths of the associations suggested by the topic model can be represented as edges. This two-mode graph (words and ‘topics’ or ‘discourses’) can be queried for deeper structure. I look at the modularity of this graph to determine ‘communities’ of ideas or discourses. I then lay this network against real geographic space by time-slice to understand changes over time and space in the Portable Antiquities Scheme data. I agree with Mimno's suggestion that this is an appropriate tool for the digital archaeologist, but try to understand the limitations, caveats, and lessons for digital humanities more generally, from this application.

The Archaeological Resource Cataloging System (ARCS): An Open-Source Solution to Digitizing an Archaeological Archive

Frey, Jon M. | Adams, Brian | Schopieray, Scott

Over the past few decades, archaeologists have begun to realize the benefits of providing archival records in digital form. Whether information is collected electronically or digitized from pre-existing materials, digital archaeological data should be readily accessible from anywhere in the world. These developments have increased the productivity of scholars who no longer need to visit the actual archive and eased the strain on those projects that must accommodate visiting researchers in addition to their normal daily operations. On the other hand, the creation of digital archives has drastically impacted the ways in which we interact with documents and artifacts that form the basis of archaeological research. Financially constrained projects have lagged behind their better-funded peers in the process of digitization and dissemination of electronic records. As a result, instead of providing greater access to a wider range of archaeological data, the process of archival digitization runs the risk of further privileging the evidence of those surveys and excavations with greater financial resources. In addition, many of the existing digital archaeological archives concentrate upon artifacts to the point that the archaeologist’s field journal, arguably the most important evidence in establishing context, is rarely presented in its original form. Even where diverse forms of information are provided, a digital archive often encourages the study of objects and documents in isolation from one another and without the benefit of the institutional memory that often aids in their interpretation. While a traditional archive allows an individual to conduct their research through physical interaction with a number of different archival materials at once, often in the presence of those who discovered and prepared them, many digital archives simply rely on keyword searches to generate lists of electronic records that to the untrained eye appear to be of equal value as forms of evidence.

In this paper, we present the Archaeological Resource Cataloging System (ARCS), an open-source digital asset management application created for the Ohio State University Excavations at Isthmia in order to address these issues. Developed at Michigan State University through an NEH Digital Humanities Startup Grant, ARCS utilizes a web-based interface that allows authorized users to

upload and “tag” digital resources consistently according to generally accepted metadata standards that can be further refined to reflect any project’s unique terminology. These resources can then be searched, sorted into collections and connected to one another through the creation of virtual links without affecting the integrity of the original data. In this way the essential interrelatedness of the various forms of archaeological data is preserved in a flexible electronic format. In order to foster a better sense of community among researchers, each resource is also provided with a discussion tool that allows users to ask questions or identify mistakes, thereby making use of others’ knowledge and experience to cultivate the development of the dataset. In addition, because ARCS depends on the collective effort of a community of users, the system generates a permanent record of all additions and modifications of resources so that errors can be easily corrected and dependable users more clearly identified.

Perhaps most importantly, because it is an open-source application that relies on multiple users to develop and manage the digital assets of an excavation or survey, ARCS offers an affordable option for archaeological projects that lack a dedicated digital archivist or IT specialist. Digital data can be added as it is made available and, once uploaded, new resources can be linked to body of evidence that continues to grow in size and detail. In the end, ARCS not only retains the many benefits of more traditional research involving physical documents at an actual archive but in many ways also speeds and simplifies the process of archaeological investigation.

“All of Us Would Walk Together": Digital Cultural Heritage and the African American Past at Historic St. Mary's City, Maryland

Brock, Terry P.

In October 2012, Historic St. Mary's City (HSMC) launched a digital exhibit and social media campaign focused on the 19th-century component of their museum. HSMC, an archaeology and living history museum, has traditionally focused its 17th-century component, which was Maryland's first capital city. The digital exhibit, however, allowed the museum to begin interpreting additional centuries without disrupting the 17th-century landscape. Additionally, the digital exhibit, HSMC is able to develop an approach that focuses on public communication and engagement, allows for transparent research methods and interpretations, and provides flexibility when integrating the content into future programming and on-site exhibit. Through a combined approach of a content based digital exhibit, research blog, and social media, the digital exhibit, called "All of Us Would Walk Together", provides an example of digital archaeology that incorporates contemporary concepts of public archaeology through digital exhibitions and research methods. This paper will discuss how this has been put into action.

During the past few decades, community engagement has become a critical component of African American archaeology. Starting with the public excavations at the African American Burial Ground project in New York City, researchers have begun to incorporate local communities and descendants in the development and implementation of research projects and museum exhibits. Establishing a transparent, reciprocal, and pragmatic back-and-forth has become a valued and integral part of the research process for many archaeologists. Online tools have also been used as a means for public engagement, in particular at the Levi Jordan Plantation and Rosewood. More recently, archaeologists have adopted social media as a means for engaging communities and stakeholders, such as at the Michigan State University Campus Archaeology Program, Florida Public Archaeology Network, and at Mt. Vernon. Although each approach has highlighted different topics and methods for engagement, each has found the use of the web and digital social media to be a beneficial means of engaging the public.

At HSMC, the interpretation of the 19th century had not been the primary focus of the museum. This was particularly evident when the structures relating to the 19th century, including a manor home, its outbuildings, and a former duplex slave and tenant quarter, were physically moved to a different location in 1992 due to its conflict with the 17th-century interpretation. While the buildings continued to be used as a bed and breakfast, they were not used as an interpretive component of the museum, causing memory of the 19th century to be lost to the public. Recently, this story has begun to resurface, due to a number of factors at the museum. Included was the reacquisition of the manor home and outbuildings from the owners of the bed and breakfast, and funding opportunities to interpret the duplex quarter through a digital exhibit and a physical exhibit. In addition to interpreting the site, the goal of the exhibits was also to build a relationship with the African American community and to reinstate the 19th-century story into the public consciousness.

The digital exhibit consists of two components: a traditional exhibit space and social media. The exhibit space presents a number of webpages devoted to the interpretation of the historical and archaeological data that has already been analyzed. These pages trace the transition from slavery to freedom for the African Americans who lived on the site, and uses historical and archaeological

evidence to develop the narrative. Interspersed are links to blog posts that discuss how the evidence was gathered or used by researchers to draw the conclusions. Additionally, each exhibit page has a comment field, where the public can ask questions, offer their own interpretation, or provide additional commentary. This allows for two-way communication between the public and researchers, while also allowing the public to engage in the interpretive efforts. Lastly, these pages will be linked to the physical exhibit through the use of QR codes or augmented reality to provide more flexibility to the interpretive efforts at the site of the duplex. In doing so, the exhibit becomes flexible, transparent, and engaged physical space, fitting within the parameters of an engaged cultural heritage project.

The use of social media, through a blog, Twitter, and Facebook, adds extra depth to these efforts. The blog provides the most flexibility and transparency by allowing the exhibit space to be amended, added to, or modified in a transparent way. This provides a great deal of flexibility to the site that only a digital exhibit can provide. For example, if the research results in a major change to the exhibit, a blog post can be written to discuss the change and why it happened. In the exhibit, a link to this post can be added to demonstrate that the research process is a fluid, ongoing process. The blog and Twitter are also used to highlight research at comparable sites or examining comparable themes. This ties the archaeological work conducted at this site to larger themes in the discipline, and begins to build relationships with other institutions, while providing additional resources and access to new scholarship to the public. The blog and Twitter are both instrumental in the preservation and exhibit building process, as they make it more transparent: the public can watch and participate in the decisions about what will be included in the exhibit, in addition to understanding what types of constraints are placed on the construction of an exhibit. Twitter also allows lab and fieldwork to be shared in realtime with the public. Lastly, Twitter provides access to a larger, global network, particularly an African American network, a demographic that uses Twitter more than others. Facebook, on the other hand, is being used to connect with HSMC's current online fans through its official Facebook account.

The project itself has a high set of goals, and is approaching it with a multifaceted approach. While the digital component is a crucial step, it is only part of a larger program of public engagement. For example, HSMC has been actively soliciting feedback from local community members and seeking to engage them through the formation of an advisory board. This reiterates an important tenant of engaging in digital public archaeology: that one cannot rely solely on one approach. Nonetheless, the use of the digital space does provide us with additional exhibit and interpretive space, and gives us a great deal of flexibility when dealing with the public, research, and presentation. Most importantly, the use of the digital arena allows this all to be transparent, reciprocal, and dynamic.

An Introduction to the Practices and Initial Findings of the Digital Index of North American Archaeology (DINAA)

Wells, Joshua J. | Anderson, David G. | Yerka, Stephen J. | Kansa, Eric C. | Whitcher Kansa, Sarah | Noack Myers, Kelsey | DeMuth, R. Carl

The Digital Index of North American Archaeology (DINAA) is a project to create interoperability models for archaeological site databases in the eastern United States, funded by the National Science Foundation (#1216810 & #1217240). The core research team consists of researchers from the Department of Anthropology and Archaeological Research Laboratory at the University of Tennessee, the Alexandria Archive Institute, and the Anthropology and Informatics programs at Indiana University. Open Context (http://opencontext.org) will be used as the primary platform for data dissemination for this project. Our aims are to work with the databases held by State Historic Preservation offices and allied federal and tribal agencies in Eastern North America, with the goal of linking data across state lines for research and management purposes. Redacted of sensitive items, such as site location, data linkages will promote extension and reuse by government personnel in state and federal agencies, and domestic and international researchers. The project will mint stable Web-URIs for each site record, and in doing so, we will help lay the foundations for future Linked Open Data applications in North American archaeology, architectural history, and historical studies.

This project repurposes government curated datasets to support innovative humanistic and social science research. Governmental archaeological site files in North America are important loci for documentary information on known archaeological sites. Their most basic function is to contain data about site types and information quality pursuant to heritage preservation legislation at the federal level, but potentially state and local levels as well. However, as a matter of practice these files, often as relational databases, contain many other data fields that describe important archaeological findings, and other data that serve environmental and bureaucratic functions for management and protection of heritage resources. The

ways in which data about archaeological sites are recorded and communicated have an important origin in theoretical models about past behavior, and also have important implications on the professional comprehension of the data at large and the use of the data to rank planning and preservation priorities.

Efforts to collect and compile archaeological data have a long history, and information about archaeological sites and collections is maintained by every state and territory. Only rarely, however, have these data been compiled and examined at large geographic scales, especially those crosscutting state lines, and never to the extent and for the research and management purposes proposed in this project. Data from some 15 to 20 states (>half a million sites) east of the Mississippi will be integrated with a common ontology, based on existing standards, and adapted in collaboration with researchers and government personnel in state and federal agencies. The ontology will classify site files according to cultural affiliation and chronology as well as agency assessments of historical significance.

Linkage of site file and other datasets will facilitate studies of past human adaptation spanning large areas, and lead to greater collaboration between archaeologists and scientists in other disciplines. As examples, the linkage of archaeological data at broad new scales will permit, for the first time, the exploration of exciting new research topics, such as how the human populations in North America responded to climate change, population growth, and/or anthropogenic environmental issues over the past 13,000 years.

The availability of output online in the form of maps and data tables (at significantly reduced spatial resolution, to protect sensitive locations) will enhance public awareness, education, and appreciation for scientific research in general and archaeology in particular. The demonstration that primary archaeological data can be integrated and used to address fundamental questions at such scales will stimulate similar efforts worldwide. Finally, by creating translating routines rather than dictating procedures, this project will foster archaeological cooperation through cyberinfrastructure with a high ratio of benefits to costs.

The project helps achieve broader archaeological concerns regarding professional data management training, research ethics and outreach education. It will foster novel networking and data integration among multiple partners, as well as research and educational activities across multiple disciplines and geopolitical boundaries. Publicly accessible data products, at coarse scales to preserve site location security, will also be available for download and reuse as shapefiles, CSV data tables, and RDF and N3 triples for other Web and desktop applications, including desktop GIS investigation. The project will provide specific instructions for the open source GIS applications QGIS gvSIG, and uDIG, and the widely used proprietary application ArcGIS, in order to foster education in geographic information science within archaeology and related disciplines. The project will fund graduate and undergraduate students, and will assist their training in critical information management skills for the 21st century. The project addresses head-on a major challenge facing research communities worldwide: how to link disconnected and incompatible data systems in such a way that the combined data are useful for important scientific research.

The integration of site file data at continental scales in a new and unique informational infrastructure will allow, for the first time, the exploration of the North American archaeological record across multiple temporal periods and geographic regions. The geographic scale and extent of data integration proposed is currently unprecedented in American archaeology yet, we believe we have demonstrated that it is readily achievable. With proper attention these data have the potential for continued growth as developed by the professional archaeological community, and as the resulting datasets become more inclusive they may transform the practice of our profession.

The Portable Antiquities Scheme: a new(ish) model for recording public discovery

Pett, Daniel

The Portable Antiquities Scheme (PAS) was inaugurated in 1997, following the revision of the ancient law of Treasure Trove and the subsequent implementation of the Treasure Act in 1996. This project has been through several funding phases all using public money and encourages the voluntary recording of archaeological objects discovered by members of the public in England and Wales (Scotland is subject to different legislation) and items that meet the stipulations of the Treasure Act.

One of the key pillars of the PAS has been its digital presence, which has now been online in some form for over 13 years and this paper will discuss the impact that the digital arm of the project has had on a national and international audience. The PAS has been hailed by many as a model of public archaeological engagement and decried by others for allowing the mining of the archaeological record for personal gain; however this paper will show some of the PAS' many successes. For example the PAS has had contributions from over 20,000 individuals worldwide, a new facility has been created for contributors to record their own discoveries through taxonomy driven interfaces and over 350 projects utilise these data for informing their research - for example Oxford University's EngLaID project. The project has also absorbed and enhanced

internationally renowned resources such as Oxford University's Celtic Coin Index and Cardiff University's Iron Age and Roman coins of Wales database to provide the largest national, single search node for the study of Roman and Celtic coinage.

The project website records a huge array of metadata about objects, that we often have but one chance to record; images, textual description, measurements, spatial data and user generated comments and audit logs. Over 820,000 objects have been recorded on the PAS database and these are made available for all to view, comment and reuse within their own research or their own websites under a Creative Commons 'by attribution share-alike' licence. This liberal approach to licensing content has not been seen widely in the UK and European archaeological sector and the launch of the PAS' Staffordshire Hoard microsite in September 2009 showed how well received this approach would be. Over 250,000 unique visitors in one day used the innovative microsite to learn more about the amazing Anglo-Saxon hoard and many more viewed the images that had been disseminated via Flickr. The PAS has also utilised and archived social media platforms with varying degrees of success and is a major case study within Lorna Richardson's forthcoming PhD and complements that of its host organisation (the British Museum.)

This paper will show how the PAS website impacts on the public with specific reference to stories of international interest – such as the Staffordshire Hoard, the Frome hoard of 52,503 Roman coins, the Crosby Garrett helmet and the Staffordshire Moorlands Patera. It will also discuss how these successes have been reached on a minimal digital budget (less than £5000 per annum) via the use of open source technology and through the buy in of its audience. The website has been internationally recognised, winning the prestigious Museums and the Web 'Best of Web' award for 'research or online collection' (recent winners include the V&A and the Metropolitan Museum of Art.)

The author will demonstrate how a variety of digital techniques that have been employed to complement and enhance data collected via our network of archaeologists, volunteers and citizen scientists; for example geo enrichment through the use of Yahoo!, Geonames, Pleiades and Google Maps, text extraction through OpenCalais and Autonomy, the implementation of a Solr based faceted multi-core search engine and also how a wide variety of application programming interfaces (api) have been employed throughout the site. It will also show how the author has been influenced by ground breaking projects such as Open Context, Pleiades and Pelagios and a variety of Museum sector projects.

The paper will also touch on the recent steps towards providing the PAS data through linked data (for example the release of over 50,000 annotations for Pelagios), towards integrating linked data into the site (for example dbpedia enriched resources) and the CIDOC-CRM mapping process. It will also discuss 3D , computed tomography and PTM/RTI imaging projects conducted by Brighton and Southampton universities that have used PAS sourced objects and hoards of coins to provide research material. If time allows, the paper will also demonstrate the success achieved through the use of Flickr as a disseminator for a wide variety of archaeological images; how the author has leveraged news sources such as the Guardian and UK parliamentary records for background debate on the portable antiquities debate (ethical and fiscal) and how it has impacted on the public psyche within the UK and further afield.

msu.seum: A Location Based Mobile Application for Exploring the Cultural Heritage and Archaeology of Michigan State University

Watrall, Ethan

The spaces we inhabit and interact with on a daily basis are made up of layers of cultural activity that are, quite literally, built up over time. While museum exhibits, historical and archaeological narratives, and public archaeology programs can communicate this cultural heritage, they do not generally allow for rich, place-based, and individually driven exploration by the public. In addition, museum exhibits rarely explore the binary nature of material culture and the preserved record of human activity: the presented information about material culture and the process by which scholarly research has reached those conclusions. In short, the scholarly narrative of material culture, cultural heritage, and archaeology is often hidden from public consumption.

In recent years, mobile devices as well as the development and maturation of augmented reality (broadly construed) have offered both platforms and models for mobile cultural heritage applications to address the former issue. Mobile applications such as The Museum of London's Streetmuseum Londinium, Florida Public Archaeology Network's Destination: Civil War, and the forthcoming CHESS Acropolis Museum mobile application facilitate public interaction with cultural heritage and archaeology in a place-based context. However, the latter issue, the scholarly narrative of the process by which cultural heritage and archaeological information was uncovered and information was generated, is often left unaddressed and unexplored in mobile cultural heritage applications.

It is within this context that this paper will introduce and explore msu.seum. Developed during the 2011 Michigan State University Cultural Heritage Informatics Fieldschool directed by Ethan Watrall in collaboration with the Michigan State University Campus Archaeology Program and Campus Archaeology Fieldschool, msu.seum is a mobile application that allows users to interact with the rich cultural archaeological heritage of the historic Michigan State University campus, and explore the processes by which the Campus Archaeology Program helped reveal it. Building on the idea of "campus as museum," msu.seum connects cultural heritage directly to place, highlighting both what is known about the MSU Campus and the scholarly narrative of the associated archaeological and historical research. msu.seum functions less like a "check-in" app and more like a rich, exploration-based museum tour guide that exposes the cultural heritage and archaeology of Michigan State University to staff, faculty, students, alumni, and the general public.

Currently available for iOS, msu.seum's content is organized into a series of thematic "exhibits" that reflect the development of Michigan State University: Beginnings: 1855-1870, Foundation: 1870-1900, Expansion: 1900-1925, and Legacy: 1925-1955. A fifth exhibit, Discovery: Archaeology, explores more recent research and excavations carried out by the MSU Campus Archaeology Program. The idea behind this exhibit-based model is that the user's experience mirrors the physical layout of a museum. Instead of being contained within a physical structure, however, exhibits are distributed around the MSU campus. Each exhibit contains a collection of locations that users are free to visit and experience at their leisure (either within the context of a visit to the Michigan State University campus or during their regular, daily campus activities)

Each exhibit location contains information and rich media (video, audio, and imagery) about that location and well as the narrative about the work (excavation, survey, etc) carried out at the location by the MSU Campus Archaeology Program.

The paper will explore the features of the application itself, as well as the process by which it was designed. Special attention will be paid to discussing the unique and highly collaborative environment in which msu.seum was developed.

Beyond its value as a tool to allow the public to interact with and explore the cultural heritage and archaeology of the Michigan State University (and the associated scholarly narrative as to how that knowledge was developed), we feel very strongly that msu.seum can act not only as a model for the design and development of other mobile and place-based campus cultural heritage applications, but for mobile and place-based cultural heritage applications in general.

Papers

Freedom and Flow: A New Approach to Visualizing Poetry

Abdul-Rahman, Alfie
alfie.abdulrahman@oerc.ox.ac.uk
University of Oxford, UK

Coles, Katharine
k.coles@english.utah.edu
University of Utah, USA

Lein, Julie
jkglein@gmail.com
University of Utah, USA

Wynne, Martin
martin.wynne@it.ox.ac.uk
University of Oxford, UK

Much research in the digital humanities has been data-driven and quantitative, and while these methodologies and projects have led to substantial scholarship and vastly improved access to texts, such approaches can also risk diverting us from established values and concerns within the humanities. Under an international Digging Into Data Challenge grant, our team of computer scientists, a linguist, and poet/scholars from the University of Oxford and the University of Utah have been working to move beyond counting to create, through computation and visualization, a richer understanding of how poems work: one that embraces qualitative as well as quantitative components and that engages the perspectives and research needs specific to the humanities in general and to literature, especially poetry, in particular. This paper will describe the challenges and opportunities presented by such multidisciplinary collaboration, and demonstrate how the new visualization tool we are developing provides literary scholars freedom to explore individual poems and bodies of poetry in ways traditional scholarship and other text analysis software cannot. Specifically, we will show how our new approach, by treating poems as complex dynamic systems, opens new interpretive directions via the metaphor of *flow*.

Dixon (2012, 200) notes that within digital humanities, 'researchers often appear to be dragged towards more scientific interpretations and methods,' observing that reasons for these shifts might arise from researchers' own inclinations, but also as a result of the very tools often utilized in this new field. Tools not only influence scientific or humanist orientations, however; they can also augment or diminish our understanding of the subjects we study. A quarter century ago, in their seminal book *Understanding Computers and Cognition*, Winograd and Flores (1986, xi) perceived that 'in designing tools we are designing ways of being'. And for this reason, scholars like Drucker (2011) and Schmidt (2012) have argued forcefully that humanists should actively participate in software design, bringing our disciplinary perspectives to bear on shaping the digital tools we want to use. In the case of literature, available tools have oversimplified poetry, requiring scholars to contort and contract their scholarship. The closest existing tool to the one we are working on is Myopia, presented by Laura Mandell at last summer's Digital Humanities conference (Chaturvedi et al. 2012). Myopia deals with more complexity in poetry than other programs, attending to meter, sound, syntax, and other poetic devices like metaphor and personification. However, it does so through coding structures developed for other disciplines, rather than incorporating into its design the disciplinary perspective we seek. Fortunately, the computer scientists in our group have not only been open to the input of the poets, they have persistently sought it, leading the poets to new insights about poetry and language. At the same time, the interests and needs the poets identified have pushed the visualization scientists to make groundbreaking advances in their own field.

Poetry scholars see poems as living and relational. The strong reader produces innovative interpretations even of poems that may already have been subject to many previous interpretations by looking at how language is operating across multiple poetic elements, from rhyme and meter, to the use of such figures as metaphor and metonymy, to tone, to the emotional affect that may inhere in specific words. Though we can list these elements as if they are discrete, of course they are not: the emotional affect of a word will influence the poem's tone, for example, just as metaphors and sonic features work to create and enhance emotion. To adequately address these various elements, any visualization tool must treat poems as multi-dimensional. They not only exist in space but also exist in and change through time, which operates both forward (left-to-right across the line and down the page) and backward. Though almost every poetic device works in this way, the simplest example of poetry's double movement through time is end rhyme: if two quatrains of a Shakespearean sonnet rhyme abba/abba, each recurrence of the rhyme once it is established not only moves the reader forward, through anticipation, toward the next occurrence, but also sends the reader backward by evoking all previous occurrences. Such nuanced dynamism presents individual poems as highly complex and uncertain, with each word, each sound, each poetic figure and device

relating and responding to one another. It demands that visualizations be time dependent and able to handle various poetic elements not only independently but relationally, something we believe has not been accomplished before.

Our current visualization tool, PoemViewer, focuses on the specific but complex task of tracing the movement of sound through the defined space of almost any poem or poem fragment (Abdul-Rahman et al. 2013). In response to a questionnaire produced by the Oxford team, the poets determined that this single poetic device involves twenty-six distinct variables. As Ware (2012) explains, though, visual design for such high dimensional data is challenging due to limitations of human visual perception and the number of visual channels available. For these reasons, visualizations—excluding approaches like parallel coordinates and scatterplot matrices (Inselberg and Dimsdale 1990; Wilkinson et al. 2005)—typically are constrained to no more than nine dimensions (Kindlmann 2004). PoemViewer encodes twenty-six dimensions using d3.js (Bostock, Ogievetsky & Heer 2011; Bostock 2013), a JavaScript library that utilizes SVG, HTML and CSS for data visualization. PoemViewer is a web-based application that can run on all modern browsers. It uses common existing linguistic models and tools, such as the International Phonetic Alphabet (1999) for classifying sounds, and the CLAWS (UCREL 2013) part-of-speech tagger for morpho-syntactic word classes.

Still, not all poetic features need to be visualized at once. In order to create visualizations relevant to particular interests in a particular poem (and to avoid becoming visually overwhelmed), readers can choose which linguistic variables to view—in isolation or relation—at any given time (see Fig. 1). For instance, our interactive tool can identify assonance (the recurrence of vowel sounds) and consonance (the recurrence of consonant sounds) as they occur within a specific poem in time. But by tracking each phonetic placement as it is formed in the mouth, our software can also reveal how sounds relate to and change each other; it can, for example, reveal modulation between long and short vowels in a poem like Louise Bogan's "Night" (1968, 130). During our presentation we will share additional images, including more screenshots of the tool and its components (menus, legends, etc.) as well as visualizations it can produce—both of singles poems and in comparing multiple poetic and non-poetic texts. Of specific interest are insights drawn from our tracking of how sounds are formed in the mouth and about how poetry behaves differently from prose in its use of sound.

Among the other breakthroughs we have gained through this work on sound is the development of a powerful metaphor for poetry that will move us forward as we continue our work. Specifically, our work on this tool has led us to think of the poem as *a fluid (or fluids) moving via its linguistic elements, devices and figures through a self-defined space*. Given that a poem, unlike a fluid, has no material analogue in the world, one of our challenges will involve identifying and digitizing those specific elements of the poem in addition to sound that we believe make it behave metaphorically as a fluid does, with the goal of adding new capabilities to our existing model. Elements that are of particular interest to poetry scholars include, among others, syntax, meter, and figure. Eventually, we expect to extend our original tools to analyze different poetic features as they interact within poems, across bodies of work by individual poets, and finally across large numbers of poems representing different poets in different literary periods. We anticipate applying this software to large poetry corpora to detect formal patterns, and conduct explorations such as, for example, recognizing, examining, and beginning to explain widespread poetic responses to historical events, social phenomena, or technological invention.

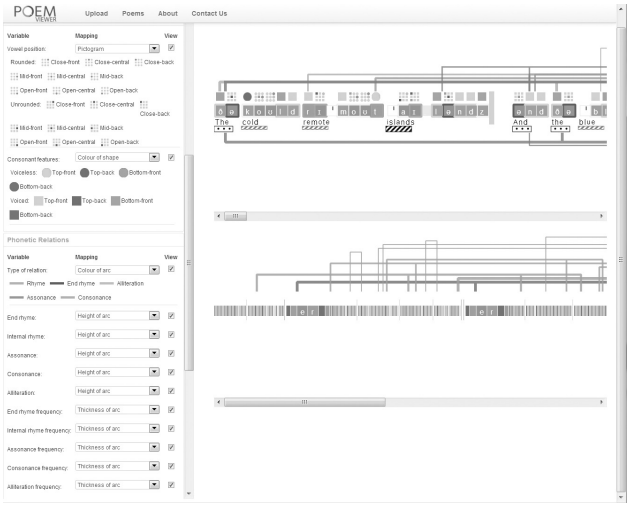


Figure 1:
Screenshot of PoemViewer visualization of Night by Louise Bogan © 2013 Alfie Abdul-Rahman, Oxford e-Research Centre

Funding

This work was supported by a Digging Into Data Challenge grant: in the US, by the National Endowment for the Humanities; and in the UK by the Arts and Humanities Research Council, Economic and Social Research Council, and JISC. We would like to acknowledge Min Chen, Christopher Johnson, Eamonn Maguire, and Miriah Meyer for their collaboration on this project.

References

Abdul-Rahman, A., J. Lein, K. Coles, E. Maguire, M. Meyer, M. Wynne, A. Trefethen, C. R. Johnson, and M. Chen (2013). Rule-based Visual Mappings—with a Case Study on Poetry Visualization. Paper conditionally accepted to *EuroVis*, Leipzig, Germany.

Bogan, L. (1968). "Night" *The Blue Estuaries: poems 1923-1968*. New York: Farrar, Straus, and Giroux.

Bostock, M. (2013). Data-Driven Documents. http://d3js.org/ (accessed 4 February 2013).

Bostock, M., V. Ogievetsky, and J. Heer (2011). 'D3: Data-Driven Documents. Transactions on Visualization and Computer Graphics'. *Proceedings of InfoVis 2011 IEEE*, 2301-2309. http://vis.stanford.edu/papers/d3 (accessed 7 January 2013).

Chaturvedi, M., G. Gannod, L. Mandell, H. Armstrong, and E. Hodgson (2012). Myopia: A Visualization Tool in Support of Close Reading. *Digital Humanities 2012*. University of Hamburg, July 2012. http://lecture2go.uni-hamburg.de/konferenzen/-/k/13930 (accessed 6 September 2012).

Dixon, D. (2012). Analysis Tool or Research Methodology: Is there an Epistemology for Patterns?. In Berry, D. M. (ed). *Understanding Digital Humanities*. New York: Palgrave Macmillan. 191-209.

Drucker, J. (2011). Humanities Approaches to Graphical Display. *Digital Humanities Quarterly* 5:1 http://digitalhumanities.org/dhq/vol/5/1/000091/000091.html (accessed 8 September 2012).

Inselberg, A., and B. Dimsdale (1990). 'Parallel coordinates: A tool for visualizing multi-dimensional geometry'. *Proceedings of the 1st Conference on Visualization* held in 1990. 361-378.

International Phonetic Association. (1999). *Handbook of the International Phonetic Association: A Guide to the Use of the International Phonetic Alphabet*. Cambridge: Cambridge University Press.

Kindlmann, G. (2004). *Superquadric tensor glyphs, Proceedings of the 6th Joint Eurographics- IEEE TCVG Conference on Visualization* held in 2004. 147-154.

Schmidt, B. (2012). 'Reading Genres: Exploring Massive Digital Collections from the Top Down'. in *Big Data and Uncertainty in the Humanities*. held at Institute for Digital Research in the Humanities. Lawrence, KS.

UCREL. *Claws Part-of-Speech Tagger for English*. http://ucrel.lancs.ac.uk/claws/ (accessed 7 January 2013).

Ware, C. (2012). *Information Visualization*, 3rd edn. San Francisco: Morgan Kaufmann.

Winograd, T., and F. Flores (1986). *Understanding Computers and Cognition: A New Foundation for Design*. Norwood, NJ: Ablex.

Dyadic pulsations as a signature of sustainability in correspondence networks

Aeschbach, Michael
michael.aeschbach@unil.ch
University of Lausanne, Switzerland

Brandt, Pierre-Yves
pierre-yves.brandt@unil.ch
University of Lausanne, Switzerland

Kaplan, Frédéric
frederic.kaplan@epfl.ch
Ecole Polytechnique Federale de Lausanne (EPFL), Switzerland

In this paper, we introduce the concept of dyadic pulsations as a measure of sustainability in online discussion groups. Dyadic pulsations correspond to new communication exchanges occurring between two participants in a discussion group. A group that continuously integrates new participants in the on-going conversation is characterized by a steady dyadic pulsation rhythm. On the contrary, groups that either pursue close conversation or unilateral communication have no or very little dyadic pulsations. We show on two examples taken from Usenet discussion groups, that dyadic pulsations permit to anticipate future bursts in response delay time which are signs of group discussion collapses. We discuss ways of making this measure resilient to spam and other common algorithmic production that pollutes real discussions.

Can a discussion group be characterized by looking solely at the interaction patterns of its participants? Symmetrically, can the patterns of interaction of a given participant identify its role in a discussion group? Can we predict from these patterns the evolution of the interaction inside a group, spotting for instance the early signs of a group decomposition process? Our research aims to establish a new mathematical approach to distinguish between different types of discussion group participants as well as between different types of discussion groups,

both with their typical ways of interacting (interaction "signatures") and life cycles.

The analysis of interaction dynamics has recently received an increased focus of attention in the network science community. Mathematical methods (Newman, Barabási and Watts 2006) have been used to identify signatures characterizing the mode of exchanges of famous scholars, comparing for instance patterns in the correspondence of Charles Darwin and Albert Einstein (Oliveira and Barabási 2005). Fluctuation patterns and delays in letter responses indicate prioritization strategies that can be modeled and simulated. These methods permit also to draw comparisons with modern forms of electronic exchanges where similar patterns, corresponding to universal scaling laws (Barabási 2005; Bunde, Eichner, Havlin and Kantelhardt 2004), can be found. Interestingly, all these analyses can be conducted without considering the semantic or pragmatic nature of the exchanges.

Patterns in correspondences networks are of great interest for research in Digital Humanities. For instance, the Stanford's Republic of Letters project (http://republicofletters.stanford.edu) use "big data" and "distant reading" approaches to offer new visualization tools and test various hypotheses about the Enlightenment. However, mathematical analysis of such networks is not yet common in the Digital Humanities community.

Drawing on research on social networks (Wasserman and Faust 2009 [1994]) and computer mediated communication (Smith and Kollock 1999; Turner, Smith, Fisher and Welser 2005; Welser, Gleave, Fisher and Smith 2007), we analyze the activity and correspondence pattern of participants in Usenet newsgroups about religion and spirituality. Our paper reports an on-going analysis of large data set that consists of more than 1.5 million unique Usenet messages. Usenet is one of the Internet's oldest discussion systems still in widespread use. Unlike more recent platforms like Facebook and Twitter, Usenet hasn't stored its messages in a single central location and offers due to its open nature easier access to its data. Furthermore, Usenet's threaded conversations are organized by topics, with the advantage of allowing comparisons between topics.

In this paper, we introduce the concept of dyadic pulsation (DP) as a complementary measure to reply time for measuring the vitality of a given group. In our representation, a pulsation corresponds to the creation of new communication dyads, i.e. the first direct communication between two users A and B. When B first replies to A, a pulsation of type *A* (for asymmetric) is emitted. When A replies again to a message of B, a second pulsation of type *M* (for mutual) is produced. A group that continuously integrates new members in the on-going conversation is characterized by a steady dyadic pulsation rhythm, mixing type *A* and type *M* pulsations. On the contrary, groups that either pursue close conversation or unilateral communication (e.g. news feeds, announces without discussion) have no or very little dyadic pulsations.

Our working hypothesis is that evolutions in the pulsation rhythms are earlier predictors of the evolution of group dynamics. Figure A shows an example of a group maintaining a good average response time for a long period followed by an apparently unanticipated explosion of response time. At some point group members simply stop to answer timely to the messages of the discussion group. Interestingly, although no anticipated sign of this evolution could be spotted in the delay time graph, the pulsation graphs shows a progressive reduction of the frequency on the creation of new communication dyads. The grey lines in the lower bar of show pulsations of type A while the black lines show pulsations of type M. Figure B shows the change in delay time after the appearance of a succession of type *A* and type *M* pulsations, indicators for the formation of new links between discussion participants.

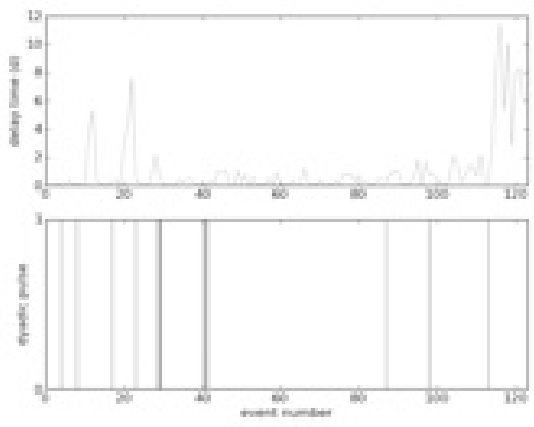


Figure A and Figure B:

For dyadic pulsations to be a reliable predictor, it is mandatory that they are resistant to the various forms of spam common to digital communication. This is one reason for the distinction of the two types of dyadic pulsations. Indeed, dyadic pulsations of type A can result of spam bots posting messages to a group. However, it is very unlikely that any real user answers to those messages. Thus, the presence of type M pulsations guarantees that new correspondence partners entered the group studied.

We are conducting a larger study to test the relevance of this measure in the particular case of different online discussion groups related to topics about religion and spiritualty. Our hope is to validate the hypothesis that dyadic pulsations on a group level permit to distinguish between different modes in a discussion group's life cycle.

More generally, we believe that this measure can be relevant to characterize the rises and declines of activity in correspondence networks, including literary correspondence networks.

References

Barabási, A.-L. (2005). The origin of bursts and heavy tails in human dynamics. *Nature*, 435(7039), 207–211. doi:10.1038/nature03459.

Bunde, A., J. F. Eichner, S. Havlin, and J. W. Kantelhardt (2004). Return intervals of rare events in records with long-term persistence. *Physica A 342*. 308–314.

Newman, M., A.-L. Barabási, and D. J.Watts (eds.). (2006). The Structure and Dynamics of Networks. *Princeton Studies in Complexity*. Princeton: Princeton University Press.

Oliveira, J. G., and A.-L. Barabási (2005). Human dynamics: Darwin and Einstein correspondence patterns. *Nature*. *437.7063*. 1251–1251. doi:10.1038/4371251a.

Smith, M. A., and P. Kollock (eds.). (1999). *Communities in Cyberspace*. London: Routledge.

Turner, T. C., M. A. Smith, D. Fisher, and H. T. Welser (2005). Picturing Usenet: Mapping Computer-Mediated Collective Action. *Journal of Computer-Mediated Communication*. 10.4. doi:10.1111/j.1083-6101.2005.tb00270.x

Wasserman, S., and K. Faust (2009). Social Network Analysis: Methods and Applications. Cambridge: Cambridge University Press.

Welser, H. T., E. Gleave, D. Fisher, and M. Smith (2007). Visualizing the Signatures of Social Roles in Online Discussion Groups. *Journal of Social Structure*. 8. http://www.cmu.edu/joss/content/articles/volume8/Welser/

An Evaluation of the Involvement of General Users in a Cultural Heritage Collection

Agosti, Maristella
agosti@dei.unipd.it
University of Padua, Italy

Benfante, Lucio
benfante@dei.unipd.it
University of Padua, Italy

Manfioletti, Marta
manfioletti@dei.unipd.it
University of Padua, Italy

Orio, Nicola
orio@dei.unipd.it
University of Padua, Italy

Ponchia, Chiara
chiara.ponchia1@studenti.unipd.it
University of Padua, Italy

Introduction

Digital tools are becoming increasingly important research aid in the humanities. Scholars in different disciplines related to cultural heritage are currently exploiting digital resources at different levels, including the simple storage of digital acquisitions of cultural objects in an online collection, the 3D rendering of complete sites, the use of advanced search functions to browse multimedia collections, and the possibility of annotating the digital objects or personalizing interaction with the system.[cit]

Once a digital system has been populated by researchers, a natural step is opening it to a wider public, which is enabled to access culturally relevant content. This extension poses a number of issues that are strictly related to the interest a specialized collection can raise. In this paper we present the results of a user study carried out over two years on two user groups: specialists in the domain and members of the general public.

Background

In 2003, we designed and have since continued to develop a digital archive (Agosti et al., 2003) of illuminated manuscripts called IPSA (*Imaginum Patavinae Scientiae Archivum*). The design was carried out with a user-centered approach, according to the user requirements of professional researchers. The focus of the project was on a collection of scientific manuscripts of Medieval and Renaissance periods, which were influenced by the new spread of scientific culture that saw Padua as one of its main centers. The IPSA collection includes herbals and astrological manuscripts. After scholars had used IPSA as a research tool for a number of years, in 2010 we started a new project to open up IPSA to different kinds of users among the general public. This effort aims at promoting the interest towards

ancient manuscript and illumination but, at the same time, is intended as a case study for the dissemination of scientific research in the humanities.

To this end we started to re-evaluate IPSA functions with different groups of users that could have an interest in the IPSA content: students in history of art, students in archival science, and researchers in humanities (but not in history of illumination). In this paper we focus on general trends in the evaluation, with the aim of generalizing the outcomes also to other collections (Sweetnam et al., 2012). Although IPSA is actually restricted online, a subset of the collection will be freely available online once additional functions will be in place.

Methods

Evaluation was carried out during the academic year 2011-2012, using a task-oriented approach and based on the triptych model of interaction (Fuhr et al., 2007), which considers a system consisting of three elements: collection, technical infrastructure, and people for which the collection is built. Participants were asked to perform simple research tasks on the digital collection, requiring about one hour of interaction that involved the use of the main IPSA functions. A total of about 60 users divided into three groups were involved in the evaluation.

Each user group participated to two evaluation sessions which were two weeks apart. This time span was needed because we decided to implement the main suggestions gathered during the first session in the IPSA interface and functions for the second session. In particular, the interface provided clearer contextual information during image analysis and the query results were presented as a wall of image, thus in a way more familiar to the general public. Moreover, the second evaluation session was introduced by a short lecture on the research methods adopted in history of illumination. The goal was to improve users' motivation by giving them direct feedback of their suggestions on the interface and, at the same time, to raise their interest in the collection. Users' comments were gathered using both questionnaires on user satisfaction and open-form interviews. The evaluation process ended with a focused-group discussion on the main aspects of the IPSA functions. Comments were also gathered while users were performing the tasks.

Outcomes

The evaluation highlighted a number of common characteristics on how non-specialized users consider the IPSA digital archive, and gave some clues to possible threats in proposing to the general public a digital archive that has been developed for professional users. Firstly, all users showed an appreciation on the quality of the digital content, yet they do not express a sustained interested towards the collection. Although almost all users commented on the beauty of the images, only a few of them continued browsing the digital archive for mere personal interest after they had finished the tasks proposed for evaluation.

Moreover, a large number of users focused their comments on details on the user interface, probably because they were not comfortable in providing comments on a subject they were not acquainted with — as they have no background in history of illuminations. Thus, the evaluation suffered from the "bike-shed effect", which is well-known in software design and is related to the inverse proportionality between the importance of a subject and the time spent discussing it.

A third outcome regards users' requests to provide more background information on the digital content. Since it was developed for domain specialists, who are mainly interested in images, IPSA provides only a basic set of descriptive metadata. This information, although relevant and congruent with library standards, was considered insufficient. Users asked for more involvement with the digital content by being provided with access to additional information, such as the research results produced by specialists. Furthermore, we found a substantial difference in user involvement when evaluation sessions were introduced by a lecture on the methods and aims in history of illumination.

During the final discussion, we stimulated users to suggest additional functions that may improve their enjoyment of the digital collection. As expected, most suggestions were about providing IPSA with functions — in particular on searching — that are usually available in popular web services and social networks. In general, the evaluation was biased by a continuous comparison with systems not expressly developed for cultural heritage collections, notwithstanding the clear differences in the selection of the content, the quality of completeness of representation between general-purpose systems and specialized collections such as IPSA.

Discussion

A comparison of the results of this evaluation with the comments gathered by domain specialists after using IPSA reveals that the primary aspect to take into account is the substantial difference in user interest towards — for instance — digital collections of artistic images. This difference exists also between domain specialists and scholars and students in related domains. This aspect affects all the main outcomes described in the previous section. A marginal interest towards the illuminations obviously explains the

relative short interaction, which for most users only lasted for the length of the evaluation session, even though they were free to continue exploring the collection. At the same time, the requirement of additional content — extending the simple collection of images and their metadata — may also be an expression of an insufficient involvement towards the content and, to a certain extant, the focus on minor details when evaluating the interface.

These outcomes are completely in line with the trends in dissemination of cultural heritage. The application of 3D technology to interact with digital artifacts and navigate inside virtual spaces (Koller et al. 2010), the development of serious games (Falk Anderson et al, 2010) for dissemination purposes, the increasing exploitation of portable and interactive devices — including users' portable devices — all suggest that the cultural content itself is not sufficient to raise interest among the general public.

We believe that all these strategies, although not always paired by extensive user studies on their actual effectiveness, are useful for promoting cultural heritage and improving enjoyment of digital collections even in the presence of marginal user interest.

However, an important outcome of our user study shows a direction that may need further exploration. A relevant part of the participants in our user study showed more interest towards the research process carried out by scholars than towards the subject of the research. That is, users seemed much more involved in understanding the different steps of the research work on illuminated manuscripts, its tools, the motivations of the choices made by scholars, the use of different sources for providing evidence of new hypotheses. This consideration suggests a shift in the focus of a digital collection: from a system for disseminating cultural content to a system for involving general users in goals, methods and results of scientific research on the cultural content.

Such a shift requires a novel design of digital systems in order to provide access to a variety of additional content — sources, measurements, analyses — that is routinely used by scholars but which is seldom accessible to the public. Moreover, this approach requires major involvement of domain experts, who would need to allow the system to track their research activity.

This is the approach we have chosen for IPSA, which was created as a research tool and provides methods for keeping track of research results. The further steps will be to make this valuable information available to all users in order to improve their involvement with cultural heritage content.

Funding

This work was supported by the CULTURA project (reference: 269973), and by the PROMISE network of excellence project (reference: 258191), within the Seventh Framework Programme of the European Commission.

References

Agosti, M., L. Benfante, and N. Orio. (2003). I*PSA: A Digital Archive of Herbals to Support Scientific Research. Proceedings of the Asian Conference on Digital Libraries*. 253-264.

Sweetnam, M. S. et al. (2012). *User Needs for Enhanced Engagement with Cultural Heritage Collections. Proceedings of Theory and Practice of Digital Libraries*. 64-75.

Fuhr, N. et al. (2007). Evaluation of Digital Libraries. *International Journal on Digital Libraries*, 8(1): 21-38.

Koller, D., B. Frischer, and G. Humphreys. (2010). Research challenges for digital archives of 3D cultural heritage models. *Journal of Computing and Cultural Heritage*. 2(3). 7.1-7.17.

Falk Anderson, E. et al. (2010). Developing serious games for cultural heritage: a state-of-the-art review. *Virtual Reality*. 14(4): 255-275.

A Comparative Kalendar: Building a Research Tool for Medieval Books of Hours from Distributed Resources

Albritton, Benjamin
blalbritton@gmail.com
Stanford University Library, United States of America

Sanderson, Robert
azaroth42@gmail.com
Los Alamos National Laboratory, United States of America

Ginther, James
ginthej@slu.edu
Saint Louis University, United States of America

Bradshaw, Shannon
sbradsha@drew.edu

Drew University, United States of America

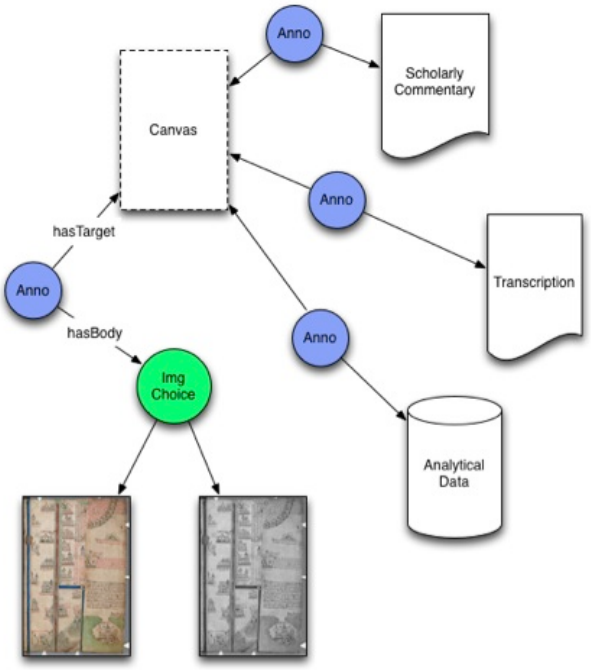
Foys, Martin

mfoys@drew.edu
Drew University, United States of America

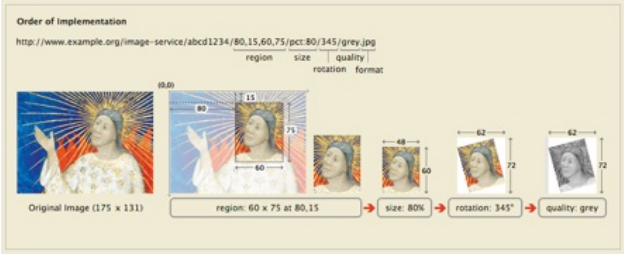
The Book of Hours is perhaps the most ubiquitous type of surviving medieval manuscripts. Each unique book contains a wealth of information about liturgical practices, social and private use of books, manuscript production and decoration, and personal and local choices for the creation of private devotional collections. Each Book of Hours contains a Kalendar that lists feast days for the entire year. These document the standard days for liturgical celebration, but are filled with local and temporal variants that paint a picture of the ways in which practices change according to regional interests over the centuries. Any attempt to curate and control a project devoted to cataloging Kalendar variants in a single team or location suffers from the sheer volume of potential inputs to the project. Further, the standard form of the Kalendar invites widespread contributions from interested scholars and the broader public. This paper describes a comparative Kalendar built from distributed resources served from multiple participating tools and repositories.

This distributed environment relies on repository/tool interactions to share image data in multiple tool environments, and harvest the tool output to build a discovery and viewing environment that can grow as more data is added from any participating tool or repository. These interactions rely on two principle interchange protocols: the SharedCanvas data model, used to aggregate distributed resources; and the International Image Interoperability Framework image API for standardized image resource access.

The SharedCanvas data model (http://www.shared-canvas.org) specifies a mechanism for representing a real-world manuscript object as a collection of two or more blank “canvases” each with a URI and an aspect ratio. Information, including digital image surrogates, is then associated with this canvas using the OpenAnnotation specification. This approach decouples information “about” the real-world object from the digital surrogate that could be used to illustrate that object on the web. Because it is quite possible that a single real-world resource will have multiple digital surrogates all purporting to represent the same thing, this approach makes it possible to build a digital facsimile utilizing a choice of digital surrogates without affecting the other information associated with the object (like transcriptions, scholarly commentary, or other metadata).



The International Image Interoperability Framework’s image API (http://lib.stanford.edu/iiif) specifies a RESTful web service to deliver a requested image with a number of specified parameters. These include: image format, size, rotation, cropping, and color. The API allows standardized image retrieval from disparate repositories, leading to the ability of tool developers to pull content in an efficient way from any participating host repository. More importantly, this API allows contributors to the Kalendar project to work with from a number of different interfaces and use image data served from a number of different repositories, allowing scholars with a regional or temporal focus to compare books held in different modern repositories.



The participating repositories in this effort are the Stanford University Libraries (serving content from the Walters Art Museum manuscript collection, the Parker Library, Corpus Christi College, Cambridge manuscript collection, and Stanford’s own collection), along with other participants of the Digital Manuscript Technical Council

(http://dms.stanford.edu/). This selection of content offers a variety of geographically and chronologically diverse Books of Hours, leading to an initial dataset that highlights both the consistency of some Kalendar elements over three centuries (13th, 14th, and 15th centuries), and the variants that provide insight into localized practices.

The tools used for this project are T-PEN (http://t-pen.org/TPEN) and DM (http://ada.drew.edu/dmproject/), specialized tools for transcription and annotation of medieval manuscript content. A user of either tool has full access to any of the manuscripts served from the participating repositories, and can add transcriptions of Kalendars while also adding structured data to help organize the material for further analysis and display. Both tools produce outputs that link user-generated information about the Kalendars to a region-constraint on the canvas that represents the page of the real-world manuscript. The tools provide a web service for exposing user-generated information that is then aggregated into a discovery and display interface.

This approach reduces the need for any one institution or project to host and serve every image that might be used in the Comparative Kalendar, to provide the user interface tools for transcription, commentary, and analysis, or to control and curate the input and output. The mechanism for user-interaction allows a user in any of the participating tools to choose content from any of the participating repositories. This allows a user to choose the working environment that most suits the tasks they wish to pursue, and do their analytical work within that tool. Since the tools allow standardized extraction of data for additional display, data management, navigation, or additional analysis and commentary, a user can move from tool environment to tool environment to achieve separate tasks (transcription, addition of structured metadata, commentary and notes, etc.).

The end result of this project is a user interface that draws image resources from the participating repositories and user-generated data produced by the participating tools, hosted by each respective participant. New data from any participating node will be automatically added to the interface without the need for human-mediated interaction, leading to a dynamically growing resource. This resource will allow a user to browse a Kalendar and see image, transcription, and commentary data about each entry, or compare across Kalendars to observe clusterings of continuity and variance over the liturgical year depending on location of book use and production, or chronological period. The distributed approach to a standardized text, where variants are of great interest across a large number of disciplines, and which benefits from a broad array of participants providing small bits of detailed data to build up a useful dataset, could serve as a model for further work on other parts of the Book of Hours, or other medieval texts that were copied frequently an reflect changing social, political, or liturgical practices across Western Europe over a period of several hundred years.

Tropes, Context and Computation: An approach to digital poetics

Algee-Hewitt, Mark Andrew
mark.algee-hewitt@stanford.edu
Stanford University, United States of America

Hauser, Ryan
heuser@stanford.edu
Stanford University, United States of America

While text mining, and similar quantitative analyses of lexical and semantic frequencies, have proven to be highly informative about aspects of literary history, the origin of these methods within the “hard" sciences creates hurdles for their application within the humanities (Pasanek and Sculley 2008, Bei 2008). Fundamental to these statistical strategies is the assumption of communicative equivalence among words as types: to most kinds of frequency analysis, each instance of a particular word retains the same meaning, force and valence of all other instances of that word. A chapter of a book in which the word “money” is relatively frequent suggests, to these methods, that the author is concerned with economics and it can therefore be related to other texts densely populated by economic terms. A critical problem in literary analysis, however, occurs within the field of tropes: in a seventeenth-century poem, a flea may be a small insect, or it may be a complex metaphor for a particular kind of romantic relationship. While newer, mixture-based, text mining methods, such as topic modeling, can use semantic clusters to reveal the difference between homonyms (the associated “topic” words of “bow”— applause, cheer, crowd — would be different than those of “bow” - arrow, fletch, string) they remain unable to detect the nuance of metaphors that are critical to our understanding of literary effect (Ross 2003, Blei 2012). Indeed, if most of the work of a poem comes not through the communicative value of the literal meaning of the words themselves, but instead, through the complex tropology of the formal effects, then we, as digital scholars of literary texts, require a new way to incorporate an understanding of the work of tropes within the semantic field. In particular,

at the level of genre identification, where much text mining or frequency analysis finds its home, the genres of allegory and satire, as extended tropes, remain unavailable to these methods. This problem is particularly acute when bringing to bear new quantitative methods on studies of poetry: with its highly figurative language and complex communicative intent, poetry remains on the fringe of quantitative analyses of literature. Most such studies focus exclusively on prose writing, and, in particular, the novel (Stanford Literary Lab 2011, Clement 2008). Given the importance of poetry to literary history, we believe that this lacuna represents a critical problem for the use of digital humanities in the study of literature.

Our project is an attempt to address this challenge through a new approach that combines a digital analysis of both the formal and lexical features of a constrained sample of poetry to test if it is possible to detect and identify the presence of tropological structures within poetic writing. In particular, this paper explores whether trope-based genres, specifically satire and allegory, are identifiable from linguistic and formal artifacts that can be recovered through digital analysis. In literary criticism, these genres are highly dependent upon context: a traditional analytic approach insists that a detailed knowledge of the poem's intended reception is necessary to recognize the satiric intent of a particular poem, or the allegorical framework that informs it (Fletcher and Bloom 2012). We contend, however, that there are generic markers within the formal and semantic fields of the poem that can be used to identify the presence of these tropes. To perform this analysis, we will draw upon a corpus of eighteenth-century poems selected for their comprehensive use of either satire or allegory, as well as a corpus with the same distribution across the period composed of poems that do not participate in either genre. By carefully choosing period and genre specific texts, our project, the first effort in this direction, seeks to limit the scope of its inquiry. As a way of approaching the larger question of poetic tropes, we ask if within the specific genres of eighteenth-century allegoric and satiric poems, we can train an algorithm to recognize the presence of either genre. A key to this project is our belief that the greatest challenge for the analysis of poetry using quantitative methods, that is, its highly figurative and stylized language, can be turned into an advantage in our approach. Both genres, we argue, operate on a semantic, as well as a formal level, and, therefore, the regularities of these formal structures can aid us in revealing key generic identifiers. Our method operates by identifying both the formal features and lexical regularities within a poem and comparing their relationship: we argue that the match, or more importantly, the *mismatch* between the formal structure and the semantic fields within poems reveals the presence of these tropological genres. In our project, this relationship becomes one of mediation: what patterns emerge when a topic is mediated through a form for which it is traditionally unsuited?

This is a particularly relevant question for the poetry of the eighteenth century, which formalized many of the generic tropes within specific combinations of style and meaning. Satire, for example, borrowed the rhetorical tropes of high-minded discourse (heroic couplets, iambic meter and poetic epithets) which are undercut by the inappropriateness of the poetic object (a lady's dressing room, the theft of a lock of hair) to the poetic medium. Similarly, allegory uses specific formal constraints to layer meaning within its lexical patterns. It is this formal shift towards different models of mediation that our project measures by combining text mining strategies, such as frequency analysis and topic modeling, with an automated recognition of poetic structure (like meter or rhyme scheme). While the results that we will present are specific to the genres of eighteenth-century poetic allegory and satire, we nevertheless believe that this represents a crucial first step towards automating the ways in which we recognize figurative language and developing flexible methods that can be extended to other contexts of use.

This project builds upon ongoing work within the Stanford Literary Lab on the subject of poetic discourse. Our project depends upon a corpus of poetry tagged as either satire or allegory (or neither, as a control), which can be used to both train our algorithm and test its efficacy. For this project, we have assembled a corpus of over 1500 poems, written between 1700 and 1800 that have been identified as belonging to one of our two particular genres. As this project seeks to test the ability of the methodology we have devised to identify the presence of satire or allegory within an entire poem, we have limited our corpus to poems that participate fully within the genres of satire or allegory, rather than those that have satiric or allegoric passages only. As the identifying markers of these genres are already highly nuanced, we feel that restricting our study to these two examples better enables us to identify a working model of poetic tropes that can later be improved for finer-grained and more detailed applications. Similarly, this will be the first practical application of an automated method for identifying and cataloguing meter and rhyme scheme within poetic writing that we have developed for the lab. As we argue that the artifacts of poetic tropes exist within the interaction between semantic fields and their formal mediation, we are able to bring our new ability to identify formal aspects of the poem to bear on the analysis of the patterns that these interactions take.

The stakes of this project involve both a push to incorporate poetic writing within our methodology of quantitative analysis, and a challenge to our understanding of the ways that figurative language operates within eighteenth-century poetry. Scholars, such as Fletcher and

Bloom on allegory (2012) and Connery and Combe on satire (1995), argue that we are only able to recognize these genres through a detailed understanding of the historical and cultural contexts in which the poem was written. Our methodology, however, argues for an expanded understanding of context: instead of a detailed knowledge of eighteenth-century political or economic history, context, in our study, is a function of the relationship between semantic fields and their formal mediation within a poem. This, we argue, is equally as important to the context of the poem as the geo-political and historical world from which it emerged. While the necessary specificity of our study limits our project's conclusions to the genres of eighteenth-century allegory and satire, we believe that our results have implications for a wider understanding of how figurative language can be incorporated within a quantitative analysis of literature. By carefully limiting our genres and period, we are offering this study as a proof-of-concept test whose results will aid in future quantitative and digital work on poetry. We anticipate that this pilot study can be extrapolated into a wider understanding of figurative language and the tropes in which it operates and that our results from this project can be incorporated into future work that seeks to undertake a wider and more comprehensive analysis of poetic language across multiple centuries.

Identifying the Real-time impact of the Digital Humanities using Social Media Measures

Alhoori, Hamed M
alhoori@gmail.com
Texas A&M University, United States of America

Furuta, Richard
furuta@cse.tamu.edu
Texas A&M University, United States of America

Introduction

Rankings of academic articles and journals have been used in most disciplines, although concerns and objections about their use have been raised, particularly when they affect appointments, promotions and research grants. In addition, journal rankings may not represent real research outcomes, since low-ranking journals can still contain good work. Arts and humanities scholars have raised additional concerns about whether the various rankings accommodate differences in cultures, regions and languages. Di Leo (2010) wrote that "journal ranking is not very useful in academic philosophy and in the humanities in general" and one reason is the "high level of sub-disciplinary specialization". Additionally, Di Leo notes there is "little accreditation and even less funding" in the humanities when compared with business and sciences.

In a *Nature* article entitled, *"Rank injustice",* Lawrence (2002) notes that the "Impact factor causes damaging competition between journals since some of the accepted papers are chosen for their beneficial effects on the impact factor, rather than for their scientific quality". Another concern is the effect on new fields of research. McMahon told *The Chronicle of Higher Education,* "Film studies and media studies — they were decimated in the metric because their journals are not as old as the literary journals. None of the film journals received a high rating, which is extraordinary" (quoted by Howard 2008).

Although the Australian government dropped rankings after complaints that they were being used "inappropriately", it will still offer a profile of journal publications that provides an "indication of how often a journal was chosen as the forum of publication by academics in a given field" (Rowbotham 2011). Despite concerns over rankings, educators and researchers agree there should be a quality management system. By publishing their results, researchers are not just talking to themselves. Research outcomes are for public use, and others should be able to study and measure them. However, the questions are how can we measure the research efforts and their impact, and can we get an early indication of research work that is capturing the research community's attention. A second question is whether measures appropriate for one research area also can be applied to publications in a different area. In this study, we seek initial insights on these questions by using data from a social media site to measure a real-time impact of articles in the digital humanities.

Research Community Article Rating (RCAR)

Citation analysis is a well-known metric to measure scientific impact and has helped in highlighting significant work. However, citations suffer from delays that could span months or even years. Bollen, *et al.,* (2009) concluded

that "the notion of scientific impact is a multi-dimensional construct that cannot be adequately measured by any single indicator". Terras (2012) found that digital presence in social media helped to disseminate research articles, and "open access makes an article even more accessed".

An alternative approach to citation analysis is to use data from online scholarly social networks (Priem and Hemminger 2010). Scholarly communities have used social reference management (SRM) systems to store, share and discover scholarly references (Farooq *et al* 2007). Some well-known examples are Zotero[1] , Mendeley (Henning and Reichelt, 2008) and CiteULike [2] . These SRM systems have the potential to influence and measure scientific impact (Priem *et al.,* 2012). Alhoori and Furuta (2011) found that SRM is having a significant effect on the current activities of researchers and digital libraries. Accordingly, researchers are currently studying metrics that are based on SRM data and other social tools. For example, *Altmetrics*[3] was defined as "the creation and study of new metrics based on the social web for analyzing and informing scholarship". PLOS proposed article-level metrics (ALM) [4] that are a comprehensive set of research impact indicators that include usage, citations, social bookmarking, dissemination activity, media, blog coverage, discussion activity and ratings.

Tenopir and King (2000) estimated that scientific articles published in the United States are read about 900 times each. Who are the researchers reading an article? Does knowing who these researchers are influence the article's impact? Rudner, *et al.,* (2002) used a readership survey to determine the researchers' needs and interests. Eason, *et al.* (2000) analyzed the behavior of journal readers using logs.

There is a difference between how many times an article has been cited and how many times it has been viewed or downloaded. A citation means that an author has probably read the article, although this is not guaranteed. With respect to article views, there are several viewing scenarios such as intended clicks, unintended clicks or even a web crawler. Therefore, the number of views has hidden influential factors. To eliminate the hidden-factors effect, we selected the articles that researchers had added to an academic social media site. In this study, we ranked readers based on their education level. For example, a professor had a higher rank than a PhD student, who in turn had a higher rank than an undergraduate student.

Zotero's readership statistics were not available to the public, and in CiteULike, the most cited articles in *Literary and Linguistic Computing (LLC)* were shared by few users. Therefore, we were unable to use either system's data. Instead, we obtained our data from Mendeley, using its API[5] . We measured the research community article rating (RCAR)using the following equation:

$$RCAR = \frac{\sum R + \sum (P * K) + \sum D + \sum C + \sum A + \sum G}{\log (y_c - y + 2)} \quad (1)$$

RCAR uses the following quantities:

- R = researchers that had added an article to their digital library.
- $\sum (P * K)$ = percentage (P) of researchers who had an article, multiplied by their rankings (K).
- $\sum D$= number of different academic disciplines of the R.
- $\sum C$= number of different countries of the R.
- $\sum A$= number of authors to an article.
- $\sum G$= number of online groups that shared an article.
- y_c = current year.
- y = year the article was published.

Citations, Readership and RCAR

We looked at seven digital humanities journals that were added to Mendeley and also mentioned in Wikipedia[6] . Of the seven journals, Google Scholar had an h5-index for only two journals: *Digital Creativity* (h5-index = 7) and *LLC* (h5-index = 13). We calculated the RCAR and compared the top-cited *LLC* articles with Mendeley readerships, as shown in Table 1. The number of citations were significantly higher than the number of Mendeley readerships for *LLC* (p-value>0.05).

Article title	Citations	Readerships	RCAR	Year
Quantitative Authorship Attribution: An Evaluation of Techniques	73	42	84.85	2007
If You Build It Will They Come? The LAIRAH Study: Quantifying the Use of (	37	16	46.03	2008
An evaluation of text classification methods for literary study	32	16	40.97	2008
Bigrams of Syntactic Labels for Authorship Discrimination of Short Texts	35	17	41.11	2007
All the Way Through: Testing for Authorship in Different Frequency Strata	25	11	29.24	2007
Function Words in Authorship Attribution Studies	28	16	39.37	2007
Use of the Chi-Squared Test to Examine Vocabulary Differences in English	24	9	27.05	2007
Interpreting Burrows's Delta: Geometric and Probabilistic Foundations	23	13	38.10	2008
Supporting Annotation as a Scholarly Tool—Experiences From the Online (	19	20	48.10	2007
Modelling Space and Time in Narratives about Restaurants	20	8	22.35	2007
Reassessing authorship of the Book of Mormon using delta and nearest sh	21	17	44.91	2008
The Identification of Spelling Variants in English and German Historical Te	16	9	31.52	2008
The effect of author set size and data size in authorship attribution	20	18	81.24	2011

Table 1:
Google citations, Mendeley readerships and RCAR for LLC

We investigated how the digital humanities discipline is different from other disciplines. We compared *LLC* with a journal from a different area of research, *Library Trends*, which had a similar h5-index. *Library Trends* received more citations and readerships than *LLC*. Three of its top articles also had more Mendeley readerships than citations, whereas *LLC* only had one such case. However, there was no significant difference between *Library Trends* citations and readerships. Next, we tested the *Journal of the American Society for Information Science and Technology* (JASIST) and the *Journal of Librarianship and Information Science* (JOLIS). We found that JASIST and JOLIS readerships of articles published in 2012 were higher than the citations with significance difference. This indicates that computer, information, and library scientists are more

active in academic social media site than digital humanities researchers. By active we mean that they share and add newly published articles to their digital library.

Citations and altmetrics

In order to better understand different socially-based measures, we compared *LLC* articles using altmetrics and citations. We used an implementation of *altmetrics* called *Altmetric* that gave "each article a score that measures the quantity and quality of attention it has received from Twitter, Facebook, science blogs, mainstream news outlets and more sources". We found that most of the articles that received social media attention were published during the last two years. A number of articles that were published four or more years ago were exceptions to this finding. These older articles had received at least four citations, as shown in Table 2. We also found similar correlations with articles in *Digital Creativity*.

Finally, we compared readerships and Altmetric. We found no significant difference between LLC citations of articles published in 2012 and readerships. However, we found a significant difference between Altmetric and citations ($p<0.05$) for articles that were published in 2012. This shows that the researchers who are interested in digital humanities are more active in general social media sites (e.g. Twitter, Facebook) than academic social media sites (e.g. Mendeley).

Article title	Altmetric	Citations	Year
Transcription maximized; expense minimized? crowdsourcing and editing T	17.55	2	2012
Longitudinal detection of dementia through lexical and syntactic changes i	12.45	4	2011
A rationale of digital documentary editions	6.45	4	2011
Computational analysis of the body in European fairy tales	6.3	1	2012
Reassessing authorship of the Book of Mormon using delta and nearest shi	5.35	22	2008
Experiments in 17th century English: manual versus automatic conceptual l	4.35	0	2012
Improving record matching in imprecise and uncertain datasets	3.75	0	2012
Managing and Growing a Cultural Heritage Web Presence. A strategic guid	3.25	0	2012
Natural language processing and early-modern dirty data: applying IBM Lar	2.75	0	2012
Scalability Issues in Authorship Attribution.Kim Luyckx.	2.75	4	2011
Detecting authorship deception: a supervised machine learning approach u	2.25	1	2012
Co-occurrence-based indicators for authorship analysis	2	0	2012
A thing not beginning and not ending': using digital tools to distant-read Ge	2	13	2008
It's a team if you use "reply all" ': An exploration of research teams in digit	2	15	2009
Who wrote Shamela? Verifying the Authorship of a Parodic Text	2	4	2005
The Density of Latinate Words in the Speeches of Jane Austen's Characters	1.85	9	2001
The inadequacy of embedded markup for cultural heritage texts	1.85	9	2010
Visual Interface Design for Digital Cultural Heritage. A Guide to Rich-Prosp	1.75	0	2012
The Tesserae Project: intertextual analysis of Latin poetry	1.75	0	2012
Ce qui compte. Méthodes statistiques. Ecrits choisis, tome II. Etienne Bru	1.75	0	2012
The Potosi principle: Religious prosociality fosters self-organization of large	1.75	0	2012
Text and Genre in Reconstruction. Effects of Digitalization on Ideas, Behavi	1.6	0	2012
How To Do Things With Videogames. Ian Bogost.	1.6	0	2012
Digital Research in the Study of Classical Antiquity.Gabriel Bodard and Sim	1.6	0	2012
In Memoriam Charles Douglas Bush (1948-2011)	1.6	0	2011
A trip around the world: Accommodating geographical, linguistic and cultur	1.5	0	2012
Poetics of crisis or crisis of poetics in digital reading/writing? The case of S	1.25	0	2012
Expressing complex associations in medieval historical documents: the Her	1	10	2008
Narrative rules? Story logic and the structures of games	1	0	2012
Social network visualization from TEI data	1	0	2011

Table 2:
Altmetric score and citations to LLC articles

In this paper we describe a new multi-dimensional approach that can measure in real-time the impact of digital humanities research using academic social media site. We found that RCAR and altmetrics can quantify an early impact of articles gaining scholarly attention. In the future, we plan to conduct interviews with humanities scholars, to better understand how these observations reflect their needs and the standards in their fields.

References

Alhoori, H., and R. Furuta (2011). Understanding the Dynamic Scholarly Research Needs and Behavior as Applied to Social Reference Management. *TPDL'11 Proceedings of the 15th international conference on Theory and practice of digital libraries.* Berlin Heidelberg: Springer. 169–178.

Bollen, J., H. Van de Sompel, A. Hagberg, and R. Chute (2009). A principal component analysis of 39 scientific impact measures. *PloS one.* 4(6). e6022.

Eason, K., S. Richardson, and L. Yu (2000). Patterns of use of electronic journals. *Journal of Documentation.* 56(5): 477–504.

Farooq, U., Y. Song, J. M. Carroll, and C. L. Giles (2007). Social Bookmarking for Scholarly Digital Libraries. *IEEE Internet Computing.* 11(6): 29–35.

Henning, V., and J. Reichelt (2008). Mendeley — A Last.fm For Research? In *2008 IEEE Fourth International Conference on eScience.* IEEE. 327–328.

Howard, J. (2008). New Ratings of Humanities Journals Do More Than Rank They Rankle. *The Chronicle of Higher Education.* http://chronicle.com/article/New-Ratings-of-Humanities/29072 (Accessed 5 March 2013).

Lawrence, P. A. (2002). Rank injustice. *Nature.* 415(6874): 835–6.

Di Leo, J. (2010) Against Rank. *Inside Higher Ed.* http://www.insidehighered.com/views/2010/06/21/dileo (Accessed 5 March 2012).

Priem, J., and B. M. Hemminger (2010). Scientometrics 2.0: Toward new metrics of scholarly impact on the social Web. *First Monday.* 15(7).

Priem, J., C. Parra, H. Piwowar, P. Groth, and A. Waagmeester (2012). Uncovering impacts: a case study in using altmetrics tools. *Workshop on the Semantic Publishing (SePublica 2012) at the 9th Extended Semantic Web Conference.*

Rowbotham, J. (2011). End of an ERA: journal rankings dropped. *The Australian*. http://www.theaustralian.com.au/higher-education/end-of-an-era-journal-rankings-dropped/story-e6frgcjx-1226065864847 (Accessed 5 March 2013).

Rudner, L.M., J. S. Gellmann, and M. Miller-Whitehead (2002). Who Is Reading On-line Education Journals? Why? And What Are They Reading? *D-Lib Magazine.* 8(12).

Tenopir, C., and D. W. King (2000). *Towards Electronic Journals: Realities for Scientists, Librarians, and Publishers.* Washington, D.C.: Special Libraries Association.

Terras, M. (2012). The impact of social media on the dissemination of research: results of an experiment. *Journal of Digital Humanities.* 1(3): 30–38.

Opening Aladdin's cave or Pandora's box? The challenges of crowdsourcing the Medici Archives

Allori, Lorenzo
lorenzo.allori@gmail.com
The Medici Archive Project, Italy

Kaborycha, Lisa
lkaborycha@medici.org
The Medici Archive Project, Italy

Introduction

This year the Medici Archive Project (MAP), based in the State Archive in Florence, launches BIA, an interactive on-line digital platform, thanks to the generous funding from the Andrew W. Mellon Foundation. This ambitious project is in the process of digitizing around four million handwritten letters dating from the sixteenth to eighteenth centuries which are part of the Medici Granducal Archive, 1537-1743. Eventually MAP will publish digital images of these historical documents on BIA, which were formerly accessible only onsite in Florence. Along with each image, BIA will also provide a transcription of each document in its original language, accompanied by an English-language synopsis, which places the document within a historical context. BIA has been developed using Java following the J2EE standards, utilizing the Spring Framework. It is based on a relational database and uses Apache Lucene for indexing the data for full-text searching.

Existing models

The completion of this vast undertaking has become a real possibility by using crowdsourcing and digital editing techniques. Projects such as the University College London's award-winning *Transcribe Bentham* [1] and the *The Civil War Diaries & Letters Transcription Project* at the University of Iowa[2] have demonstrated how successful crowdsourcing can be, having accomplished a large proportion of their transcriptions in a relatively small amount of time. However, these projects encompass relatively small collections: *Transcribe Bentham* is part of a collection that contains around 60,000 folios, arranged into 174 boxes; the *Civil War Diaries* project has 172 documents, with a total of 18,270 scanned pages. Like these, the Center for History & New Media at George Mason University's The Papers of the War Department Project 1784-1800[3] , the National Archives' *You Can Transcribe It!* [4] , and the New York Public Library's *What's On The Menu?* [5] , among others, also invite members of the general public to volunteer to transcribe documents.

Challenges

Unlike the above projects, however, the Medici Archive Project faces specific challenges based on the following:

- Size of collection: ca. four million handwritten letters
- High level of technical expertise: paleography, language, historical training
- Varying languages, nationalities, and cultural backgrounds of community

Comparably large digitized projects, which make use of crowdsourcing, such as the Australian Newspapers Digitisation Program[6] , can take advantage of OCR technologies to expedite the data entry phase. However, because the documents in the Medici Granducal Archive are handwritten, each document needs to be manually transcribed. Moreover, the technical expertise in paleography required to read these documents excludes the feasibility of broad-based crowdsourcing of data entry. Among the challenges we face, is that of finding the broadest possible group of community contributors with sufficient expertise to transcribe and provide historical context. Thus, MAP's user base tends to be high-level academic researchers, professors and graduate students, who are specialists in early modern history and art history, who are at least bilingual in English and Italian. There are other languages used in the granducal documents as well—

Spanish, German, Latin, French, etc.—because the Medici court had connections throughout all of Europe. Thus the BIA community will involve scholars from around the globe; each scholar bringing not only varying levels of linguistic ability, but also different cultural approaches to collaborative work.

There were two main cruxes that our IT Team had to resolve, while building BIA: firstly create a stable platform suitable for proper data-entry and transcription of digitized manuscripts; and secondly coming up with a forum-like tool which could be employed by the scholars to communicate while working on BIA.

The community-sourcing model

Thus, rather than crowdsourcing, MAP's approach is one of community-sourcing, creating a hierarchy of levels of contributors to the BIA system, where issues of gatekeeping are foremost. Furthermore, unlike many opensourced projects, where a user's anonymity or pseudonymity is the standard, the academics who make use of BIA require the assurance of the authority behind transcriptions and contextualizations. They need to be assured that fact-checking procedures have been rigorously followed, that disambiguation with regard to people and places have been correctly ascertained, and these can only be guaranteed by providing the name of the scholar beside his or her work. In addition to these accountability mechanisms, the scholarly community presents further challenges to a project of this kind. There is often resistance to sharing documents, especially if they have not yet been published. In order to encourage trust in sharing raw, early-stage data before publication, MAP must create a spirit of open collaboration between peers, with mutual accountability, as well as enforce scholarly citation norms and occasionally issuing warnings when abuses are reported. The system was first tested by a restricted group of off-site scholars (30) a few months before the official launch. The results of this testing phase have been very successful. BIA has now been made available for the entire scholarly community: by July 2013, we shall have a clearer picture of the evolution of this DH project.

Conclusion

This paper describes the challenges faced within the first nine months of this innovative project, examining the successful strategies implemented, as well as improvements that have been introduced. There will be discussion of the future not only of the Medici Archive Project BIA system, but of the long-term prognosis of community-sourced digital humanities projects as a whole.

References

Causer, Tim, and V. Wallace (2012). Building A Volunteer Community: Results and Findings from Transcribe Bentham DHQ, 6(2). http://www.digitalhumanities.org/dhq/vol/6/2/000125/000125.html

Deegan, M., and W. McCarty (2012). *Collaborative Research in the Digital Humanities*. London: Ashgate.

Dormans, S., and J. Kok (2010). An Alternative Approach to Large Historical Databases: Exploring Best Practices with Collaboratories. *Historical Methods*, 43:3. 97-107.

Organisciak, P. (2010). *Why Bother? Examining the Motivations of Users in Large-Scale Crowd-Powered Online Initiatives*. MA Thesis, University of Alberta.

Parry, M. (2012). Historians Ask the Public to Help Organize the Past. But is the crowd up to it? *The Chronicle of Higher Education*, 3 September 2012.

Ridge, M. (2012). On the Internet, nobody knows you're a historian: exploring resistance to crowdsourced resources among historians *Digital Humanities Conference*, Hamburg, 20 July 2012. http://lecture2go.uni-hamburg.de/konferenzen/-/k/14007

Sample Ward, A. (2011). Crowdsourcing vs Community-sourcing: What's the difference and the opportunity? *Amy Sample Ward's Blog*, 18 May 2011. http://amysampleward.org/2011/05/18/crowdsourcing-vs-community-sourcing-whats-the-difference-and-the-opportunity/

Siemens, L., R. Cunningham, W. Duff, and C. Warwick (2011). More minds are brought to bear on a problem: Methods of Interaction and Collaboration within Digital Humanities Research Teams *DS/CN*, 2(2).

Terras, M. (2010). DH2010 Plenary: Present, Not Voting: Digital Humanities in the Panopticon. Digital Humanities 2010 plenary talk manuscript. *Melissa Terras' Blog*, 10 July, 2010. http://melissaterras.blogspot.com/2010/07/dh2010-plenary-present-not-voting.html .

Winters, J. (2011). Digital editions and crowdsourcing *Blog ReScript, the Institute of Historical Research*, 9 June 2011. http://rescriptihr.blogspot.it/2011/06/digital-editions-and-crowdsourcing.html

Notes

1. http://www.ucl.ac.uk/Bentham-Project/transcribe_bentham
2. http://diyhistory.lib.uiowa.edu/transcribe/
3. http://wardepartmentpapers.org/index.php
4. http://www.archives.gov/citizen-archivist/

5. http://menus.nypl.org/
6. http://www.nla.gov.au/content/newspaper-digitisation-program

Representing Texts Electronically in Lesser-used Languages: Current Issues and Challenges in Character Encoding

Anderson, Deborah
dwanders@sonic.net
UC Berkeley, United States of America

A premise of the Digital Humanities 2013 conference theme, "Freedom to Explore," is that users of the languages of the world should be able to express themselves and search for documents in their own language (or one of their heritage), using the script employed to write that language. For many widely used languages, this is generally possible today, because its script (and characters) can be represented by the international character encoding standard Unicode (Unicode Consortium, 2012) and its ISO mate, ISO/IEC 10646 (ISO/IEC, 2012), and because these scripts are supported on today's computers and electronic devices.

But those scripts used for lesser-used languages — both modern and historic — that are *not* in Unicode are subject to problems in representation and searching, since they are not part of the standard. One consequence is that this will leave gaps in the world's cultural, linguistic, and historical legacy. Even after the scripts are accepted and published in the Unicode Standard (and ISO standard), the languages and their scripts must then be supported in fonts and rendering engines, and locale data information is needed to implement a modern script on cell phones or computer devices. Because of these factors, the goal of making electronic text communication truly global and multilingual remains a high bar, but one that the digital humanities should continue to strive to achieve.

This short paper will report on recent developments in the effort to make the text of lesser-used languages accessible electronically, focusing primarily on issues involved in getting characters into Unicode but also touching on the social aspects of character encoding.

In the past few years, Unicode (and ISO/IEC 10646) has become more widely accepted and better understood, even by laypeople. Indeed, forty-two modern and historic scripts have been added in the various releases over the past six years[1] , most through the UC Berkeley Script Encoding Initiative (Script Encoding Initiative, 2012). In the past, it was often challenging to explain to language communities the importance of getting a script into Unicode. However, today user communities are themselves approaching the Unicode Consortium (or others involved in character encoding) in order to get their script included in the standard. The Script Encoding Initiative at UC Berkeley has, for example, been approached to help in encoding the ADLaM alphabet in Guinea and the Mandombe script in the Democratic Republic of Congo, two recently devised scripts intended for modern use. Historic script users — who may be heritage language users, "hobbyists" or come from academia or a liturgical setting — have also been active in getting scripts encoded. For example, experts of the Hatran script of the Middle East recently sought help in getting their script encoded.

While it is useful to have the user communities directly involved in the proposal process, proposals written by new authors often require extensive assistance. Typically, the approval process takes much longer for proposals written by new authors in order for all the required technical information to be included in the proposal. A more effective approach, adopted by the Script Encoding Initiative (SEI), has been to encourage user communities to work with a veteran Unicode proposal author, who is familiar with the standards committees' requirements. However, relying on an outsider to help on a proposal may be looked on with suspicion. One way to allay this fear, an approach also advocated by the SEI project, is to try to encourage collaborative work between an expert in the encoding process and the users, making it clear the ownership of the script clearly belongs to the user community, but the encoding expert (and the standards committees) will help make sure the script will be implementable on today's computers and fonts. Currently, work on the Nepaalalipi/ Newar proposal is following this approach and a report on its success (or failure) will be relayed.

One issue in the encoding process that has repeatedly been an issue for user communities of modern scripts is what name to assign the script. A script name in Unicode is meant to be as an identifier (such as in programming), and typically one that is commonly found in English.2 The English requirement has been problematic, since the English name is often not that of the user communities. Also, the name for a script can vary across languages (and countries), so there can be a tug-of-war between groups as to which name to use. One way to deal with the name has been to encourage user communities to translate character names

in the code charts into their own language. Likewise, once the script is approved, fonts can use their preferred name for the glyphs. In dealing with competing names in different languages, one solution has been to get the groups to try to agree on a name that is acceptable to all. This approach was successful for Tai Tham ("Lanna" in Thailand), but has not worked out well for Old Hungarian/Szekely-Hungarian Rovas. (Typically, names for historic scripts are not controversial or contentious, though disagreements have arisen *within* in the standards committees on a particular name or character.)

The current process of script encoding involves approval by the Unicode Technical Committee (UTC) and the subcommittee on coded character sets in the International Organization for Standardization (ISO). Because ISO represents character encoding in the governmental arena, the subcommittee chair on Coded Character Sets in the past three years has tried to encourage countries of lesser-used scripts to participate in ISO, particularly those in West Africa. However, many countries have not been able to participate due to lack of funding. As a way to involve the user communities, the SEI project has worked with user communities, without direct support or involvement of the government.

Once a script is approved and published in a version of the Unicode Standard (and ISO/IEC 10646), a text in that script is not necessarily ready for interchange: fonts with the correct codepoints must be created as well as an input method, the rendering engine must be updated to support the script (particularly if the script is complex and has special processing for display), and locale data is needed for implementations, including electronic devices. In these areas, one of the more promising areas of assistance for modern lesser-used languages and scripts is the ScriptSource project sponsored by SIL International (SIL International, n.d.), The talk will include a discussion of the Cherokee script as it progressed from initial Unicode proposal to iPad and iPhone implementations, as a model for other lesser-used scripts. The paper will close with thoughts from the speaker on how to improve the current process and possible next steps, such as:

- engage users and relevant national body standards bodies early in the process of encoding a script (so as to prevent a situation that arose with Khmer3)
- support funding for projects in which a dedicated person can work with the user communities and instruct them on the encoding process
- encourage users to vote (via their national standards body) on relevant ISO script encoding ballots.

References

Bauhahn, M., and M. Everson (2001). Response to Cambodian official objection to Khmer block (N2380). (=ISO/IEC JTC1/SC2/WG2 N2385). http://std.dkuug.dk/JTC1/SC2/WG2/docs/n2385.pdf (accessed February 2013).

Committee for Standardization of Khmer Characters in Computers (2001). Cambodian official objection to the existing Khmer block in UCS. (=ISO/IEC JTC1/SC2/WG2 N2380). (2001). Web. http://std.dkuug.dk/JTC1/SC2/WG2/docs/n2380.pdf (accessed 19 February 2013).

International Organization for Standardization/ International Electrotechnical Commission (2012). ISO/IEC 10646:2012(E). Information technology — Universal Coded Character Set (UCS). Third ed. http://standards.iso.org/ittf/PubliclyAvailableStandards/index.html (accessed: 12 February 2013).

Script Encoding Initiative. (2012). List of SEI-Supported Proposals Completed and Published in Unicode and ISO/IEC 10646. UC Berkeley. http://www.linguistics.berkeley.edu/sei/alpha-script-completed.html. (accessed 19 February 2013).

SIL International Writing systems, computers and people. Scriptsource. http://scriptsource.org. (accessed 19 February 2013).

Unicode Consortium (2012). Character Properties, Case Mappings and Names FAQ. http://www.unicode.org/versions/Unicode6.2.0/. (accessed 3 November 2012).

Unicode Consortium (2012). Unicode 6.2.0. http://www.unicode.org/versions/Unicode6.2.0/. (accessed 3 November 2012).

Notes

1. See the charts for the various editions of the Unicode Standard, such as: http://www.unicode.org/charts/PDF/Unicode-6.1/ , http://www.unicode.org/charts/PDF/Unicode-6.0/ , http://www.unicode.org/charts/PDF/Unicode-5.2/ , http://www.unicode.org/charts/PDF/Unicode-5.1/ , http://www.unicode.org/charts/PDF/Unicode-5.0/ .
2. See the Unicode Consortium's Frequently Asked Question on this topic, Unicode Consortium 2012b.
3. The encoding of the Khmer script in 1999-2000 resulted in the Khmer community feeling acutely left out of the process, (cf. their objections in Committee for Standardization of Khmer Characters in Computers, 2001), although the situation was probably somewhat more complex (cf. Bauhahn and Everson, 2001)

Optimized platform for capturing metadata of historical correspondences

Andert, Martin
martin.andert@informatik.uni-halle.de
Computer Science Department, Martin-Luther-Universität Halle-Wittenberg, Germany

Ritter, Joerg
joerg.ritter@informatik.uni-halle.de
Computer Science Department, Martin-Luther-Universität Halle-Wittenberg, Germany

Molitor, Paul
paul.molitor@informatik.uni-halle.de
Computer Science Department, Martin-Luther-Universität Halle-Wittenberg, Germany

Various projects around the world focus on analyzing and visualizing the correspondences and social networks of scholars(academics, writers, etc), e.g., the 'Mapping the Republic of Letters' project centered at Stanford University (Edelstein et al., 2008) and the project 'Vernetzte Korrespondenzen'(engl.: Networked Correspondences) (Burch et al., 2012) run by Trier Center for Digital Humanities, German Literary Archive Marbach, and Martin-Luther-University Halle-Wittenberg. In addition to the design and implementation of adequate tools for analyzing and visualizing the big data, an indexing scheme for the content of the letters of the corpus under consideration has to be developed and the letters have to be captured, annotated, and indexed with respect to the indexing scheme.

The capturing, annotating, and indexing process should be supported by an interactive browser based platform which allows the professional philologists (as well as 'citizen scholars' from non-academic backgrounds) to quickly enter metadata (e.g., date, name/identification of senders and addressees) in an accurate und simple manner. To meet the simplicity requirement, the platform should accept unformatted entries as far as possible; to meet the accuracy requirement, the entries have to be clear without ambiguity. The two requirements appear to be competitive and conflicting. To sort out the problem, we propose to support unformatted data entry by auto-completion based on external catalogues, e.g., the 'Deutsche Nationalbibliografie'(DNB, 2013b) provided by the German National Library. The auto-completion tool should provide additional biographical information on the proposed scholars as soon as the philologist enters first data, to help the philologist uniquely identify scholars.

In preparation for the above mentioned 'Vernetzte Korrespondenzen' project(Burch et al., 2012) which investigates the correspondences and social networks of German scholars which went into exile during Nazi Germany, we have implemented such a browser based platform.

- The platform is generic in the sense that it can be trained on different centuries by using one or several external catalogues. In principle, each catalogue which provides a web service that offers infix search of names with the possibility of limiting the search to a specified time period can be used for training the platform. The illustrations presented in the next paragraph assume that the platform has been trained on scholars of the 18th and 19th Century.
- The platform only propounds scholars based on PND/GND-normed data together with additional accurate biographical and historical data which are extracted from existing audited catalogues. These supplementary data should allow the philologist/historian to uniquely identify the scholar. After identification the system files the scholars in a standardized manner, i.e., normed data are inserted in the metadata.
- The platform provides capturing of accurate dates by nearly unformatted entries.

As soon as the philologist enters a part of the name of the sender or the addressee, e.g., 'Joseph', the tool makes proposals of scholars of the 18th and 19th Century that match the requirement, e.g., 'Eichendorff, Joseph von', 'Hübner, Joseph Alexander von' and so on, together with their dates of birth and death, identification number of the name authority file ('Gemeinsame Normdatei') published by the German National Library(DNB, 2013a), and short vitae. Especially these secondary data are of benefit if several scholars with similar(or same) name exist. Once the philologist made her choice, the system pins the scholar's biographical data to the work space. If the person is unknown to the system which may be the case if the letter was sent, e.g., to the pharmacist next door (whose name is not specified in the letter), the philologist enters this new person into the system. The autocompletion mechanism of the platform uses these new data during future requests.

The entry of the letter's date poses a further challenge as the determination of the date is harder than you think. For instance, a letter can be written over several days, the date may be hardly readable, it can be given by a

holiday, or it can be specified imprecisely. That's why our platform allows the philologist input the date of the letter in various formats. For instance, the letter can be dated 'early January 1798', 'Eastern 1770', 'mid-April 1770', '19/04/1770', or '15/04/1770-20/04/1770'. The conversion of the entry to an exact date or a period is automatically done by the underlying system. 'Early January 1798', e.g., is converted to the period 01/01/1798-10/01/1798, 'Eastern 1770' to the period 12/04/1770-16/04/1770, 'mid April 1770' to 11/04/1770-20/04/1770. After the date had been entered, the platform provides additional information: It runs plausibility checks, e.g., whether the date of the letter is in life of the correspondents, and provides information on historical events that took place immediately before or immediately after that date. For instance, in the case of Eastern 1770, the system would provide the information that the German evangelical theologian August Wilhelm Reinhart passed away on 17 April 1770 and Marie Antoinette got married per procurationem Dauphin Louis-Auguste in the Augustinian Church in Vienna on 19 April 1770. Especially, later on, such data should help the philologist annotate the letters. These latter data can be extracted from digital chronicles, e.g., the 'Chronik deutscher Zeitgeschichte'(Overesch et al., 1982) for adapting the platform to the project 'Vernetzte Korrespondenzen'(Burch et al., 2012).

If possible, we will distribute the tool in different digital research infrastructures for the humanities, e.g., DARIAH(DARIAH-EU, 2012), CLARIN(CLARIN, 2011).

Funding

This work was supported by the German Federal Ministry of Education and Research (BMBF) [grant number 01UG1354C]

References

Edelstein, D., P. Findlen, and N. Coleman (2008). Mapping the Republic of Letters. Stanford University. https://republicofletters.stanford.edu/ (accessed 15 February 2013)

Burch, T., V Hildenbrandt., S. Kamzelak, P. Molitor, C. Moulin, and J. Ritter (2012). Vernetzte Korrespondenzen — . BMBF-Projekt 01UG1354. http://kompetenzzentrum.uni-trier.de/de/projekte/projekte/briefnetzwerk/ (accessed 15 February 2013)

Deutsche Nationalbibliothek (2013a). Homepage. http://www.dnb.de/DE/Home/home_node.html (accessed 15 February 2013)

Deutsche Nationalbibliothek(2013b). Deutsche Nationalbibliografie. http://www.dnb.de/EN/Service/DigitaleDienste/DNBBibliografie/dnbbibliografie_node.html (accessed 15 February 2013)

Overesch, M., and F. W. Saal (1982). Chronik deutscher Zeitgeschichte. Droste Geschichtskalendarium. Droste Verlag 1982-1986.

DARIAH-EU (2012). Introducing DARIAH-EU. Information Brochure. http://www.dariah.eu/(accessed 15 February 2013)

CLARIN(2011). CLARIN ERIC STATUTES. http://www.clarin.eu/external/index.php?page=about-clarin (accessed 15 February 2013)

An Interactive Interface for Text Variant Graph Models

Andrews, Tara Lee
tara.andrews@arts.kuleuven.be
KU Leuven, Belgium

Van Zundert, Joris Job
joris.van.zundert@huygens.knaw.nl
Huygens ING, Royal Netherlands Academy of Sciences (KNAW)

In the last three years there has been an increasing adoption of the variant graph as a suitable computer-internal data model for textual variation in tools and applications such as NMerge (Schmidt & Colomb 2009), CollateX (Dekker & Middell 2011), and Stemmaweb (Andrews & Macé 2013). The variant graph model can be visualized easily and succinctly; these visualizations become for the text researcher the first interface through which to approach the particular similarities and differences in a given text. In this short paper we consider from the theoretical perspective some hermeneutic risks inherent in the static visualization most commonly used at present, when the interface stands between the scholar and the data model (whether produced by a collation tool such as CollateX or NMerge, or generated from a published critical apparatus or text collation). The result of our application of the theory is the tool we present, meant to address these risks in part. Implemented in JavaScript[1] , the tool allows the user to interact with and modify the model of a particular text from within the browser window. By doing so the scholar can move beyond the 'black box' nature of visualizations produced by analysis tools and test alternate hypotheses for understanding a text.

Most visualizations of data analysis results are by their nature static, and the variant graph has so far been no exception. A typical example of a variant graph visualization is given in Figure 1, which shows part of a collation result of a line of text obtained from CollateX. (Example adapted from the prologue of the Canterbury Tales (Solopova, 2000; Robinson 2000).):

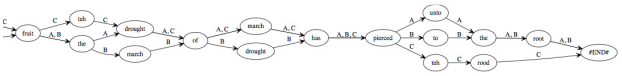


Figure 1:
Partial collation of a line of text of the Canterbury Tales' prologue.

Each of the manuscript witnesses A, B, and C takes a single path through the possible readings, indicating the version of the text that appears in the witness itself. The variant graph is a compelling and concise means by which we can display the phenomenon of text variation. However, from the given example it is clear that there remains room for further interpretation: should spelling variants such as 'teh' and 'the' always be represented as having equal impact on the text with transpositions such as the one between 'drought' and 'march'? Does the alignment of witness C's 'teh rood' make sense as it stands?

The initial model is thus not entirely satisfactory; there is a clear need for the scholar to express more information or to express the variation more exactly. As long as the graph remains static, our example reflects a general danger inherent to visualizations of digital models of humanities data. An interface is meant to make an internal computer model tractable to a user, but paradoxically it also imposes a barrier: it most literally stands between the user and the model. The user can not change the model; he or she can basically only inspect it insofar as the interface visualizes it. In this sense the interface becomes a display practice enforcing one-way communication (Rijcke & Beaulieu 2011). The extent to which a visualization is interactive and alterable will consequently have an inverse relationship to the impression of finality or correctness that the user will come away with. A high degree of interactivity arguably conveys an impression of interpretability and malleability of the data and results visualized. In contrast, a static visualization produced by a computer tool carries with it an aura of correctness or immutability — even if it is not the only such visualization. If the scholar/user is unable easily to interact dynamically with the model, or even to modify it, then the danger arises that the visualization as posited by the computer analysis will be regarded uncritically and accorded a problematic or even false authority.

Current automated collation practices are a good example of this phenomenon. If an automatic collation algorithm such as that employed by CollateX chooses arbitrarily between two equally good text alignment scenarios, then unless the user is able not only to visualize the results but to manipulate them and explore the equally valid alternative scenarios, the 'answer' produced by the computer will tend to be seen as the 'correct' collation. Interactivity can thus be seen as an inherent and indispensable part of any sort of modeling of humanities data if the model is to be accepted, refined and used more widely in the field. In the case of variant graphs resulting from computed collation, essential interactions include:

- annotating variants with information about how they are related
- combining multiple words into one reading (e.g. 'what/so/ever') to better express the mutation as it occurred
- splitting words into multiple readings to show more finely-grained variation
- altering or correcting an initial collation

To enable and enhance such interactivity we present a self-contained and generalized JavaScript library for the interaction with and modification of variant graphs produced by online collation services directly in the browser. The library, currently in use by the Greek New Testament Editio Critica Maior project at the Institut für Neutestamentliche Textforschung (INTF)in Münster[2] and the 'Stemmaweb' application of the Tree of Texts project at KU Leuven[3] , can be applied to any research context wherein a variant graph would be appropriate. Users can relate parallel variants, merge or split variant readings, and regularize variants where appropriate to more easily home in on the variants under investigation. Figure 2 shows our variant graph example annotated with different types of relationship.

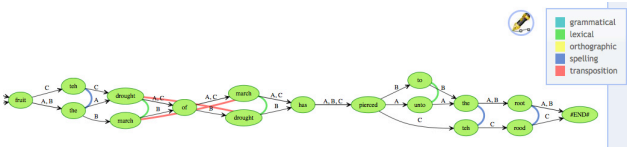


Figure 2:
Partial collation of a line of text of the Canterbury Tales' prologue annotated with various types of relationship.

The scholar may make explicit his or her model of variation by defining the types of relationships that can occur between variants, as well as criteria concerning when these variants can or cannot apply (must the variant readings occur in parallel within the text? must they *not* occur in parallel, e.g. for transpositions? Does the existence of one relationship imply the existence of another?) These

interactions with particular instances of a variant text model allow users to adapt, correct, and extend the results of an automated collation engine, thus making the results produced by the software less fixed and more open to scholarly interpretation. To further support interactivity with the model we experiment with methods of machine recognition of user adjustments in context — for instance, we scan the graph using a vector cosine measure that takes into account the two nodes of a user defined relationship and their existing direct adjacent nodes to identify contexts in the graph that are similar to the context where the user adapts the graph, producing automatic suggestions for where the same relationship might apply under the same conditions elsewhere in the graph.

The clear advantage of enabling interactivity is that the problem of one-way communication between visualization and user is resolved: the variant graph model of a particular text becomes a much more tangible and expressive means to engage with and to conduct research into the text in question.

The interactive engagement with a specific instance of a computer model is a first step in our trajectory to enhance interactivity with current state-of-the-art text collation tools. Future research and development work will focus on feeding instance-based manipulations of collation results back into the reference models of online collation services. As a scholar works with a particular graph, the interactions with that instance of the collation model will be captured by decision trees. These trees can be used by the processes backing the collation model to amend the collation and graph construction, not only for specific instances but also for the general model, thus changing collation engines from rather static rule-based automatons into adaptive algorithms that learn from expert input and feedback.

References

Andrews, T. L., and C. Macé. (2013). Beyond the tree of texts: Building an empirical model of scribal variation through graph analysis of texts and stemmata. *Literary and Linguistic Computing.*

Haentjens-Dekker, R., and G. Middell (2011). *Computer-Supported Collation with CollateX: Managing Textual Variance in an Environment with Varying Requirements. Supporting Digital Humanities 2011.* University of Copenhagen, Denmark, 17–18 November 2011.

De Rijcke, S., and A. Beaulieu (2011). Image as Interface: Consequences for Users of Museum Knowledge *Library Trends*. 59(4): 663–685.

Robinson, P. (2000). New methods of editing, exploring, and reading *The Canterbury Tales.* http://www.canterburytalesproject.org/pubs/desc2.html (accessed 23 October 2012).

Schmidt, D., and R. Colomb (2009). A data structure for representing multi-version texts online. *International Journal of Human-Computer Studies* 67(6). 497–514. http://dx.doi.org/10.1016/j.ijhcs.2009.02.001

Solopova, E. (ed.) (2000). Chaucer: The General Prologue on CD-ROM. The Canterbury Tales on CD-ROM. Cambridge: Cambridge University Press.

Notes

1. http://en.wikipedia.org/wiki/JavaScript
2. http://egora.uni-muenster.de/intf/aecm/aecm_en.shtml
3. http://treeoftexts.arts.kuleuven.be/

Using the Social Web to Explore the Online Discourse and Memory of the Civil War

Appleford, Simon
simonja@clemson.edu
Clemson University, United States of America

Thatcher, Jason
jason.b.thatcher@hotmail.com
Clemson University, United States of America

Overview

Some humanities scholars argue that the social web "poses a grave threat to the humanities because it lacks the depth, nuance and permanence that make genuine, meaningful interactions about the human condition possible." They fear that "we risk becoming a world without the comprehensive communication tools needed to keep the humanities alive" (Adamek 2010). Even those who embrace social media all too often dismiss it as a tool that is used primarily for community engagement and self-promotion (Terras 2012). Yet the social web — which we broadly define as the array of technologies that allow individuals to post their thoughts, pictures, and comments in a public forum — when coupled with recent advances in cloud computing, data management and statistical/visual analysis

offers significant potential to explore new and enduring humanities questions.

Through careful, rigorous analysis, we believe that the social web affords opportunities for the humanities to realize a contemporary, nuanced understanding of how the public believes our past informs modern society. For example, although there is a strong consensus amongst historians, the broader American public remains conflicted, divided, and confused about the causes of the Civil War. This is evident in any number of recent public polls, such as an April 2011 CNN poll which found that 42% of Americans believed that cause was something other than slavery, and the April 18, 2011 issue of Time, whose cover said "Why We're Still Fighting the Civil War: The Endless Battle over the War's True Cause Would Make Lincoln Weep" (CNN 2011; von Drehle 2011). Where polls provide a snapshot of Americans' views, analysis of online discourse and social media activity affords opportunities to explore modern attitudes towards issues surrounding the Civil War.

Hence, against the backdrop of the on-going sesquicentennial commemorations, this paper examines the intersection of humanities, social sciences, social media, and computing to probe enduring questions around the legacy of the Civil War. To do so, we will illustrate how an ongoing study that uses the Civil War serves as a testbed for examining and developing techniques to conduct traditional types of humanistic inquiry in the context of the social web. Our results demonstrate how careful analysis of the online discourse that occurs across the social web enables deeper understanding of how the Lost Cause Ideology continues to recast the origins of the war and minimize the role of slavery. Simultaneously, we explore how stereotypes of Southerners continue to be propagated and used to shape a memory of the Civil War.

To conduct this study, we demonstrate how contemporary tools developed by industry may be used to analyse the social discourse around the Lost Cause and stereotypes of Southerners. Most scholars researching the social web rely primarily on the use of single keyword search and user-specified hashtags to narrow the size of their datasets to only posts that are immediately relevant to the topic under discussion (for example, Graham 2012; Ross, et al. 2010). However, these methods can be inadequate or problematic for understanding widely diffused social phenomena. For example, when attempting to study a topic such as the Civil War or especially a concept such as "the South" this methodology isn't adequate as the majority of users do not tag their casual online conversations with these types of metadata nor do they restrict their conversations to a single platform (e.g., Twitter, Facebook etc.). To elicit a holistic view of conversations on the social web, the researcher has to employ strategies that reach across social media platforms and go beyond simple hashtag sets. To overcome these limitations, we leverage software originally developed for business analytics to aggregate content from across the social web (Radian6, 2012) — our search includes not just Twitter and Facebook, but also content from sources such as mainstream news sites, blog posts, and forum posts.

Method

Through our analysis of the social web, we highlight issues and opportunities for scholars who work at the intersection of the humanities, social science, social media, and computing. One issue that must be overcome is that, on a global scale, there are many Souths: not only must we find a way of eliminating references to countries or continents such as South Africa, South America, and the South Pole, we must also filter out conversations related to places such as the South Side of Chicago and even the television show South Park, none of which would be relevant for this research. Similarly, while most Americans would refer to the American Civil War as simply "The Civil War," there are many similarly named historic and ongoing conflicts that we must filter from our results. The solution is to create "topic profiles" — collections of words that fall into one of three categories: words or phrases that must be present in a post to be included in the results; phrases that must be present along with words from the first category to help categorize results into different topics; and, finally, phrases that, if present in a post, will result in that post being excluded from our results. So, when conducting research into online discussions of southern gender ideals, we create a topic profile similar to the following:

1. Posts that CONTAIN any of the following: "the south", "thesouth", "southern"
2. AND CONTAINS any of the following: "belle", "lady", "gentleman", …
3. But DOES NOT CONTAIN any of: "south pole", "south park", "south side", "south africa", …

By using keywords that were selected following a survey of users of the H-South discussion list, a request from leading scholars (including several former presidents of the Southern Historical Association, of the American South, surveys of a graduate and an undergraduate history class on Southern History at Clemson University, as well as a non- random sample of black and white southerners who were not academics, we have constructed topic profiles for five different themes that might be used to segment contemporary discussion of the South online: Southern Culture (which includes concepts such as honor, hospitality, and accents), Southern Food, Discussions of Gender, Religion, and Southern History. This set of keyword groups

results in approximately 300,000 unique hits from the social web each month. Figure 1 illustrates the relative volume of each keyword group during September 2012 and shows how the number of hits varies for each topic across the month. Similar keyword groups have also been constructed for specific events that occurred during the Civil War, such as the attack on Fort Sumter, the Emancipation Proclamation, and the Battle of Gettysburg.

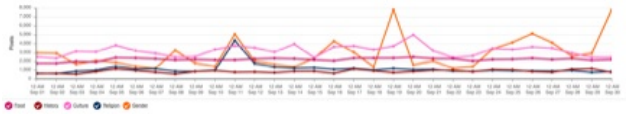


Figure 1:
Relative Volume of Five Topic Profiles for September 2012

Analysis

Drilling into this data reveals numerous insights into how users of the social web conceptualize the South and what topics are of interest. Figure 2(a), for example, presents a word cloud of the major terms associated with civil war commemorations, again for the month of September 2012. This word cloud immediately reveals several potential avenues for further investigation, including the centrality of Gettysburg to the online discourse surrounding the Civil War, but also suggests other useful topics to explore. Indeed, although word clouds have frequently been criticized for divorcing words from their context (Harris, 2011), we are able to maintain the connection between the words represented in a cloud and the underlying data. Therefore, if we are interested in further exploring the appearance of the word "history", we can, as shown in Figure 2(b), drill further into the data to reveal new levels of insight to our topics. At all times, the original posts remain available both to view at an individual level to ensure the relevancy of content and to download for further textual and statistical analysis with dedicated software packages such as R and Gephi.

Figure 2:
(a) Word Cloud of Frequently Used Words in the Civil War Profile for September 2012; (b) The same word cloud, instead focused on the word "history"

We can also analyse demographic information, such as age, gender, and location, from the posts we have collected, allowing for a more nuanced understanding of our results. Figure 3 illustrates the location of users in the United States whose posts and tweets were collected from the Civil War topic profile over a seven-day period in mid- September 2012, while Figure 4 presents word clouds of the conversations that occurred in our Civil War topic profile on the 150th Anniversary of the Battle of Antietam by (a) people located in Maryland; (b) those in North Carolina; and (c) users aged between 36 and 45. These charts reveal not only the geographic distribution and volume of conversations, but also regional and demographic differences in the topic, tone, and scope of this online discourse.

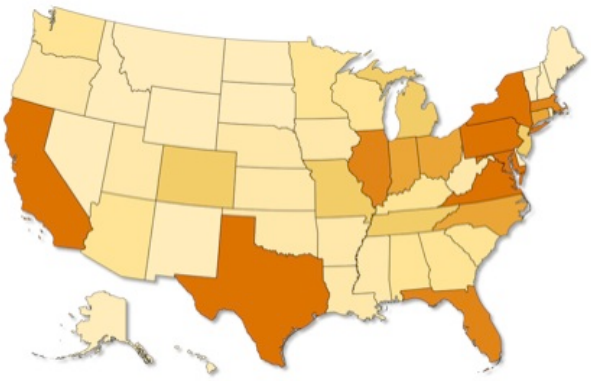


Figure 3:
Geographic Distribution of Unique Hits from the Civil War Topic Profile in September 2012. A darker shade of orange indicates a greater number of posts from that state.

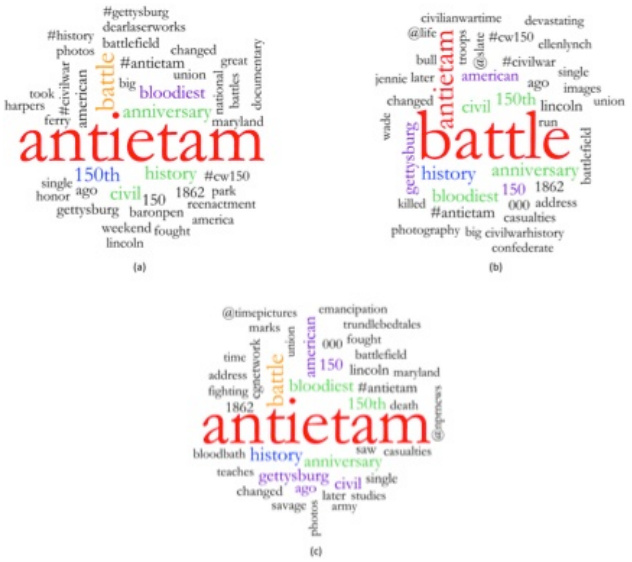


Figure 4:
Word Cloud of conversations that occurred on the 150th Anniversary of the Battle of Antietam by (a) people located in Maryland; (b) those in North Carolina; and (c) users aged between 36 and 45.

References

Adamek, D. (2010). Social Media Flaws and the Humanities. *Valley Advocate,* 13 December. http://www.valleyadvocate.com/blogs/home.cfm?aid=12863 (accessed 10 October 2012).

CNN Political Unit. (2011.) Civil War Still Divides Americans. *CNN,* 11 April. http://politicalticker.blogs.cnn.com/2011/04/12/civil-war-still-divides-americans/ (accessed 11 April 2011).

Graham, S. (2012). Mining the Open Web With Looted Heritage — Draft. *Electric Archaeology,* 8 June 2012. http://electricarchaeology.ca/2012/06/08/mining-the-open-web-with-looted-heritage-draft/ (accessed 15 October 2012).

Harris, J. (2011). Word Clouds Considered Harmful. *Nieman Journalism Lab,* 13 October. http://www.niemanlab.org/2011/10/word-clouds-considered-harmful/ (10 October 2012).

Radian6 (2012). http://www.radian6.com. accessed 10 October 2012.

Ross, C., M. Terras, C. Warwick, and A. Welsh. (2010). Pointless Babble or Enabled Backchannel: Conference Use of Twitter by Digital Humanists. in*2010 Digital Humanities Conference.* http://www.ucl.ac.uk/infostudies/claire-ross/Digitally_Enabled_Backchannel.pdf

Terras, M. (2012). The Impact of Social Media on the Dissemination of Research: Results of an Experiment. *Journal of Digital Humanities,* 1(3) http://journalofdigitalhumanities.org/1-3/the-impact-of-social-media-on-the-dissemination-of-research-by-melissa-terras/ (accessed 10 October 2012).

Von Drehle, D. (2011). 150 Years After Fort Sumter: Why We're Still Fighting the Civil War. *Time,* 18 April. http://www.time.com/time/magazine/article/0,9171,2063869,00.html (accessed 18 April 2011).

Memoragram, a service for collective remembrance

Arauco Dextre, Renzo
renzoarauco@gmail.com
Memoragram, Perú

Introduccion

Desde la década de los 80s hasta los 2000 el Perú vivió en medio de un conflicto armado entre dos movimientos terroristas, Sendero Luminoso y el Movimiento Revolucionario Túpac Amaru (MRTA), los cuales buscaban derrocar el sistema democrático peruano, desatando uno de los más violentos pasajes de la historia moderna.

Estos movimientos fueron vencidos militarmente, desarmados y encarcelados, pero a un alto costo en vidas humanas. A pesar de la desarticulación de estos grupos aún se reportan células armadas y organizadas en los lugares más alejados de la selva peruana. Al mismo tiempo, en las ciudades buscan impunidad criminal y política para sus líderes arrestados a través de mecanismos legales que les permitirían introducirse en la política activa.

Hoy, casi 10 años después del fin de este episodio de la historia, el museo que conmemora el recuerdo de las víctimas de la violencia, el Lugar de la Memoria del Perú, no ha sido terminado. Las nuevas generaciones no conocen claramente lo sucedido debido al deficiente programa educativo; tal que estos grupos subversivos aprovechan la desinformación para entregar una imagen distorsionada de esta parte de la historia. (Pisa, 2009). Sin embargo, somos la octava comunidad con mayor uso de las redes sociales (Comscore Inc. 2012) e Internet en el mundo. Asimismo, Perú ya ha sobrepasado el 100% de penetración móvil. (Budde.com 2012)

Estos índices muestran que Perú tiene una oportunidad única y sin precedentes de desplegar servicios que asistan al desarrollo y mantenimiento de una memoria colectiva usando tecnologías móviles.

La Evolucion De La Idea

La misión del proyecto evolucionó con el tiempo y se ha trabajado y conversado con expertos del tema.

Un museo virtual

Memoragram nace a partir de la idea de crear un museo geo-localizado de la memoria, accesible desde cualquier dispositivo sin importar la hora ni el lugar. Los usuarios podrían explorar eventos históricos cercanos a su ubicación actual o hacia la fecha que deseen consultar, revisar los personajes involucrados, los lugares (en su mayoría demolidos), organizaciones e incluso revisar qué libros, revistas, periódicos, películas, reportajes de TV, cómics, entre otras fuentes, hacen referencia al hecho.

Por el lado tecnológico, los usuarios podrían usar smartphones o tablets para que, al acceder a una web móvil, esta solicite permiso para usar el GPS del dispositivo y así mostrar contenido relevante. La tecnología que hace posible este acceso es HTML5, CS3 y Javascript. Asimismo, los usuarios podrían descargar una tercera aplicación móvil llamada Junaio, la cual, usando el GPS, giroscopios, brújula y cámara puede mostrar contenido en una vista de tiempo real.

Si bien esta parte del alcance no ha sido modificada, surgieron otros problemas fuera del campo tecnológico.

El problema de las fuentes de datos

Con la idea y el alcance del proyecto se empezó por buscar fuentes de información fiables, entre ellas el Banco de Imágenes de la Comisión de la Verdad y Reconciliación Nacional del Perú (Comisión de la Verdad y Reconciliación del Perú), la cual se basa en los archivos fotográficos de los principales diarios y revistas locales peruanos. Este mismo banco se usó para formar la muestra *Yuyanapaq* (Programa de las Naciones Unidas para el Desarrollo — PNUD Perú) (término quechua que significa 'Para recordar'). Este archivo posee alrededor de 1560 fotografías.

Luego de una evaluación de esta fuente y del mapeo de eventos posibles a registrar se concluyó que el número de eventos registrables sería mucho menor a 1560. De encontrarse otras fuentes de datos el número sería más limitado. Esta situación conllevaría a que la oferta de contenido fuera fija, lo cual podría evitar que los usuarios no se familiaricen con el sistema o no tengan una razón para regresar al servicio.

El fenómeno social del traspaso de la memoria y LoSoMo

La memoria colectiva se da como un proceso social y natural en el que las generaciones mayores traspasan conocimiento a las generaciones más jóvenes a través de relatos, documentos, mapas, elementos multimedia como fotos y videos, entre otros.

Este acto puede tener un efecto terapéutico en las personas que vivieron hechos traumatizantes.

A la vez, en el entorno de marketing en Internet (campo en donde se desenvuelve el autor) se hace presente la tendencia llamada "LoSoMo" que son las siglas de tres elementos que se recomienda considerar en una campaña digital: "Local", "Social" y "Mobile". Este enfoque se puede encontrar en las más exitosas start-ups como Foursquare y Groupon. A este momento, el alcance de Memoragram sólo abarcaba el primer y último aspecto. El fenómeno de traspaso de la memoria entonces se postulaba como el elemento social para generar una comunidad activa de usuarios.

Nuevas posibilidades

La inclusión de un fenómeno social en el alcance del proyecto abrió un nuevo espectro de posibilidades para la idea de formar un museo virtual. El sistema empezó a evolucionar: pasó de un servicio en el cual los usuarios sólo consumen contenido a una plataforma donde pueden aportar con sus propias memorias, sus propios recuerdos.

Este nuevo tipo de contenido obliga a generar nuevas acciones que se pueden expresar y medir con la plataforma, por ejemplo: los usuarios exploran el contenido y pueden registrar y medir el efecto que este genera en ellos, logrando que esta medida de rankings sea otro criterio de organización de los eventos listados.

El Proyecto

Con las consideraciones listadas previamente se definió el alcance del proyecto, el cual se alcanzará a través de una serie de prototipos funcionales en proceso de perfeccionamiento.

El alance elegido

Los usuarios podrán explorar los eventos históricos más cercanos a su ubicación actual, la fecha en que sucedieron, los personajes, organizaciones y las publicaciones que

las mencionan como fotografías, video, audios, archivos periodísticos, libros, películas, entre otros. Asimismo podrán dejar sus propios recuerdos usando texto, audio, fotografías e incluso video, a manera de mensaje a las futuras generaciones (recuerdos públicos).

El contenido al inicio será creado con la asesoría de especialistas basados en fuentes confiables.

La Solucion

Eligiendo el gestor de contenido base

Para construir la plataforma se requiere de un paquete de software web modular, preferentemente gratuito para ahorrar costos y open-source para aprovechar el avance de la comunidad que lo mantiene. Se eligió entonces el CMS (Sistema de gestión de contenido) Drupal 7 basado en el lenguaje PHP, el cual se conecta a una base de datos MySQL (Metaio Inc.) Este sistema ya incorpora funciones de manejo de registro de usuarios, creación de perfiles de usuario, tipos de contenido y una robusta estructura de etiquetas así como soporte de distintos tipos de archivos multimedia.

Interfaces del usuario

En la versión 7 de Drupal ya se maneja HTML versión 5, llamadas asíncronas AJAX e incluso soporte de diseño responsivo, el cual permite crear una sola interfaz que adapta el contenido mostrado según el dispositivo desde el que se accede a él.

Asimismo, se decidió experimentar con una interfaz de realidad aumentada para móviles; para ello se eligió la aplicación móvil Junaio www.junaio.com, disponible para iOS y Android, a la cual basta alimentar con contenido desde una interfaz entre servidores usando XML.

Desarrollo de prototipo

A continuación pueden observarse unas capturas de pantalla. Todas las imágenes son de elaboración propia, se realizaron desde una PC con una navegador web y en un Apple iPhone 3GS.

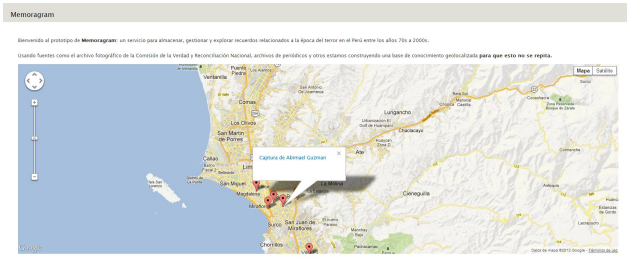


Figura 1.
Home del sitio web prototipo con mapa de los últimos eventos registrados (Mostrando un pin del mapa con burbuja de información)

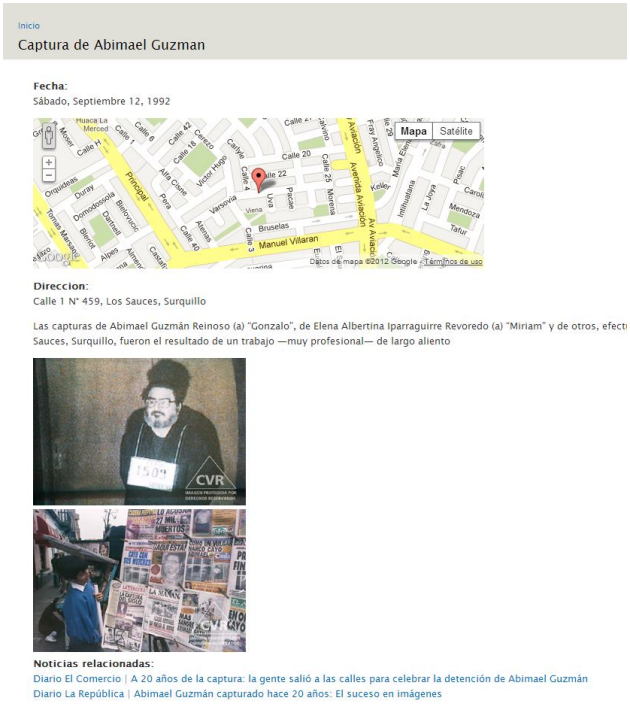


Figura 2.
Captura de pantalla web del detalle de un evento

Figura 3.
Vista en vivo en interfaz de realidad aumentada móvil usando la aplicación Junaio en un iPhone 3GS desde la Plaza Mayor de Lima

Figura 4.
Vista de mapa usando la aplicación Junaio (navegando a través de eventos)

Figura 5.
Vista de lista usando la aplicación Junaio (navegando a través de eventos)

Figura 6.
Vista de detalle de un evento usando la aplicación Junaio (los usuarios pueden realizar distintas acciones como obtener la ruta para llegar al punto, ver la imagen, entre otras)

Conclusiones

Es técnicamente viable crear una plataforma que llegue al alcance propuesto, sin embargo, a la fecha sólo se ha implementado el acceso web y realidad aumentada móvil sin la interfaz de creación de memorias por parte de usuarios. En el futuro cercano se irán extendiendo las funcionalidades y luego se pasará a realizar pruebas de usuario para diseñar una interfaz adecuada.

References

1. PISA 2009 Results: Executive Summary (Figura 1), OECD, 2010. Consultado el 2012-11-04. http://www.oecd.org/pisa/46643496.pdf.

2. Comscore Inc. “It’s a Social World: Social Networking Leads as Top Online Activity Globally, Accounting for 1 in Every 5 Online Minutes”. Consultado el 2012-11-04. http://tinyurl.com/bgn3sln

3. BuddeComm. “Peru - Telecoms, Mobile, Broadband and Forecasts”. Consultado el 2012-11-04. http://www.budde.com.au/Research/Peru-Telecoms-Mobile-Broadband-and-Forecasts.html.

4. Comisión de la Verdad y Reconciliación del Perú. Banco de Imágenes de la Comisión de la Verdad y Reconciliación Nacional del Perú. http://www2.memoriaparalosderechoshumanos.pe/apublicas/galeria/index.php

5. Programa de las Naciones Unidas para el Desarrollo - PNUD Perú. “Yuyanapaq. Para recordar”. Consultado el 2012-11-04. http://www.pnud.org.pe/yuyanapaq/yuyanapaq.html

6. Metaio Inc. “Junaio Demo Book”. http://www.junaio.com/fileadmin/upload/documents/Promo_Booklet/DOC-junaio_promo_book-EN-DIGI.pdf

CULTURA: Supporting Professional Humanities Researchers

Bailey, Eoin
baileyeo@scss.tcd.ie
Knowledge and Data Engineering Group, Trinity College Dublin, Ireland

Sweetnam, Mark
sweetnam@tcd.ie
Department of History, Trinity College Dublin, Ireland

Ó Siochrú, Micheál
OSIOCHRM@tcd.ie
Department of History, Trinity College Dublin, Ireland

Conlan, Owen
owen.conlan@cs.tcd.ie
Knowledge and Data Engineering Group, Trinity College Dublin, Ireland

A key challenge facing professional researchers in the domain of cultural heritage across Europe and worldwide is the interrogation of growing digital humanities collections. However, the full value of these heritage treasures is not being realised. After digitisation, these collections are typically monolithic, difficult to navigate and can contain text which is highly variable in terms of language, spelling, punctuation, and consistency of terminology. These difficulties are compounded by a lack of normalised spelling in most European languages before the eighteenth century. This means that search across these digital collections tends to return sub-par results as multiple spellings for many common words are treated as independent document keywords. CULTURA is a corpus agnostic environment with a suite of services, including personalisation, annotation, and recommendation, providing necessary supports and features for a diverse range of professional researchers.

In order to empower communities of researchers with personalised mechanisms which support the collaborative exploration, interrogation and interpretation of complex digital cultural artefacts, the adaptivity provided in CULTURA is required to be integrated and intelligent. The next generation adaptivity provided by CULTURA supports the dynamic composition and presentation of digital cultural heritage resources. Automated adaptivity however, is not enough on its own. Ensuring that the user is in control of the personalisation process is essential. Such user-centred control is enhanced through correlating usage patterns with self-expressed user goals; pre-defined strategies (e.g. research strategies, investigation strategies, discovery strategies, explanatory strategies etc.); and the provision of appropriate tools for users to explore and navigate large cultural heritage information spaces.

A central aspect of the CULTURA environment is its use of rich metadata (user generated, computer generated and expert generated) coupled with natural language processing, entity extraction and social network analysis techniques, in order to support collaborative exploration, interrogation and interpretation of the underlying cultural resources. The Qviz (http://www.qviz.eu) project has some similarities in approach to the CULTURA project in that it makes explicit recognition of the value of users as members of communities, and as contributors to digital cultural heritage collections, however Qviz, however, does not incorporate a personalised or adaptive aspect.

The manual determination of descriptive metadata across a large corpus is too time-consuming to be practical. For example, the process of metadata identification for the 8,000 1641 depositions took a research assistant 12 months (http://1641.tcd.ie). An automated entity extraction process is used by the CULTURA project (Carmel et al. 2012). This process interprets words and combinations of words to identify entities in the corpus such as people, places, events, and dates. Complex entities are then constructed from these entities and the corpus allowing the identification of events, such as *WHO* did *WHAT* on such a *DATE* at a specific *LOCATION*.

Entity extraction is most powerful on a corpus in which all the content has been normalised as entities can be matched across multiple documents. To enable normalisation a ground-truth is manually generated across a proportion of the corpus, approximately 10% in the case of the 1641 depositions (Hampson et al. 2012). The ground-truth assigns non-normalised terms in the corpus to normalised terms. The ground-truth is applied in the generation of a statistical model that is then utilised on the entire corpus outputting a normalised corpus. It is this normalised content from which entities are identified and extracted. The linguistic model used in the normalisation of the 1641 Depositions has proved highly reusable, and provides robust results when applied to a range of other material contemporary to the corpus used to generate the model, enabling the re-use of the model on additional content collections.

The output of these processes ensures each individual piece of content from the corpus has descriptive metadata in the form of individual entities such as a person or place, and complex entities such as where a person lived (i.e. compound entities) as well as a normalised variant. This enables a simple keyword search to provide results that cover the entire corpus and all variants of spelling of the terms entered. Additionally, and more significantly, individual pieces of content from within the corpus can be linked to other content that contains information on the same event or similar events, based on the date, location, people, and type of event. These links enable professional researchers to quickly identify content that is related to their current research topic, by way of visualisations of these links (Hampson et.al. 2012).

While the automated tools are providing useful results for professional researchers, these state-of-the-art tools cannot replace the insight and experience of a professional researcher. These insights can be captured via an annotation tool that can be used to annotate specific aspects of any piece of content, both textual and visual. Annotations can be at any level, from a single word up to the entire document, or any user-identified region of an image. Annotations also allow the researcher to link the identified content to any other document within the corpus. Links such as this feed a professional researcher's knowledge into the system, and can be used to aid in the adaptivity and personalisation of the system for all users.

Collaboration between professional researchers occurs in an implicit manner via the use of annotations as link generation between content, which feeds into recommendations and personalisation. Explicit collaboration is also present in CULTURA. Researchers can share annotations with other users, enabling the propagation of insights and discoveries in the content. As annotations are anchored to specific elements of a document they provide a powerful mechanism for inserting detailed and relevant knowledge. This knowledge can be made available to either groups of researchers or all researchers thus enabling a greater collaboration across all users of the environment.

The CULTURA environment has already been used by professional researchers in the course of their research. Services including annotations, document level notes, and multi-dimensional recommendations of content were enabled within the environment.

A number of the professional researchers involved in trials of the CULTURA environment were initially wary of the potential impact of adaptivity on their research. They expressed concern that the recommendation system would create a 'filter bubble', distorting the appearance of the collection. As they engaged with the environment, however, these concerns were abated. Research into the user requirements of these end-users had identified this concern, and a concomitant need for a high level of scrutability. In line with this, recommendations made by the CULTURA engine were very explicitly presented in a way that made it easy for researchers to decide whether to use them, or to ignore their suggestions. In addition, the interface made it clear why a given user was being recommended a particular piece of content.

After some initial caution, professional researchers agreed that the adaptive recommendations provided by CULTURA were genuinely useful. This was especially true of researchers who were previously unfamiliar with the CULTURA corpora. These users reported that the recommendations facilitated their mastery of the collection. Researchers who had greater familiarity with the collection rated the utility of the recommendations less highly, but still expressed an appreciation of their potential usefulness. In particular, they suggested that the recommendations were valuable in encouraging them to look at the collection in new ways.

The user model for the professional researchers was evaluated and determined to be an accurate reflection of the researcher's interests and topics under research. While some anomalies were identified in the user model, the areas of interest that were weighted most highly correlated with the topics of the research.

The CULTURA environment is currently engaged in additional evaluations with trainee researchers utilising a broader set of documents, tools, and adaptivity.

References

http://www.qviz.eu

http://1641.tcd.ie

Carmel, D., N. Zwerdling, and S. Yogev (2012). 'Entity oriented search and exploration for cultural heritage collections: the EU cultura project'. In *Proceedings of the*

21st international conference companion on World Wide Web. New York: ACM. 227-230 .

Hampson, C., M. Agosti, N. Orio, E. Bailey, S. Lawless, O. Conlan., and V. Wade (2012). CULTURA Project: Supporting Next Generation Interaction with Digital Cultural Heritage Collections. EUROMED 2012, International Conference on Cultural Heritage. Lemesos, Cyprus. To Appear.

Inferring Social Rank in an Old Assyrian Trade Network

Bamman, David
dbamman@cs.cmu.edu
Carnegie Mellon University

Anderson, Adam
aganders@fas.harvard.edu
Harvard University

Smith, Noah A.
nasmith@cs.cmu.edu
Carnegie Mellon University

1. Introduction

In the early 20th century, the attention of Assyriologists and archaeologists was directed to a number of cuneiform tablets coming from a remote archaeological tell in Kültepe, Turkey. After the first series of excavations, archaeologists discovered a large collection of texts and the remains of a Bronze Age trade colony, referred to in the texts as kārum Kaneš. Once these initial ca. 5,000 texts were deciphered, the field of Old Assyrian studies was born. In 1948 official Turkish excavations began at Kültepe and added over 17,000 tablets to the Old Assyrian text corpus, which now totals ca. 23,000 cuneiform tablets (Michel, 2008). These texts document the intricacies of thriving Bronze Age trade networks, comprised of Old Assyrian merchants from the ancient city of Assur approximately 4,000 years ago (ca. 1950-1750 BCE) (Barjamovic et al., 2012). The texts further show how the merchants acted as the middle-men in a large series of inter-connected networks which, among other things, linked the natural resources of tin (in Iran and Afghanistan) and copper (in Turkey) in order to produce bronze in Anatolia.

However, one thing the texts do not make clear is the scope and structure of the colonial trade network, in terms of the people involved and their organization. Although the high degree of literacy among the inhabitants of the colony at Kaneš helped create an extremely rich source of texts illustrating the daily life of the people involved, the practice of paponomy (naming a son after his grandfather) has obscured the identities of the merchants for modern scholarship. Thus, due to the density and ambiguity of the names mentioned in these texts, it has been too difficult to gain an understanding of the scope of the colonial society on the basis of the textual record at Kültepe.

Our work therefore focuses on jointly inferring the unique individuals as well as their social rank within the Old Assyrian trade network, using a novel probabilistic latent-variable model that exploits partial rank information contained in the texts.

2. Data

Of the 23,000 tablets unearthed at Kültepe (Kaneš), 5,691 published texts (known as the "old texts" (Veenhof and Eidem, 2008)) have been digitized and transcribed into machine-actionable text as part of the Old Assyrian Text Project (Old Assyrian Project). While these tablets include mostly economic and legal transactions, 2,094 of them are letters between merchants. Along with the body of the Akkadian text, each of these letters includes a highly stylized introductory formula (an "epistolary formula") which lists the senders and recipients using strict dominance rules concerning the order of the names. For example:

umma	Aššur-idī	Aššur-nādā	ana	Amur-Ištar	Alāhum	Aššur-taklāku
from	Aššur-idī	and Aššur-nādā	to	Amur-Ištar,	Alāhum	and Aššur-taklāku

These formulae have a consistent internal structure from which we can draw relative social ranks among the individuals involved. Each formula can be divided into two parts (a *receiving* rank and a *sending* rank), and an individual placed linearly after another *within* one of these ranks cannot be socially higher than any mentioned before (whether the first is higher or equal is ambiguous). Additionally, one individual (who may be either among the senders or recipients) is mentioned first in the letter, a marked position signifying the highest social status of those mentioned in either rank.

These partial orderings provide a rich source of evidence for the global social structure; from this example, we can extract seven pairwise partial social orderings: Amur-Ištar ≥Alāhum and Aššur-taklāku; Alāhum ≥ Aššur-taklāku; and Aššur-idī ≥ all four of the others.

If all such partial orderings were to be trusted, if each observed name in such a formula were unambiguous, and if social power were a stationary quality that remained constant over time, inferring a consistent global rank over all individuals would be easy (though more than one such rank may be possible). Unfortunately, however, none of these assumptions are true. The rank we observe in one letter is a subjective judgment by the author, and we can easily imagine that complex social dynamics are involved in the choice of who to rank highest (which can vary by author); names are indeed ambiguous with one name potentially referring to multiple people, and the same person can be known by several names; and the letters span a period of ca. 200 years, over which time a young individual with low social rank can age and accrue power[1] .

3. Technical Approach

Our goal is to find the social ranking over individuals (possibly not in a one-to-one relation with names) that best explains the observed data. To illustrate the intuition behind our approach, consider a simple example. Suppose we are trying to establish the temporal rank of a set of individuals with the names ADAMS, JEFFERSON and MADISON, and we have the following evidence (where > indicates "was president before").

- ADAMS > JEFFERSON
- JEFFERSON > MADISON

Assuming transitivity, a sound global rank among these three is: ADAMS > JEFFERSON > MADISON; while we never directly witness a statement of the sort ADAMS > MADISON, we can infer it through intermediary relations. Now suppose we observe an additional piece of evidence:

- MADISON > ADAMS

If we assume that the three names only refer to three distinct individuals, transitivity breaks down: putting all three statements together results in a circular rank, leading to the contradiction that ADAMS > JEFFERSON while at the same time JEFFERSON > ADAMS. However, we can establish a sound global order if we allow the three names to refer to four individuals (e.g., two people both have the name ADAMS), resulting in the rank: ADAMS1 > JEFFERSON > MADISON > ADAMS [2] . In fact, given the inconsistency of the evidence under the assumption of only three people, the existence of four underlying people is in fact more likely. Here, our method offers an informed hypothesis—that Adams refers to two distinct individuals rather than one—that can be verified (or refuted) in consultation with the data. In this simple case, the hypothesis is supported by the fact that Adams can refer to both JOHN ADAMS and JOHN QUINCY ADAMS.

In the case of our Old Assyrian dataset, the evidence takes the form of 4,191 pairwise observed ranks of 717 individual names in 1,657 letters, along with the Akkadian text of those letters. Our task is to find the most likely overall social rank of a fixed set of actors that best explains the pairwise ranks in letters that we see. In casting the problem in this way, we are building a *probabilistic generative model* of the data. Latent Dirichlet allocation (Blei et at., 2003) is a familiar example of a generative model that seeks to explain data in text documents by inferring latent topic assignments to individual words. In our case, the latent variables are a.) the identities of the people named in each letter; and b.) the social rank of those people, represented as continuous values.

Figure 1 illustrates the graphical model in detail using standard notation alongside our inference algorithm. In brief, we use a randomized algorithm known as Monte Carlo Expectation Maximization (Wei and Tanner, 1990). The algorithm alternately a.) samples a value of the latent entity *z* for each instance of a name *x* in a given letter, conditioned on current values of all other latent variables and parameters, cycling through the name instances, and b.) uses those accumulated samples to optimize the values of the social rank β that generated the pairwise ranks *y* in evidence. In this way, we alternate between picking probable latent individuals referred to in letters (given some fixed social rank), and determining the best social rank given that estimate of who those names refer to.

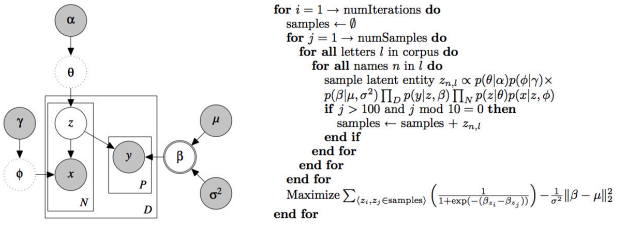


Figure 1:
Graphical model and MCEM algorithm. Shaded circles represent observed or fixed variables, empty circles are latent variables, dotted circles represent variables that are integrated out via collapsed Gibbs sampling, and the double-circled β is optimized via MCEM. φ, θ - Dir; z - Categorical; x - Multinomial; y - Logistic; and β - Normal.

In our case, we fix the hyperparameters $\alpha = 100$, $\gamma = 0.01$, $\mu = 0$, $\sigma^2 = 1$, and the number of possible latent entities to 1000.

4. Results

The input to our algorithm is a set of pairwise ranks between names mentioned in a letter (of the form Aššur-idī ≥ Aššur-nā dā , as above), along with an upper bound on the number of the latent entities we expect (*K*); the output is twofold: a.) a global rank of those *K* latent entities, along with the names in letters associated with them most often; and b.) a distribution over all possible latent entities for each name mentioned in each letter.

We apply our model to *hypothesis generation*: given a set of evidence, the algorithm offers hypotheses it finds likely, which a domain expert can then validate according to established methods in the field. One such lead generated by our method about a well-studied individual concerns the name of Innāya.

4.1 Innāya

In 1991, Cécile Michel produced a two-volume work on two merchants in the colony of Kaneš named Innāya (Michel, 1991a, 1991b) . On the basis of two attested patronyms it was apparent that there were at least two individuals who were known by this name. By charting their family trees and the structure of their respective businesses, she reconstructed the separate archives and identities of these two merchants. The first and best attested individual is Innāya son of Elāli with 142 texts, who appears to have a more complete textual record. The second individual, Innāya son of Amurāya, is only attested in 74 texts; while these two merchants overlapped chronologically, the latter appears to have been a minor figure in the colony (Michel, 1991, 48). While there were a number of texts which Michel was unable to determine (ca. 57), her study illustrated the complexity involved in the Old Assyrian archives due to an active use of homonyms at that time. While the work on Innāya is in no way complete, Michel provided a basis on which future scholarship might build, and will serve as a proving ground for the purpose of this study.

Figure 2 shows the overall distribution for latent ranks associated with the name Innāya in all of the letters, as learned by our model. Given the evidence that we have seen, our algorithm has learned that this name is generally associated with a very high-ranking individual (in this, and all other plots, the highest rank is 1, with 1000 being the lowest).

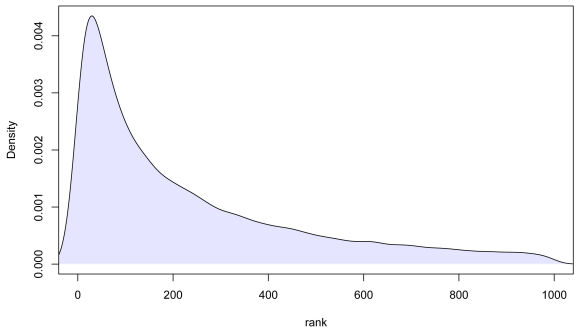


Figure 2:
Overall distribution of latent ranks for the name Innāya. The highest rank is 1; 1000 is the lowest.

If we look at the individual letters themselves, however, a more nuanced picture emerges. Figure 3 plots the distribution of latent entities for a set of 6 of the 190 letters in which Innāya is mentioned. While the majority of letters in this set recapitulate the overall distribution shown above, several letters are noticeable outliers. For example, in letter TC1,33 and BIN6,109, our model has high belief that the real person associated with the name Innāya is not in fact the high-ranking individual at all, but rather a much lower ranked one. Consulting these letters, we see that Innāya is dominated by one or more individuals with a relatively low rank, which is not consistent with a high-ranking individual. If we look at the intersection of the publicly available letters in our collection and those inspected by Michel (a total of 142 texts), we find that our method agrees with Michel's assessment of the identity of Innāya in each specific letter 80.9% of the time (discounting the level of agreement due to chance, this leads to a Cohen's κ of 0.435). These results clearly support the conclusions Michel has drawn for Innāya – two individuals, at least, each with differing social networks and hierarchical ranks – and provide evidence for the validity of our approach.

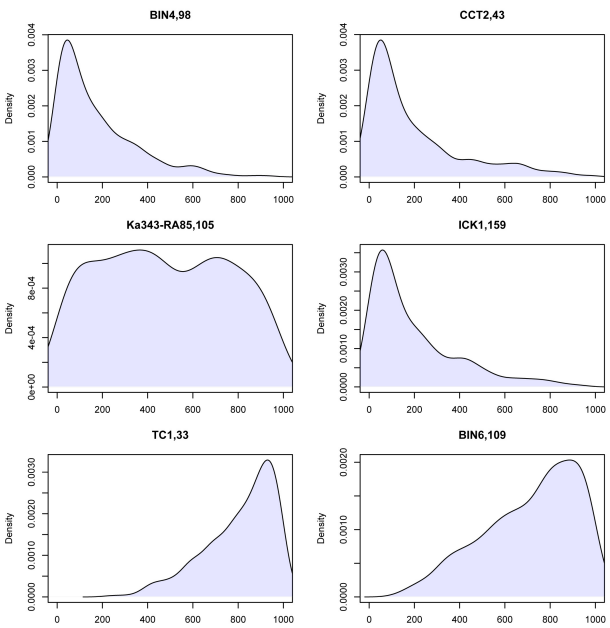


Figure 3:
Distribution over latent ranks associated with the name Innāya in various letters. The highest rank is 1; 1000 is the lowest.

5. Conclusion

The case of Innāya illustrates one of the greatest obstacles facing the field of Old Assyrian studies to date. Before any definitive statements can be made about the organization and makeup of the Old Assyrian trade network on the basis of the texts found in Kültepe, we must first determine the scope and structure of the Old Assyrian colonial society. Unfortunately, due to an active use of homonyms in the textual record, the scope of this colonial society has been obscured by a level of ambiguity too complex for any single specialist, or for that matter any group of specialists, to untangle.

As part of a solution, we present a method for aggregating small, local pieces of information—pairwise social differences between names in a cuneiform tablet—into a single underlying social order that offers the best explanation for the data we have. The latent variable model that we design allows us to be clear about our assumptions (the relationship between variables is encoded in the structure of the model; whether or not one variable is allowed to exert a direct effect on another is transparent). These kinds of generative models also allow us to add other forms of evidence; one possible extension would tie the choice of latent entity for each name in a letter with all of the other text observed (so that, for example, if one Innāya is often associated with letters mentioning *tin* while another trafficks in *textiles*, we have further evidence that the two are different individuals).

In applying this method to our Old Assyrian dataset, we are engaging in exploratory data analysis, offering informed hypotheses that are driven by data, and that our model believes are the most promising avenues for directed research by Assyriologists. Grounding these hypotheses in the data allows us to return to the source of our induction—the letters themselves—and validate whether or not we have sufficient evidence to support our claims. In our particular case, the agreement between our model's beliefs and those in the published literature are encouraging. We leave to future work to explore the differences that remain.

6. Acknowledgments

We would like to thank Mogens Larsen, Thomas Hertel, and other contributors to the Old Assyrian Text Project for providing the digitized texts we analyze. The research reported in this article was supported in part by Google (through the Worldly Knowledge Project at CMU) and by an ARCS scholarship to D.B. An extended version of this paper can be found at: http://arxiv.org/abs/1303.2873 .

References

Barjamovic, G., T. Hertel, and M. T. Larsen (2012). *Ups and Downs at Kanesh: Chronology, History and Society in the Old Assyrian Period*. PIHANS (Publications de l'institut historique-archéologique néerlandais de Stamboul), vol. 120. Nederlands Instituut voor het Nabije Oosten, Leiden.

Blei, D. M., A. Y. Ng, and M. I. Jordan (2003). Latent Dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, March.

Michel, C. (1991). *Innāya dans les Tablettes paléo-assyriennes I: Analyse*. Editions Recherche sur les Civilisations, Paris, France.

Michel, C. (1991). *Innāya dans les Tablettes paléo-assyriennes II: Edition des textes*. Editions Recherche sur les Civilisations, Paris, France.

Michel, C. (2008).The Alāhum and Aššur-taklāku Archives found in 1993 at Kültepe Kaniš. *Altorientalische Forschungen*, 35:53–67, May 2008.

Old Assyrian Text Project. http://oatp.ku.dk .

Veenhof, K. R., and J. Eidem (2008).*Mesopotamia: The Old Assyrian Period*. Orbis Biblicus Et Orientalis 160/5. Academic Press (Fribourg), Vandenhoeck and Ruprecht (Göttingen).

Wei, G. C. G., and M. A. Tanner (1990). A Monte Carlo implementation of the EM algorithm and the poor

man's data augmentation algorithms. *Journal of the American Statistical Association*.

Notes

1. Most of the letters in our corpus have not been dated to a finer level of granularity than century; a potential extension to the model described here would exploit this information when available.
2. Alternatively, JEFFERSON1 > ADAMS > MADISON > JEFFERSON2 is also valid, as is MADISON1 > JEFFERSON > ADAMS > MADISON2. In data where the orderings are not strict (i.e., ADAMS ≥ JEFFERSON), global ranks involving equalities are also possible.

The Sounds of the Psalter: Computational Analysis of Phonological Parallelism in Biblical Hebrew Poetry

Benner, Drayton Callen
drayton@uchicago.edu
University of Chicago, United States of America

1. Introduction

Parallelism lies at the artistic heart of Biblical Hebrew poetry. Traditionally, the focus of research on parallelism in Biblical Hebrew poetry was largely limited to semantic parallelism, but in recent decades scholars have responded to Roman Jakobson's challenge to explore the grammatical and phonological aspects of parallelism as well (1966). These recent treatments of phonological parallelism have represented an important addition to the study of Biblical Hebrew poetry (Watson 1984, 1994; Pardee 1988; Alonso-Schökel 1963, 1988; McCreesh 1991). However, the number of phonemes in Biblical Hebrew is naturally limited, so repetition of phonemes and phonemes with similar features is inevitable. In the rare instances in which scholars have offered objective criteria by which to discriminate between intentional and unintentional phonological parallelism in ancient Semitic poetry, the criteria have been arbitrary, divorced from any nuance required to make them statistically meaningful (e.g. Margalit 1979). Criteria proposed by scholars for other languages have been better but still not designed to responsibly find clusters of any phonemes over an arbitrarily large swath of poetry (Clement 2012; Plamondon 2001, 2005, 2006; Barquist 1987, 1991; Hidley 1986; Leavitt 1976; Jackson 1942; Skinner 1941; Chisholm 1976, 1981; Magnuson 1962; Hervás 2007).

There are three main ways in which ancient Israelite poets used phonological parallelism. First, they used a cluster of a single phoneme or a group of phonemes with similar features in a small section of a poem for artistic effect, making lines memorable and binding together poetic units tightly. Second, they clustered the sounds from a key word in the poem, reinforcing the theme. Third, they used phonological parallelism in creative ways across large sections of a poem to reinforce the structure of the poem and/or its function in Israelite society. In this paper, I explore computational techniques for studying the first two of these categories in a single, sizable corpus of Biblical Hebrew poetry, namely the Book of Psalms.

This paper limits its investigation to the consonants on account of the difficulties in reconstructing the precise vowels of the biblical period, particularly in the Book of Psalms, an anthological corpus with compositions from many different centuries. For the consonants, I assume that the consonantal orthography represents the phonology well, with three exceptions: the representation of /ʕ/ and /ɣ/ by ע (ʿ), the representation of /ħ/ and /x/ by ח (ḥ), and the occasional quiescence of א (ʔ). I also assume that the Massoretes correctly distinguished between two phonemes in marking שׁ as שׁ (ʃ) and שׂ (ɬ).

2. Source of Data

The J. Alan Groves Center for Advanced Biblical Research maintains an accurate digital representation of the text of the Hebrew Bible, following the Leningrad Codex, and has also tagged each word with lexical and morphological information. Version 4.14 is used herein after having stripped the vowels, cantillation marks, and matres lectionis—vowels represented in the orthography using consonant symbols—from the text.

3. Visualization

Two types of visualizations have been employed. Both rely on a mapping from the features of each phoneme onto a three dimensional RGB color space so that similar sounds are mapped to similar colors. In the first visualization, each consonant of the text of the Hebrew Bible is colored. In the second visualization, each consonantal phoneme is represented by a short vertical line, with one row of these vertical lines per chapter. The user can view the vertical

lines for all consonantal phonemes or any subset of them, with the rest forming a black background.

Figure 1.
Visualization for Psalm 37, showing just the ר (ʁ) phoneme. The white block indicates the end of the psalm. Note the cluster of lines near the right side (verses 34-36), in which eighteen consecutive words contain a ר (ʁ).

4. Statistical and Computational Techniques

With a limited consonantal phonemic inventory A and the hypothesis that a particular poem or section thereof contains a cluster of a set of phonemes P, the simplest way of testing that hypothesis is to calculate the probability that a poetic section that size would contain at least the observed number of occurrences of phonemes in P if the phonemes were chosen randomly in accordance with their respective frequencies in the corpus as a whole. Thus, let n be the number of consonantal phonemes in the poetic section, let x be the observed number of phonemes in P in the poetic section, and let p be the probability of any given consonantal phoneme in the entire corpus being an element of P. Then, based on the cumulative distribution function of the binomial distribution:

$$\Pr(X \geq x) = \sum_{i=x}^{n} \binom{n}{i} p^i (1-p)^{n-i}$$

Where computationally necessary, the normal approximation of the binomial distribution or the Poisson approximation of the binomial distribution can be used to estimate this probability.

Greater nuance can be added to this statistical test by weighting certain words in the text more heavily than others, taking into account all of the following:

a) The relative frequency of lexemes. A list of nouns with the definite article ה (h) does not represent a cluster of the ה (h) phoneme. By contrast, a rare lexeme may have been chosen over a more common one precisely on account of its sound.
b. The repetition of lexemes. The repetition of a particular lexeme does not signal that the poet was specifically seeking to employ phonological parallelism over against another type of parallelism. A clause such as פַּחַד וָפַחַת וָפָח עָלֶיךָ (pḥd wpḥt wpḥ ʕlk; "terror, pit, and snare are upon you"; Isaiah 24:17), is rhetorically effective on account of its use of different lexemes with similar phonemes.
c. Parts of speech. Phonological parallelism using content words is more effective than that which uses function words.

These factors are used to weight each word so that each consonantal phoneme in that word counts not as a single Bernouilli trial but rather as a trial with weight in the range (0.0, 1.0]. Since this no longer results in a discrete binomial distribution, the distribution can be approximated via the continuous Gamma distribution, the cumulative distribution function of which can be used to find the probability of a given poetic section having at least the observed number of phonemes in P.

When a scholar proposes the artistic use of sound in a text, one can evaluate this claim using this technique. In addition, this technique can be used to identify poems or sections thereof that deserve the researcher's consideration. I have written software that, given a set of consonantal phonemes and constraints on the size of text, computes this metric for every stretch of text in the Hebrew Bible and lists all of them that score above a given value, with care taken so that only the highest scoring passage is shown among qualifying passages that overlap. In other words, this software finds the passages within a given size range that are most likely to contain an artistic clustering of the desired sounds.

5. Results

The Hebrew Bible, and even the Book of Psalms, is large enough that given a set of phonemes, our software is bound to find many false positives alongside any legitimately artistic uses of sound. Indeed, estimates concerning the number of expected false positives can be discerned by two methods. First, one can compare the number of prose texts found with a given minimum score to the number of poetic texts found. In addition, one can rearrange at random the words of the Hebrew Bible and of the psalms in particular and then re-run the algorithm to see how many passages score above a given level. Doing these tests indicates that we should expect the majority—but not all—of the high-scoring passages to be false positives. Thus, a critical eye and ear are required to determine when phonological parallelism is not merely statistically plausible but also rhetorically plausible, to the point that it is likely that the poet used sound intentionally.

Looking through the output of the software tool has produced some compelling examples, of which three are presented briefly in conclusion. In Psalm 37:34-36, the ר (ʁ) phoneme appears precisely one time in each of eighteen consecutive words. Psalm 37 is an acrostic poem, and these eighteen words span the ק (kʾ) and ר (ʁ) sections. This phonological parallelism helps to reinforce the acrostic structure of the poem and also serves to bind the ק (kʾ) and ר (ʁ) sections together. Psalm 46 is generally classified by modern interpreters as a song of Zion, one of a series of psalms with Jerusalem/Zion as its central theme. The psalmist only actually mentions the city in verses 5-6 and does not even name it. Yet in verses 3-4, there is a cluster of the sounds of the word יְרוּשָׁלַם (jʁʃlm; "Jerusalem"), climaxing with the two words יִרְעֲשׁוּ הָרִים (jʁʕʃ hʁm; "mountains quake"), which together sound very much like יְרוּשָׁלַם (jʁʃlm; "Jerusalem"). Similarly, Psalm 122 is all about Jerusalem and its temple, and it makes heavy use of the phonemes in יְרוּשָׁלַם (yʁʃlm; "Jerusalem") throughout the psalm, with an especially high density in verse 6.

References

Alonso Schökel, L. (1963). *Estudios de Poética Hebrea*. Barcelona: J. Flors.

Alonso Schökel, L. (1988). *A Manual of Hebrew Poetics*. Translated by Adrian Graffy. Subsidia Biblica 11. Rome: Biblical Institute Press.

Barquist, C. R. (1987) Phonological Patterning in *Beowulf*. *Literary and Linguistic Computing* 2: 19-23.

Barquist, C. R., and D. L. Shie (1991). Computer Analysis of Alliteration in Beowulf Using Distinctive Feature Theory. *Literary and Linguistic Computing* 6. 274-280.

Berlin, A. (1985). *The Dynamics of Biblical Parallelism*. Bloomington, Ind.: Indiana University Press.

Chisholm, D. (1976). Phonological Patterning in German Verse. *Computers and the Humanities*. 10. 5-20.

Chisholm, D. (1981). Phonology and Style: A Computer-Assisted Approach to German Verse. *Computers and the Humanities*. 15. 199-210.

Clement, T. (2012). "Methodologies in the Digital Humanities for Analyzing Aural Patterns in Texts." Proceedings of the 2012 iConference. New York: ACM, 2012. 287–293.

Hervás, R., J. Robinson, and P. Gervás. (2007). Evolutionary Assistance in Alliteration and Allelic Drivel. *Lecture Notes on Computer Science*. 4448. 537-546.

Hidley, G. R. (1986). Some Thoughts Concerning the Application of Software Tools in Support of Old English Poetic Studies. *Literary and Linguistic Computing*. 1. 156-162.

Jackson, E. (1942). The Quantitative Measurement of Assonance and Alliteration in Swinburne. *American Journal of Psychology*. 55. 115-123.

Jakobson, R. (1966). Grammatical Parallelism and Its Russian Facet. *Language* 42. 399-429.

Leavitt, J. A. (1976). On the Measurement of Alliteration in Poetry" *Computers and the Humanities*. 10. 333-342.

Magnuson, K. (1962). Consonant Repetition in the Lyric of Georg Trakl. *Germanic Review*. 37. 263-281.

Margalit, B. (1979). Introduction to Ugaritic Prosody. *Ugarit Forschungen*.11. 289-313.

McCreesh, T. P. (1991). *Biblical Sound and Sense: Poetic Sound Patterns in Proverbs 10-29*. Sheffield, England: Sheffield Academic Press.

Pardee, D. (1988). *Ugaritic and Hebrew Poetic Parallelism: A Trial Cut 'nt I and Proverbs 2*. Vetus Testamentum Supplements 39. Leiden: E.J. Brill.

Plamondon, M. R. (2001). *The Musical Aesthetics of the Poetry of Tennyson and Browning*. Ph.D. dissertation. University of Toronto.

Plamondon, M. R. (2005). Computer-Assisted Phonetic Analysis of English Poetry: A Preliminary Case Study of Browning and Tennyson. *TEXT Technology*. 14. 153-175.

Plamondon, M. R. (2006). Virtual Verse Analysis: Analysing Patterns in Poetry. *Literary and Linguistic Computing*. 21. Supplemental Issue: 127-141.

Skinner, B. F. (1941). A Quantitative Estimate of Certain Types of Sound-Patterning in Poetry. *American Journal of Psychology*. 54. 64-79.

Watson, W. G. E. (1984). Classical Hebrew Poetry: A Guide to its Techniques. *Journal for the Study of the Old Testament*. Supplement Series 26. Sheffield: JSOT Press.

Watson, W. G. E. (1994). Traditional Techniques in Classical Hebrew Verse. *Journal for the Study of the Old Testament*. Supplement Series 170. Sheffield: Sheffield Academic Press.

The German Language of the Year 1933. Building a Diachronic Text Corpus for Historical German Language Studies.

Biber, Hanno
Hanno.Biber@oeaw.ac.at

Austrian Academy of Sciences, Austria

Breiteneder, Evelyn

Evelyn.Breiteneder@oeaw.ac.at
Austrian Academy of Sciences, Austria

In the following paper a new research project undertaken within the framework of the AAC–Austrian Academy Corpus operated by the Institute for Corpus Linguistics and Text Technology at the Austrian Academy of Sciences in Vienna will be presented. The topic of this research project is focused on the questions of developing a diachronic text corpus of historical significance and establishing a corpus based research environment for language studies of the interwar period with particular emphasis on the year 1933, the year when the Nazis came to power in Germany. In the presentation of this project an overview of the necessary methodological considerations and an outline of the research perspectives based upon the principles of corpus linguistics will be given. Corpus-based approaches for analyzing the language of the historical periods before, during and after National-Socialism have been rare, despite the numerous works in the fields of historical studies as well as in German language studies. A research group of the corpus research framework of the AAC has decided to start a special investigation into this challenging question.

The AAC is an established German language text corpus of more than 500 million tokens and represents a considerably large diachronic digital text corpus comprising several thousands of German language texts of important historical and cultural significance. The core of the AAC is from the first half of the twentieth century, so that the research issue of an analysis of the German language of 1933 is not only highly appropriate, but can be addressed on a comprehensive basis. Among the sources of the AAC a large number of texts of the historical period in question have already been collected, digitized, converted into machine-readable text and fully annotated as well as been provided with metadata. Structural and thematic mark-up has been applied according to annotation and mark-up schemes based upon XML related standards. The AAC has been committed to the research field of text technology since its official foundation in 2001.

Building a diachronic digital text corpus for historical German language studies of this particular kind is a particularly challenging task for various reasons. First, the technical difficulties of corpus building in dealing with a large historical variety of different text types and genres have to be taken into consideration. Second, the specific historical parameters and the methodological scope of such an investigation has to be taken into account. The German language of the year 1933 is being considered as a historical focal point for which an exemplary corpus-based research methodology for the study of the German language could be developed. The sources of a first exemplary study will cover manifold domains and genres, not only newspapers and political journals and magazines, which will be at the core, but also several other text types representing the historical communicative strategies will be included. Among them are pamphlets, flyers, advertisements, radio programs, political speeches, but also essays and literary texts as well as administrative, scientific or legal texts, just to name a few examples, which are all difficult to collect. The AAC has started to build up a small collection of ephemera in this field.

In the overall project a special emphasis will be given to the "Dritte Walpurgisnacht" (Third Walpurgis Night), written by the satirist Karl Kraus, which will be taken (in digital format), among other sources, as a starting point for text selection for the corpus. This text is the most important contemporary text of German literature dealing with National Socialism. In the "Third Walpurgis Night" Karl Kraus has documented the murderous reality of the Nazi regime as early as May 1933 and documented and commented upon the murderous language of that time in numerous examples. Because of this text no one can claim not to have been able to know from the very start where Nazi rule would lead. However, Karl Kraus, the editor and author of his journal "Die Fackel" (The Torch), who died in 1936, did not publish his text, a text which begins with the famous line "Mir fällt zu Hitler nichts ein", because in the face of violence the deed of the word was considered inappropriate by him.

The historical period covered by the AAC is ranging from the 1848 revolution to the fall of the iron curtain in 1989. In this period significant historical changes with remarkable influences on the language and the language use in the German speaking areas can be observed. The year 1933 and the years preceding as well as following the "Machtergreifung" (seizure of power) of the National socialists is a historical period of particular interest for language studies. In this case not primarily the well-known documents and the evident language of the Nazis will be included in the analysis, but systematically the less easily visible documents and less significant lexical items will be taken into consideration as well. This methodological approach is considered as particularly fruitful by means of applying methods of corpus linguistics and by testing new strategies of the application of these methods in the context of historical language studies. For this historical period the AAC corpus holdings provide a great number of reliable resources and interesting corpus based approaches for investigations into the linguistic and textual properties of the texts in question. The digital text is going to be enriched by additional data and will be lemmatized and provided with POS data thereby making use of the tag set for this purpose of the STTS (Stuttgart-Tübingen Tagset). "Quantitative

corpus linguistics has proofed to be a valuable technique in many domains of philological, sociological and historical research. The digitized and linguistically annotated corpus is therefore an interesting source for studies in many fields and facilitates the investigation of changing patterns of language use, and how these reflect underlying cultural shifts." (M. Volk). The question is, whether corpus research methods based upon a multidisciplinary combination of corpus linguistics, lexicography, historical studies and cultural studies can be applied in order to gain insights into the textual representations of historical collections of this importance.

The AAC research group will go beyond a quantitative approach and integrate text studies into its research of the German language of 1933. The methodology of corpus based text research is determined by corpus linguistic, lexicographic and analytical procedures. The historical condition of Germany and Austria with their cultural and linguistic diversities and in particular the situation at the time of National Socialism have to be taken into consideration as historical changes with significant influences on the language. In contrast to other corpus-oriented projects, the working group proceeds from literary studies and text lexicographic premises. Corpus research and the creation of large electronic text collections have traditionally been the domain of corpus linguists. Literary digitization initiatives were quite often restricted to particular writers and many of these projects did neither produce large amounts of data nor pursue research on methods of how to tackle the problems involved in working with such data. Our perspectives parallel those formulated in the European project CLARIN which has been set up to "create, coordinate and make lan-guage resources and technology available and readily usable" (Call for Proposal), in this case also for text historians and for those interested in ideologically determined language change.

The AAC has already developed methods and tools to allow scholars to access these texts and other comparable resources. For this purpose the tools provided in order to access the corpus holdings will enable the researcher to input queries and to get a display of the results in forms provided with the necessary metadata such as information on sources, authors, date of publication etc., and with a display of the related pages of the results as digital text, also allowing access to the XML source of the texts and to define custom style sheets, alongside a feature to view facsimiles of the texts offering simultaneous access to text and facsimiles of pages. In addition to that a sophisticated navigational control tool will be provided, offering random access to the variety of different documents, journals, books, etc. to do linguistic, literary or historical research. Access to the text corpus will be given not only through query result lists but also through a structuring tool, which allows readers to navigate to any desired part of the corpus and results of queries are delivered by the server in XML format which makes it fairly easy to adjust the representation of the output, where XSLT style sheets can be used. Using such style sheet transformations also allows creating statistical analyses of the data. And it has been pointed out before (Smith, 2008) that available standard tools provide only limited support when processing query results. Building a diachronic digital text corpus for historical German language studies of this particular kind is a particularly challenging task which demands also the development of new tools and new approaches of text technology. This special research environment would be especially useful for corpus-based analyses of the language of critical historical periods such as the case of the German language in the year 1933.

References

AAC — Austrian Academy Corpus: AAC-FACKEL. Online Version: «Die Fackel. Herausgeber: Karl Kraus, Wien 1899-1936». In Biber, H., E. Breiteneder, H. Kabas, Mörth, K. *AAC Digital Edition No 1*. http://www.aac.ac.at/fackel

Biber, H. Aufbruch der Phrase zur Tat. Kommunikationsmaßnahmen und sprachliche Formungen der nationalsozialistischen Machtübernahme in Österreich 1938. In: Welzig, W., H. Biber, und C. Resch. *Anschluss. März/April 1938 in Österreich.* Wien, Verlag der Österreichischen Akademie der Wissenschaften 2010, S. 15-37

Bubenhofer, N. et al. (2007). XML-Technologien als Grundlage dynamischer Textpräsentation. Die digitale Quellenedition Der Zürcher Sommer 1968. *Jahrbuch für Computerphilologie* 9 89-110.

Mörth, K. (2000). The representation of literary texts by means of XML: some experiences of doing markup in historical magazines. In Fraser, M., Williamson, N., and Deegan M. (eds), Digital Evidence. 2002. *Selected papers from DRH 2000, Digital Resources for the Humanities Conference.* Office for Humanities Communication 14, 17-32.

Roth, T. (2009). Verteilte Korpusabfragesysteme. *Proceedings in Language and Text Corpus. Design and Linguistic. Corpus Analysis*.

Smith, N. et al. (2008). Corpus Tools and Methods. Today and Tomorrow: Incorporating Linguists' Manual Annotations. *LLC.* 23 163-180.

Volk, M. et al. (2010). Challenges in building a multilingual alpine heritage corpus, In: LREC Proceedings.

Documentary Social Networks: Collective Biographies of Women

Booth, Alison
ab6j@virginia.edu
University of Virginia, United States of America

Martin, Worthy
wnm@virginia.edu
University of Virginia, United States of America

Collective Biographies of Women (CBW) is a collaborative[1] , open-access[2] literary and prosopographical project focusing on published collections of biographies of women [3] . The literary focus of this interdisciplinary project (which also centers in women's history in Britain and the U.S.) concentrates on a popular genre that has received little critical attention. We study the narrative structure of short biographies of diverse women's lives delineated through interpretive analysis captured in a stand-aside XML mark-up using the BESS schema[4] . Having presented the theoretical and literary aspects in multiple venues[5] , in this short paper we will present emerging prosopographical interpretations of a database of over 1200 volumes, comprising more than 13,000 biographies of more than 8000 women. The women in this corpus come from all walks of life (not simply one occupation or nation, such as the women writers considered by Orlando or Brown's Women Writers Project[6]), and range from ancient and biblical figures to living contemporaries of the biographers.

In our proposed paper we introduce the phrase, "documentary social network," as a label for our prosopographical interpretations because we are interested in social networks as discovered in biographical documents[7] . Yet these networks do not always resemble those that can be found in an investigation of Twitter feeds or Facebook "friends" or even of archival materials such as provided by SNAC[8] . The social networks evidenced through our collection of collections are those of documentary grouping and reference. In specialized collections, Christian missionaries in Africa or nurses in World War I may have interacted in historical time and place, but assorted tables of contents commonly link some individuals who never actually shared a "live" event or a relationship or communication. Thus, the relationships are as perceived and presented through printed volumes collecting short versions of biographies.

Let us consider one Mrs. John Livingstone. She appears in only three CBW collections and by way of those collections her immediate documentary social network has 37 other women[9] . To visualize this network, Figure 1 is centered on Mrs. Livingstone with the three collections positioned on the innermost circle and the other women of those collections around the second circle. Of her 37 collection "siblings," nine also appear in these same three collections and nowhere else in CBW, confirming some consensus on the names that represent "Notable Women of the Scottish Reformation," a specific historical episode in one location. Other members of Livingstone's documentary "siblings" range into other collections and in Figure 2 those additional collections are positioned on the third circle. In more eclectic lists, some of the Scottish heroines of religious conflict intersect with a multi-national set of women widely recognized today, including writers such as Harriet Beecher Stowe and heroines of war such as Joan of Arc or the Countess of Montfort. Mrs. John Livingstone's 267 documentary "cousins" are displayed on the fourth circle in Figure 3[10] .

CBW's first experiments in digital analysis of narrative structure using the BESS schema have focused on two sample archives that are also productive for this paper's claims about documentary social networks. (Each sample archive is a set of all female collective biographies that include a certain woman, as in Fig. 1.) The two sample archives are "Noble Workers," 20 books that include a biography of Sister Dora, who ran hospitals in the industrial Midlands, and "Women of the World," 14 books that include a biography of Lola Montez, a celebrity in Europe, New York, California, and Australia. These Victorian women never met and never appear in the same book. The mediated interconnections between Sister Dora and Lola Montez appear in Figure 4. Only two women, Marie Antoinette (presented as a victimized queen) and Jenny Lind (the celebrated, virtuous singer), appear with Sister Dora in one collection *and* with Lola Montez in another (i.e., within one degree of separation of both Dora and Lola). The multiple versions of the lives of 141 women who "network" with Sister Dora in Noble Workers collections, compared to the versions of the 133 persons linked to Lola Montez in Women of the World books, will demonstrate the utility of our approach to analyze historical social networks and nonfiction narratives of many kinds. In spite of the vast differences between two Victorian women — one a saintly nurse, the other a notorious courtesan and performer — we discover patterns among their associates and their proximity in the CBW documentary social network.

These prosopographical networks reveal unexpected affiliations among different types of women, life stories,

and collections; with the CBW tools for searching and visualization, we are now able to calibrate not only frequencies and proximities of persons and publications over time, but gradations of rhetorical assessment of women's roles and deeds, according to the perspectives of these publications[11] . Such multivalent interconnections not only cross categories but also historical periods in significant ways. An example is the connection between Sister Dora and Anne Boleyn, shown in Figure 5. This diagram reveals that the middle-class, now-forgotten Englishwoman, Sister Dora, appears with her contemporary queen, Victoria (in seven books), and in two collections with the martyred queen, Jane Grey (one of the latter, a186 [12] , is a book that Victoria also shares), but not with Henry VIII's second wife[13] . At sufficient scale, and integrated with other elements in our analysis, data on frequency or "distance" in networks can lead to productive interpretations. Anne Boleyn, unlike good Lady Jane in ambition and sexual notoriety , but like her in being executed because of the politics of English monarchy, features in 24 collections, including 11 dedicated to English queens and three focused on the English Reformation; the latter three collections include Lady Jane Grey but not one of the obscure Scotswomen of Mrs. Livingstone's ilk. To put it simply, Boleyn serves English and Protestant history, but not the promotion of feminine virtue, heroism, or social service.

More complex documentary interrelationships can be seen in Figure 6, associating the Gothic novelist Ann Radcliffe (dead before Sister Dora was born) with Sister Dora through a range of widely respected eighteenth-and nineteenth-century women writers; these lists also commend Jeanne d'Albret, the French queen regnant who championed the Calvinist Huguenots, as well as Lady Russell, Mme. Roland, and Lady Jane Grey, represented as good, highly educated women who played indelible parts in the history of their countries in revolutionary times.

Our proposed paper will present the methods and implications of prosopographical analyses afforded by the CBW documentary social network. Our demonstration of documentary social networks has implications for any "personography," as well as for historical studies of women and studies of biographical narrative. We argue, in this paper, that digital exploration of the persons, narratives, collections, and documentary social networks in the CBW genre opens a prolific field of ramifying versions of female biography impossible to decipher through an approach to individuals or single full-length biography, and invisible through the customary lenses of historiography, time and period, place and nationality.

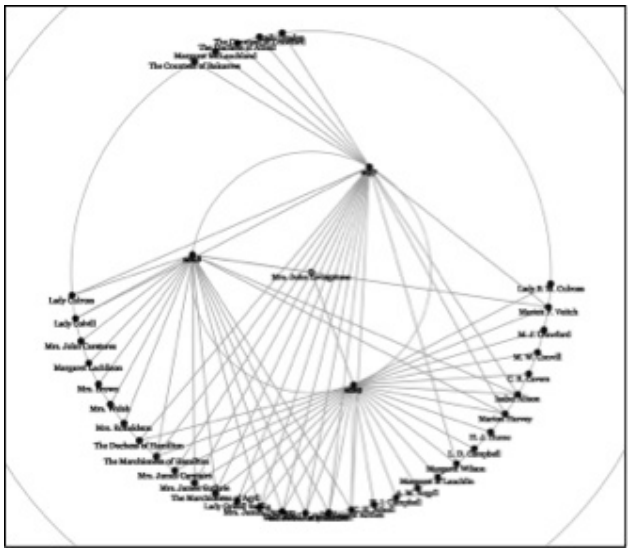


***Figure 1*:**
Mrs. John Livingstone and her 37 documentary siblings.

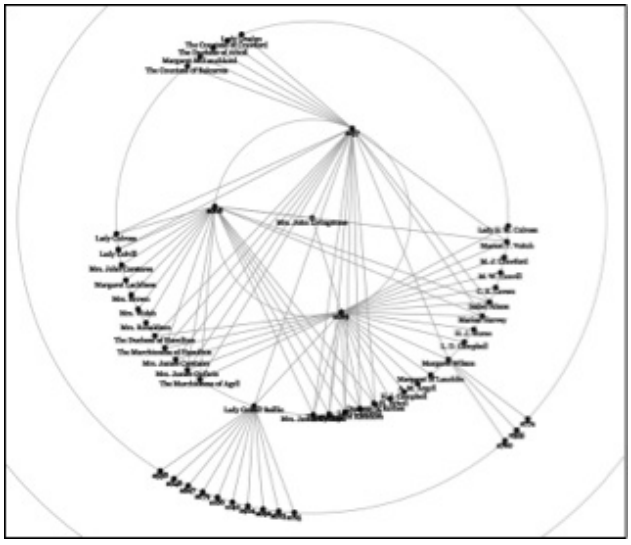


***Figure 2*:**
The 16 collections containing Mrs. John Livingstone and her documentary siblings.

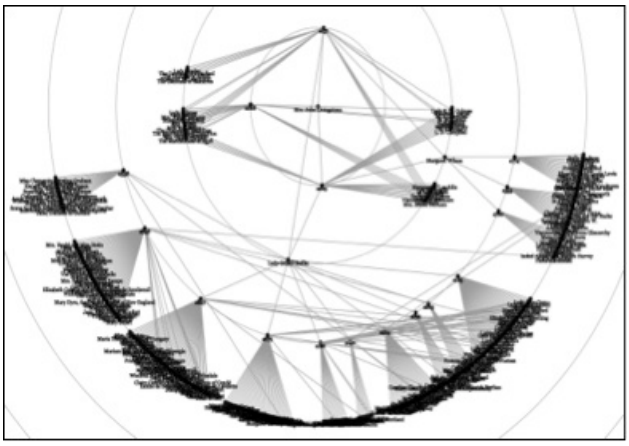


***Figure 3*:**
Mrs. John Livingstone and her 267 documentary cousins.

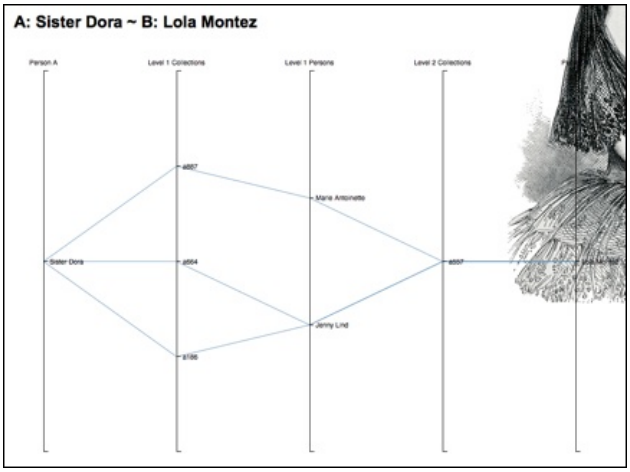


***Figure 4*:**
Connections between Sister Dora and Lola Montez.

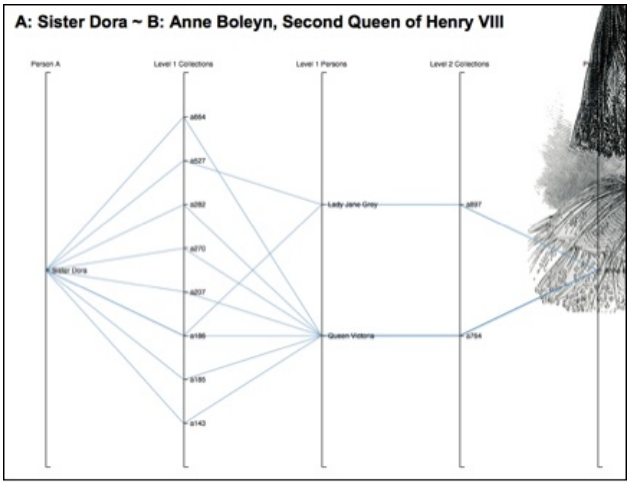


***Figure 5*:**
Connections between Sister Dora and Anne Boleyn.

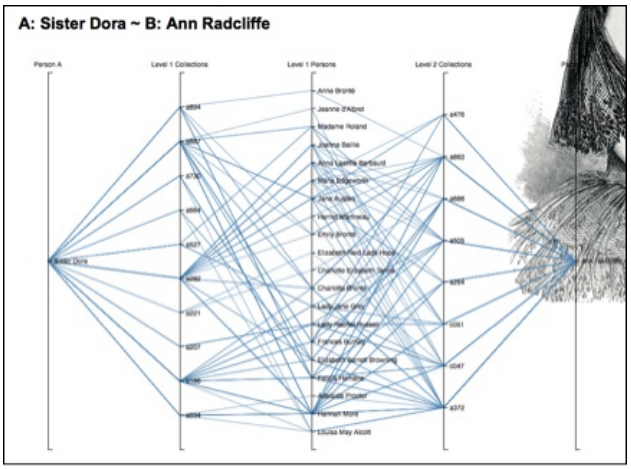


***Figure 6*:**
Connections between Sister Dora and Ann Radcliffe.

Notes

1. The collaboration is among the UVa English Department (Booth and numerous graduate students) and the Institute for Advanced Technology in the Humanities (Martin, Pitti, Ross, Girard, Brandon, and Bingler), building on earlier work with the UVa Library's Scholars' Lab (Gilbert and others).
2. To access the complete bibliography and "featured subjects" pages go to http://womensbios.lib.virginia.edu/ and to access a prototype interface to the database go to http://cbw.iath.virginia.edu/cbw_db/ .
3. The volumes are English-language collections published primarily between 1830 and 1940 (though publications of earlier and later dates are included), each containing biographies of from three to 150 women.
4. Biographical Elements and Structure Schema (BESS) is designed to reveal narrative structure and other elements in the short biographies, for measurable comparison of versions of one woman's life over many collections, and for analysis of the forms within a single collection or category of collections.
5. These include invited keynote lectures, e.g. Feminist Narrative Theories (a Project Narrative symposium, Ohio State), Life-Writing (a conference at Huntington Library), and British Women Writers Conference, as well as papers at North American Victorian Studies Association, Narrative, and Modern Language Association conferences (2009-2012).
6. See Susan Brown, Clements, Grundy, et al. *Orlando: Women's Writing in the British Isles from the Beginnings to the Present.* Cambridge: Cambridge University Press, 2006. http://orlando.cambridge.org/ , University of Alberta www.arts.ualberta.ca/~orlando; and Julia Flanders, Encoding Names for Contextual Exploration in Digital

Thematic Research Collections. NEH ODH Level II Final Report, 30 April 2010. http://www.wwp.brown.edu.
7. For innovative digital work with literary biographical networks, see Alan Liu, Research Oriented Social Environment (RoSE) http://rose.english.ucsb.edu/ . Unlike CBW, most digital prosopography projects *reconstruct* data about groups of lives in eras that predate printed documentation. See John Bradley and Harold Short, "Using Formal Structures to Create Complex Relationships: The Prosopography of the Byzantine Empire—A Case Study," in K. S. B. Keats-Rohan (ed.), *Resourcing Sources Prosopographica et Genealogica*, vol. 7., Oxford, 2002. Susan Brown, Alan Liu, John Bradley and others have committed to participate in a proposed NEH Level-I Startup Grant workshop to be held at University of Virginia in 2013.
8. The Social Networks and Archival Context Project, see: http://socialarchive.iath.virginia.edu/
9. Janet Fleming Livingstone or Mrs. John Livingstone surfaces in a cohort of late-seventeenth-century Scottish Protestant Dissenters; many are wives of Presbyterian ministers or widows of martyrs or exiles. The relatively coherent group includes noblewomen, but most are historically obscure (life dates or first names unknown). Two collections have versions of the same list of persons: a029, Anderson, Rev. James [of Edinburgh], *The Ladies of the Covenant: Memoirs of Distinguished Scottish Female Characters, Embracing the Period of the Covenant and the Persecution* (London: Blackie, 1850), with reprints through 1880; a157, Chapman, William, *Notable Women of the Covenant: Their Lives and Times* (London: Swan Sonnenschein, 1883). The third book including Mrs. Livingstone is a068, Beaton, Rev. Donald, *Scottish Heroines of the Faith: Being Brief Sketches of Noble Women of the Reformation and Covenant Times* (London and Glasgow: Catt; Adshead, 1909).
10. One near relation (two degrees of separation) of Livingstone is Sister Dora, who illustrates the exponential possibilities: she appears in 20 collections with 141 "siblings" and then a prodigious 3560 "cousins."
11. One near relation (two degrees of separation) of Livingstone is Sister Dora, who illustrates the exponential possibilities: she appears in 20 collections with 141 "siblings" and then a prodigious 3560 "cousins."
12. This is the identifying "key" for this collection in the CBW bibliography for the genre.
13. Queen Victoria, who boasts 60 biographical chapters in the CBW books, shares 10 books with Lady Jane Grey, but only five with Anne Boleyn. Lady Jane's 48 chapters include 12 in books that also represent Anne Boleyn.

Beyond the Document: Transcribing the Text of the Document and the Variant States of the Text

Bordalejo, Barbara
b.bordalejo@bham.ac.uk
University of Saskatchewan

The transcription of primary textual sources is at the center of digital editing projects. The capacity offered by the digital medium allows for the inclusion of much more detailed transcriptions than is possible within the frame of printed editions. But this capacity tempted scholars into attempts to capture every possible aspect of the document in their transcription. Unless we are very careful about the choices that we make in relation to the transcription of a document, we risk making an enormous effort and spending a great deal of money on transcriptions that end up not being very useful. This paper describes our use of an encoding system that allows scholars to present a transcription of the text of the document alongside the variant states of the text, making it possible to go beyond the encoding of documents.

Some important projects in digital humanities are focused on the transcription of documents. Notable examples of this are Transcribe Bentham (http://www.ucl.ac.uk/transcribe-bentham/), the Jane Austen's Fiction Manuscripts Digital Edition (http://www.janeausten.ac.uk) or Nietzsche Source (http://www.nietzschesource.org/). Moreover, editors such as Hans Walter Gabler who have made the transition from print to digital and who advocate critical editions are shifting their focus towards the transcription of primary sources. For example, Gabler, in his article "The Primacy of the Document in Editing," asserts that:

> …the text should be seen fundamentally as a function of the document. For, after all,… it is documents that we have, and documents only. In all transmission and all editing, text are (and, if properly recognized, always have been) constructs from documents. (Gabler 199)

This re-definition of text as construct from a document develops in Gabler's argument to become "…a set of document functions comprehensively deriving from the continuous manuscript posited…" (207). This represents a shift from what has been considered the traditional role of the editor as the scholar that establishes the text. Indeed, Peter Robinson points out that:

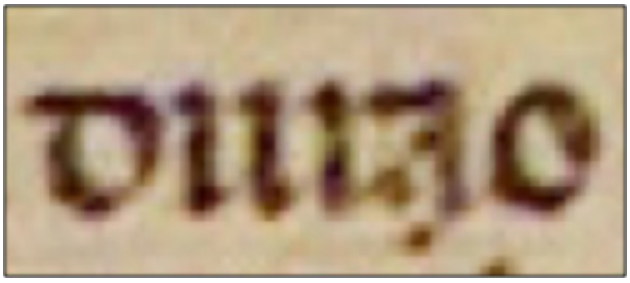
> [Gabler] proposes a complete refocusing [sic] of editorial perspective: away from a concentration on the finished product, the editorial text which is supported by reference to various documents, towards a concentration on the documents themselves, from which an editorial text may (or may not) emerge. This is an immense shift. Gabler proposes that the intense editorial effort which for centuries has seen as its goal the construction of an editorial text, should now focus on the construction of the text of the documents. (Robinson 5)

Robinson's interpretation of Gabler's text is informed by my own concept of the text of the document, as developed for the article "The *Commedia* Project Encoding System." While working on the electronic edition of the *Divine Comedy*, I developed a complex encoding system to represent, within a single file, the text of the document and the variant states of the text. In this article, I explain that:

> I use the phrase the 'text of the document' to refer to the sequence of marks present in the document, independently of whether these represent a complete, meaningful text. That is: the reader sees a sequence of letters, occurring in various places in relation to each other (perhaps between the lines or within the margins) and carrying various markings (perhaps underdottings or strikethroughs). These make up what I here refer to as the text of the document. (Bordalejo 2010).

I am not the first in referring to text of the document, G. Thomas Tanselle has contrasted the text of the document (1978; 1996) with "the text of the work" (Tanselle, 1996).

The text of the document, then, can be recorded as a sequence of meaningful marks present on a particular document. MS. Riccardianna 1005, a witness of Dante's *Divine Comedy,* presents the following example:



This could ordinarily be represented by: `<del>dura</del><add>duro</add>`

In our transcriptions, however, we represent this as:

```
<app>
<rdg type="orig">dura</rdg>
<rdg type="c1">duro</rdg> <rdg type="lit">
dur<hi rend="ud">a</hi>o</rdg> </app>
```

Here, the last reading, `<rdg type="lit">`, represents the text of the document literally. Unlike traditional TEI recommended encoding we do not use `<add>` or `<del>`, on the premise that both of these confound a statement of editorial interpretation (this is a deletion or an addition, so we should read the text with or without these characters) with a statement of what appears on the document (there is a mark under these characters; they are placed interlineally). In our view, the statement of what we see on the page is a different order of declaration from a statement of how we think this should be read. The aim of `<rdg type="lit">` is to present the meaningful marks we see on the document. The statement of how we think these marks should be read, as a sequence of readings, should be left for the other elements within `<app>`: thus, `<rdg type="orig">` and `<rdg type="c1">`, representing the different perceived stages a particular text has gone through during a process of revision, from its original ("orig") to first corrected state ("c1"). All these statements — `<rdg type="lit">`, `<rdg type="orig">` and `<rdg type="c1">` — are interpretive. However, they are interpretive in different ways and serve different purposes.

The TEI, in "An Encoding Model for Genetic Editions" (http://users.ox.ac.uk/~lou/wip/geneticTEI.doc.html#index.xml-body.1_div.1_div.1) includes some similar elements to the ones developed for our edition of the *Commedia*. Consider this example from "An Encoding Model for Genetic Editions":

> "The following example, taken from a manuscript of Jane Austen's *Sanditon*, shows a rewriting where a pencilled passage has been fixed with ink, with some modification:

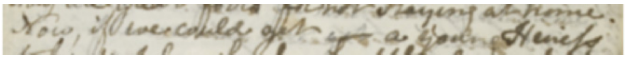


Image from page 70 of the Sanditon manuscript

> In this example, Austen sees in the fixation an opportunity to manipulate the text previously written, and thus changes the pencilled *could but get* to the inked *could get*. A simple way of encoding this might be as follows: `<ge:rewrite cause="fix" hand="ja2" stage="#s1">` Now, if we could get `<del stage="0">but</del> a young Heiress</ge:rewrite>"`

In our proposed system, not originally designed for genetic encoding but easily adaptable for this purposes, we would encode it as: `<app> <rdg type="orig">Now, if we could but get </rdg>`

```
<rdg type="c1">Now, if we could get</rdg>>
<rdg type="lit">  <seg type="overwritten">
<seg type="orig" rend="pencil">Now, if
```

we could but get </seg> <seg type="c1" rend="ink">Now, if we could get</seg> </seg> </rdg> </app> Both systems explicitly state that there was an original text, written by Jane Austen and that the text was rewritten by herself. The "lit" reading in my proposed transcription specifies that this was originally written in pencil and that the second version of the text was written in ink. In my transcription, it is clearly stated that the original form of the text was "Now if we could but get " (written in pencil) and that this was altered to "Now if we could get" (written in ink). In contrast, in the proposed TEI encoding it is not at all clear from the use of the various attributes (stages "0" and "#s1") and from the <ge:rewrite> and <del> elements what was originally written, what it was revised to, and how this was done.

For our purposes, we find that by stating explicitly what occurred to the text, that is whether it was stroke through, underdotted, underlined, erased, scrapped or rewritten, we avoid the ambiguity of the <del> element. The <add> element has also been set aside, to be replaced with markup that indicates position or mode of insertion.

The encoding system being used in our projects is TEI compliant, but it attempts to serve the purposes of various types of scholarly editions. For the purposes of textual scholarship, it is necessary to have a clear idea of what is meant by "representing the text of the document." In any case, it is not enough to represent the text of the document, we must also represent variant states of the text and give editorial interpretation a more explicit place in the transcription of primary sources.

References

Bordalejo, B. (2010). The encoding system. In Prue Shaw (ed.) *The Commedia of Dante Alighieri: A Digital Edition*. Saskatoon: Scholarly Digital Editions.

Burnard, L., F. Iannidis, E. Pierazzo, and M. Rehbein (n.d.) An Encoding Model for Genetic Editions. http://www. tei-c.org/Activities/Council/Working/tcw19.html . (Accessed November 1st, 2012).

Gabler, H. W. (2007). The Primacy of the Document in Editing. *Ecdotica*, 4, 197–207.

Pierazzo, E. (2011). A Rationale of Digital Documentary Editions. *Literary and Linguistic Computing*, 26, 463–477.

Robinson, P. (2013). Towards a Theory of Digital Editions. *Variants* 10, 105-32.

Tanselle, G. T., (1978). Editing Historical Documents *Studies in Bibliography* 31, 1-56.

Tanselle, G. T., (1996). Editing Historical Documents *Studies in Bibliography* 49, 1-60.

Mapping DH through heterogeneous communicative practices

Bowman, Timothy
tdbowman@indiana.edu
Indiana University, United States of America

Demarest, Bradford
bdemares@umail.iu.edu
Indiana University, United States of America

Weingart, Scott B.
scbweing@umail.iu.edu
Indiana University, United States of America

Simpson, Alicia
herbert_alicia@wheatoncollege.edu
Wheaton College (Norton, MA) United States of America

Neal, Grant Leyton
glsimpso@indiana.edu
Indiana University, United States of America

Lariviere, Vincent
vincent.lariviere@umontreal.ca
Université de Montréal, Canada

Thelwall, Mike
M.Thelwall@wlv.ac.uk
University of Wolverhampton, UK

Sugimoto, Cassidy R.
sugimoto@indiana.edu
Indiana University, United States of America

Objective

Digital Humanities (DH) has been exhaustively defined in the literature (e.g., Rockwell, 2002; Bellamy, 2012; Text Analysis Portal for Research, 2011; Fitzpatrick, 2011). Such definitions are sometimes at odds with each other and often represent differences based upon disciplinary

concerns. Despite the assertion that DH is a "term of tactical convenience" (Kirschenbaum, as cited in Gold, 2012), the existence of a DH community seems to be well-established; there are a dizzying array of scholars identifying themselves as digital humanists and there are others doing work that some have categorized as DH. However, a thorough investigation and description of the communicative practices of DH is lacking. We know neither the breadth of methods used, the depth to which they are used, nor the purposes to which they are put. To this end, this paper examines informal and formal communication channels used by members of the DH community to diffuse information and build communities. These communications are negotiated at a variety of levels including students and faculty at the individual level, collaborative teams at the group level, and funding agencies and institutions at the societal level (Svensson, 2010). We analyze the data from these communications to determine how these interactions connect DH community members at the individual, group, and institutional levels and across the DH landscape and helps answer the question: How does the socio-technical ecology connect or partition the landscape of the DH community?

Background

In a discussion of how qualitative research may aid bibliometric analyses of the humanities, Sula (2012, para. 18) claims that "a fuller picture of the humanities will help to clarify the ways in which the humanities and sciences differ, beyond citation patterns and authorship practices", calling for studies that look to both formal scholarly communication and informal communication from sources deriving from mentoring, peer-to-peer, and other relationships (built on interactions such as conference co-attendance, editorship, and contributorship to anthologies). Sula (2012) concludes by suggesting that these proposed studies based on expanded sets of communications look to apply the methods of network analysis and visualization. Our proposed study answers this call both in terms of the data used and the methods of analysis.

Previous studies outside the realm of formal scholarly communication in the DH domain have begun this expansion of information sources, examining DH Twitter communications (Ross, Terras, Warwick, & Welsh, 2011), syllabi (Terras, 2006; Spiro, 2012), journal citation analysis (View DHQ, 2012), and research centers (Zorich, 2008), exposing the diversity of scholarly communication activities in DH; however such studies have been limited for the most part to single channels of communication. In Terras (2011), an infographic quantifying DH produced by the UCL Centre for Digital Humanities displays DH's burgeoning internationality as well as its institutionalization. Still another dimension of diversity is addressed by McPherson (2008) via Svensson (2010) — namely, a diversity of topicality, defined as foci in digital humanities upon computing, blogging, and multimodality. The current study addresses the demands of these multiple diversities to investigate divisions in the overall DH landscape, while doing so across multiple communication channels in order to discern how different dimensions of diversity and division may or may not overlap.

Methods

This work will apply multi-dimensional network analysis to data from Twitter, *LLC* and *DHQ* journals (data taken from the Web of Knowledge database), NEH grants awarded for DH-related projects, the TEI-L and Humanist listservs, DH syllabi, and a variety of other sources (blogs, centers, and projects), employing a cumulative, normalized database composed of data from these sources to paint a wider view of the connections among people, teams, institutions, and communication channels that make up the DH landscape. Our sources will be validated through consultation with prominent members of the DH community. The resulting normalized database will be rendered as a graph connecting URLs, projects, institutions, people, publications, and grants, which will then be partitioned and analyzed using standard community detection algorithms. We will then compare community overlap over different scholarly media to explore how DH practitioners organize themselves into and across communities, specifically looking at whether certain people, technologies, or publications sit at intersection points in the network, holding communities together.

Significance

This research is innovative in its combination of both formal (syllabi, journals, grant proposals, etc.) and informal (Twitter, blogs, listserv, etc.) communication channels allowing for a broader analysis of the communication network of the DH community. Previous work has focused on single source types and has marginalized community members who communicate in other ways. There is a vacuum of formal DH connectivity and this work addresses ways in which this vacuum is being filled and what that implies about the DH community. The DH community forms a network spanning across the world (Kamada, 2010) and it's important to understand how this network is connected and how it is establishing itself in traditional academic institutions (Adams & Gunn, 2012). From a broader perspective, the methodology introduced here to

study DH is generalizable to the analysis of other fields and will hence make a valuable contribution to scholarship.

Because DH community members are situated in various locales across a wide array of institutions, there are few formalized communication channels that span the DH landscape. The lack of formalized communication channels and instructional structure indicates that multi-dimensional methods are needed to fully comprehend this network; this premise informs our selection of formal and informal data sources. It is important to note that the DH community is an area of research made up of theories, methods, and people spanning multiple domains who publish across a variety of disciplines; that said, we will not be providing an exhaustive analysis of the entire landscape of DH. Examination of a large swath of this landscape allows for a wide-ranging analysis of the various channels used to keep those in the DH community informed. It is important for members of the DH community to be made aware of the various channels of communication that are being used to spread information. As Terras (2010) stated in her plenary speech to the DH2010 conference, "digital presence and digital identity is becoming more important to Digital Humanities as a discipline." This work addresses this statement with empirical and heterogeneous evidence.

Acknowledgments

This work was funded by the Digging into Data initiative, organized by the National Endowment for the Humanities. Specific funding for this DID project comes from the National Science Foundation in the United States (Grant No. 1208804), JISC in the United Kingdom, and the Social Sciences and Humanities Research Council of Canada.

References

Adams, J. L. & D.B. Gunn (2012). Digital humanities: Where to start. *College & Research Libraries News,* 73(9): 536-569. http://crln.acrl.org/content/73/9/536.full

Bellamy, C. (2012). The sound of many hands clapping: Teaching the digital humanities through virtual research environments (VREs). *Digital Humanities Quarterly*, 6(1): http://www.digitalhumanities.org/dhq/vol/6/2/000119/000119.html

Fitzpatrick, K. (2011). The humanities, done digitally. *The Chronicle of Higher Education.* http://chronicle.com/article/The-Humanities-Done-Digitally/127382

Gold, M. K. (ed.) (2012). *Debates in the Digital Humanities.* University of Minnesota Press.

Juola, P. (2008). Killer applications in digital humanities. *Literary and Linguistic Computing,* 23(1): 73-83.

Kamada, H. (2010). Digital humanities: Roles for libraries. *College & Research Libraries News,* 71(9): 484-485. http://crln.acrl.org/content/71/9/484.full

Kirschenbaum, M. (2010). What is digital humanities and what's it doing in English departments? *ADE Bulletin,* 150: 1-6.

Rockwell, G. (2002). Multimedia: Is it a discipline? The liberal and servile arts in humanities computing. *Yearbook of the Seminar for Germanic Philology 4*. Paderborn: Mentis Verlag. http://computerphilologie.uni-muenchen.de/jg02/rockwell.html

Ross, C., M. Terras, C. Warwick, and A. Welsh (2011). Enabled backchannel: Conference Twitter use by digital humanists. *Journal of Documentation*. 67(2): 214–237

Rosenbloom, P. S. (2012). Towards a conceptual framework for the digital humanities. *Digital Humanities Quarterly* 6(2). http://digitalhumanities.org/dhq/vol/6/2/000127/000127.html

Schriebman, S. (2012). Digital humanities: Centres and peripheries. *Historical Social Research.* 37(3): 46-58.

Spiro, L. (2011). Knowing and doing: Understanding the digital humanities curriculum. (Powerpoint slides) http://digitalscholarship.files.wordpress.com/2011/06/spirodheducationpresentation2011-4.pdf

Spiro, L. (2012). Models for supporting digital humanities at liberal arts colleges. (PowerPoint slides) http://digitalscholarship.files.wordpress.com/2012/05/spirodhsupportstructureswooster2.pdf

Sula, C. A. (2012). Visualizing social connections in the humanities: Beyond bibliometrics. *Bulletin* 38(4): Retrieved from http://www.asis.org/Bulletin/Apr-12/AprMay12_Sula.html

Svensson, P. (2010). The landscape of digital humanities. *Digital Humanities Quarterly* 4(1). http://digitalhumanities.org/dhq/vol/4/1/000080/000080.html

Svensson, P. (2012). Envisioning the digital humanities. *Digital Humanities Quarterly* 6(1). http://www.digitalhumanities.org/dhq/vol/6/1/000112/000112.html

Text Analysis Portal for Research. (2011). How do you define Humanities Computing/Digital Humanities? http://tinyurl.com/9lw97hm (accessed 7 March 2012)

Terras, M. (2010). DH2010 Plenary: Present, not voting: Digital humanities in the panopticon. http://melissaterras.blogspot.com/2010/07/dh2010-plenary-present-not-voting.html (accessed 10 July 2010).

View DHQ (Data Visualization). (2012). 2012 ACH Microgrant: Citation Network Visualization for *Digital Humanities Quarterly*. http://digitalliterature.net/viewDHQ/

Wagner, C. S. (2008). The topology of science in the twenty-first century. In *The New Invisible College* Washington, D.C.: Brookings Institution Press. 15-32.

Zoirch, D. M. (2008). A survey of digital humanities centers in the United States. *Council on Library and Information Research. (CLIR)* held 12 May 2007. http://www.uvasci.org/wp-content/uploads/2008/06/dhc-survey-final-rept-2008_05_22-for-distribution.pdf, accessed on October 1, 2012.

Fitting Personal Interpretations with the Semantic Web

Bradley, John

john.bradley@kcl.ac.uk
Department of Digital Humanities, King's College London, United Kingdom

Pasin, Michele

michele.pasin@gmail.com
Department of Digital Humanities, King's College London, United Kingdom

The emergence of formal ontologies into the World Wide Web has had a significant effect on research in certain fields. In parts of the Life Sciences, for example, key research information has been captured in formal domain ontologies, like those mentioned in the Open Biological and Biomedical Ontologies website (OBOFoundary 2012). In parallel with this has been the development of the AO annotation ontology framework (AO 2012) which formalises the act of annotation as a way to connect ontologies such as those in the OBOFoundary to references to them in the scientific literature: an act sometimes referred to as "semantic annotation", and tools such as the SWAN annotation system (SWAN 2008) have emerged to support this. We will call the activity of linking references in a domain literature directly to entities in one or more domain ontologies "direct semantic annotation", and show it in schematic form in figure I. The annotations — shown as heavier lines connecting spots in the literature (to the left) to the ontologies (to the right) could be expressed in the AO annotation ontology, or something similar to it.

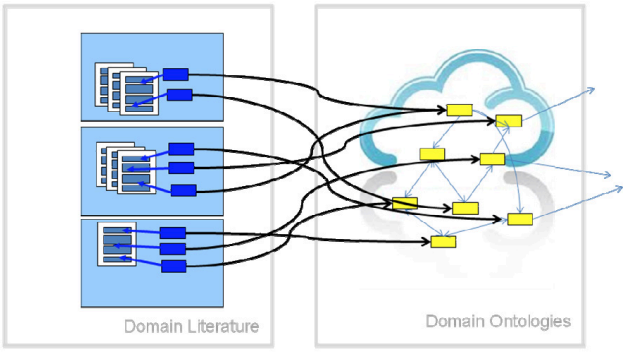


Figure 1:
direct semantic annotation

Can direct semantic annotation like this be applied to research in the Humanities? For it to work as it does in the Life Sciences, formal models of humanities materials, built using an ontology framework such as CIDOC-CRM, would need to exist and be already accepted as a useful representation of material of interest to the humanities. Not much of this has happened at present, although perhaps Linked Data initiatives (Heath and Bizer 2011) show some promise in that general direction, and we in DDH are both exploring making data from our projects available in the Semantic Web sense by, for example, setting up SPARQL endpoints, and thinking about the open-data significance of our structured prosopography model (Pasin and Bradley 2012).

Although open, linked data provides a context for exploring direct semantic annotation and even though there is evidence of this being thought useful in the Life Sciences, by itself the mere act of linking a spot in a published online journal to a relevant bit of an online ontology does not represent anything other than rudimentary research activity in either the Life Sciences or in the Humanities. Instead, almost all humanities scholars want to spend their time, not connecting things they read only to an existing, shared, understanding of things, but instead developing their *own* original interpretation of the materials they study, and they aim to subsequently explore these new concepts and paradigms in articles and books that they write. (see Brockman et al 2001 and in Palmer *et al* 2009) For traditional humanists, their scholarship does not start out only with predefined formal structures such as those provided through their community's shared concepts, but begins with a set of vague notions and insights that emerge more clearly over time in the scholar's mind as they read, and that only over time becomes clear enough to be described in original published work. For most humanists scholarship (a) is normally personal, (b) is meant to produce original ideas that must first emerge and then mature over time, and (c) even when the ideas are mature enough for publication, represents a structure that is at least "pre-

ontological", and perhaps at best only partly compatible with the clarity of ontological modelling.

Surprisingly, however, although products of humanities scholarship do not seem currently to match the formalisms of computer ontologies as perhaps some of the Life Sciences do, there is evidence of some degree of inherent structure in the process of creating them. Many researchers, including Brockman and Palmer mentioned earlier, have noted the importance of notetaking and the management of those notes in humanities research. We can see a significantly structured approach around the process of developing new scholarly ideas when we look at traditional strategies for taking and managing notes as described in books like Altick and Fenstermaker's *The Art of Literary Research* (1992). Altick and Fenstermaker describe paper-based procedures aimed to provide the new researcher with a methodology to organise notes taken while reading into a structure of topics and concepts that will eventually contribute into the writing of articles that represent publishable thinking. There is structure in Altick and Fenstermaker's approach: figure 2 shows this process in schematic form. We see the original notes created during the reading of articles and books on the left, these notes contributing thoughts in the mind of the scholar that eventually allows him to create new concepts in the middle (only 2 "concepts" are shown here, but a real user would have many more), and (towards the right) these new emerging concepts fitting with references to original sources and supporting an argument for new ideas to be presented through the writing of papers. Note the difference from direct semantic tagging in figure I: the annotations do not link directly to preexisting formal ontological entities, but first appear as informal prose notes that may, as the researcher's understanding grows, contribute to a more formal set of ideas and then emerge as entity-like objects in the form of personally developed new concepts, themes, ideas, etc.

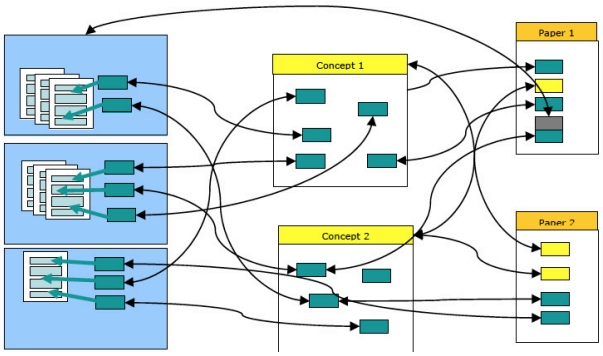


Figure II:
a model of the flow of ideas in traditional scholarship

If we wish to explore how traditional personal scholarship could connect to the formal world of RDF and ontology schemes such as the Open Annotations Collaboration (OAC 2011), we need a model that reflects some of these aspects of the act of doing personal scholarship. The models behind the Pliny project (Bradley 2008 and 2012) fit this bill, since Pliny was launched precisely to explore how computing could facilitate exactly this traditional scholarly practice. Pliny tried to be "Englebartian"— referring to Douglas Englebart's H-LAM/T paradigm (Englebart 1962) that successful software integrates with the human way of doing intellectual things so well as to almost disappear, and that this disappearing software can, paradoxically, sometimes allow its users to do entirely new things that they had been previously incapable of doing. Out of this work came two models: the interface which developed a particular view of how users might usefully interact with a note-taking and note management tool to help them develop their own interpretation, and the data model that stored the information. Bradley 2008 describes Pliny's user interface in terms of affordances: 2-dimensional space, containment and hierarchy, naming and labelling and multiple reference of notes material in different contexts, including typing of a reference.

It turns out that Pliny's data model, as well as being designed to represent aspects of traditional scholarship in its three phases, is strongly suggestive of RDF and broader ontological technologies. Like RDF, the structure is a network and the links between the network nodes can be typed in a way similar to a RDF predicate. Pliny from its first release had the ability to export its structure into a Topic Web format, and some preliminary work has been done (see Jackson 2010) to map Pliny data into RDF through the OAC ontology. Further work has been carried out by us to take Jackson's approach further and better map an interpretation as stored in Pliny into an RDF representation.

The resulting paradigm is one that, unlike direct semantic annotation (as in figure I), separates the annotation of the domain literature from the highly formal world of shared domain ontologies by injecting a personal interpretative component in-between. One introduces, in ways compatible with semantic web technologies, a personal, more informal, and emerging representation of the scholarship into the picture.

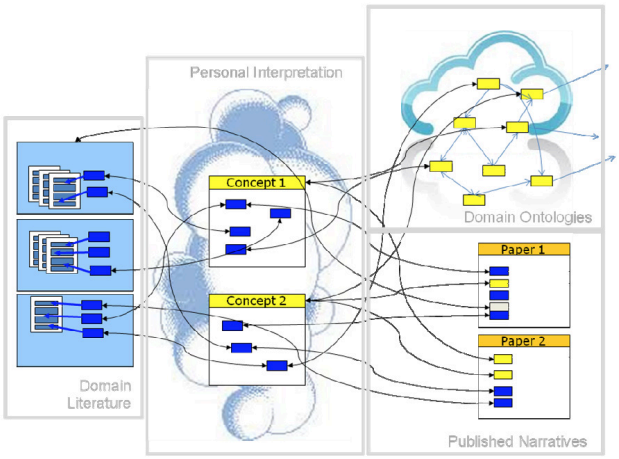


Figure 3:
The place of a structured personal Interpretation

Figure 3 is similar to the direct semantic annotation model shown in figure 1, but adds a structured area, representing aspects of the personal interpretative work of an individual, between a scholar's reading material to the left, and any shared public domain ontologies and linked data to the top right. This interjected personal interpretation "cloud" might well never be as clear-cut as formal ontologies must be, but its presence here recognises and enables the *process* towards formality that is a central part of interpretation in humanities scholarship. By interposing this somewhat-informal semantic "cloud" between the texts and the formal ontologies of the semantic web, we see a way of thinking about this central personal interpretive work that fits with the larger, more formal, semantic web picture. Although it would seem that the nature of traditional humanities research does not suit the standard direct semantic annotation model currently active in parts of the Life Sciences, we propose here an approach that, over time, encourages the researcher to turn their clouds of personal interpretation into material that might become more and more compatible with computer ontologies and the semantic web.

This presentation will describe work that was first shown in a preliminary fashion in the NeDiMaH Ontology workshop at DH2012 (Bradley and Pasin 2012), but that has continued since then and reaching a significant stage of development.

References

Altick, R. D., and J. J. Fenstermaker (1992). *The Art of Literary Research*. New York: W. W. Norton & Company.

AO (2012). *AO: Annotation-ontology*. Website at http://code.google.com/p/annotation-ontology/

Bradley, J. (2008). Thinking about Interpretation: Pliny and Scholarship in the Humanities. *Literary and Linguistic Computing*. 23(3). 263-79. doi: 10.1093/llc/fqn021. Online at http://llc.oxfordjournals.org/cgi/reprint/fqn021?ijkey=3UzJDubDB0FRQcR&keytype=ref

Bradley, J. (2012). Beyond Digital Media: Moving beyond a 'media' orientation in the annotation of digital objects. In press at the *Digital Humanities Quarterly*. A draft (under a different name) of this article is available at http://pliny.cch.kcl.ac.uk/docs/article-2011.pdf

Bradley, J., and M. Pasin (2012). Annotation and Ontology in most Humanities research: accommodating a more informal interpretation context. *DH2012 NeDiMaH Ontology Workshop*. held 17 July, 2012.

Brockman, W. S., L. Neumann, C. L. Palmer, T. J. Tidline. Council on Library and Information Resources. (2001). *Scholarly Work in the Humanities and the Evolving Information Environment.* Washington, D. C.: Digital Library Federation, Council on Library and Information Resources. (Online version at http://www.diglib.org/pubs/dlf095/

Englebart, D. (1962). Augmenting Human Intellect: A conceptual framework. Stanford CA: Stanford Research Institute. Online at http://www.bootstrap.org/augdocs/friedewald030402/ augmentinghumanintellect/AHI62.pdf (accessed March 2007)

Heath, T., and C. Bizer (2011). *Linked Data: Evolving the Web into a Global Data Space.* 1st edn. *Synthesis Lectures on the Semantic Web: Theory and Technology*. 1(1). 1-136. Morgan & Claypool.

Jackson (2010). *RDF-encoding Pliny annotations in the Open Annotation Collaboration project*. University of Illinois: GSLIS Technical Report #ISRN UIUCLIS--2010/2+OAC.

OAC (2011). *Open Annotation Collaboration*. Website at http://www.openannotation.org/

OBOFoundary (2012). *The Open Biological and Biomedical Ontologies*. Website at http://obofoundry.org/ .

Palmer, C. L., L. C. Teffeau, and C. M. Pirmann. OCLC Research. (2009). *Scholarly Information Practices in the Online Environment: Themes from the Literature and Implications for Library Service Development*. Available online at: http://www.oclc.org/research/publications/library/2009/2009-02.pdf (.pdf: 412K/59 pp.).

Pasin, M., and J. Bradley (2012). *Factoid-based Prosopography and Computer Ontologies: towards an integrated approach*. Draft paper submitted for possible publication. Available online at http://www.michelepasin.org/papers/43/

SWAN (2008). *SWAN Project: Semantic Web Applications in Neuromedicine*. Website at http://swan.mindinformatics.org/ .

Prosopography in the time of Open data: Towards an Ontology for Historical Persons

Bradley, John
john.bradley@kcl.ac.uk
Department of Digital Humanities, King's College London, United Kingdom

Along with open data about sources and places, open data about historical people provides one of the three legs for an open framework for historical information, since it is one of the three most obvious "entry points" into a potentially shared world of historical data.

DDH has been involved in a number of prosopographical projects (PBE, PBW, PASE, CCEd, PoMS, BoB[1] and most recently the *Making Charlemagne's Europe* project), and we have become convinced that prosopography is particularly well suited to the world of open and linked data, since its purpose is to establish unique digital identities for people that could be used by other researchers. Furthermore, although we here at DDH have done quite a bit of work on structured prosopography we have found that "one size does not fit all" when it comes to structuring the prosopographical materials. Several of our projects have, at their core, a common approach we are calling "factoid prosopography". However, several of them (Clergy Church of England, Early Modern London Theatres) [2] do not. Finally, of course, many other people take on prosopographical work in their research without taking up a model of their data like ours: one thinks of the "personography" model work within the TEI, for example, and we know of other colleagues with prosopographical components that have developed a different model for their data.

The issue of prosopography as a open, shared resource goes beyond the provision merely of an identity for historical persons. In the same way that CIDOC-CRM (CIDOC-CRM 2011) was developed to record information about cultural heritage objects and therefore not only help uniquely to identify them but also to allow different systems to share a much broader range of data *about* these objects, a prosopography for historical persons would work to find common ground among the different ways in which different projects collect data about historical persons so that semantic links between them could more effectively be exploited. Indeed, it would seem that much of such a prosopographical ontology would usefully be based on CIDOC-CRM, since there is often significant overlap between the two domains: prosopography often comes into the work of people working with heritage objects, and there are thus substantial elements of prosopography in CIDOC-CRM. We have done some exploring of this idea in a recently created paper by Michele Pasin and myself (Pasin and Bradley, 2012) that takes the "factoid" in our prosopographical model and explores how it might be represented primarily in CIDOC-CRM terms.

We propose calling this shared ontological model for prosopography an *Ontology for Historical Persons.* By the very nature of such a thing, it would need to be developed out of as broad a shared perspective on the different ways that prosopographical data is being collected as possible. As an early step in thinking about this ontology, we have already been in touch with not only our project partners for the prosopographies we have already done, but also with a range of colleagues who have worked on the question of structured prosopography outside of our own projects, and with colleagues who have worked on the development of CIDOC-CRM. In our talk we would also describe the current state of our ontology, and explain how we think it would link with CIDOC-CRM. However, we'd be most interested in hearing of other people who are working on structured prosopography, and who would be interested in working with us towards enabling the open data potential that an ontology for historical persons might provide.

References

CIDOC-CRM (2011). *The CIDOC Conceptual Reference Model website.* At http://www.cidoc-crm.org/

Pasin, M., and J. Bradley. (2012). *Factoid-based Prosopography and Computer Ontologies: towards an integrated approach.* Submitted for possible publication. Draft available at http://www.michelepasin.org/papers/43/

Notes

1. PBE: Prosopography of the Byzantine Empire, PBW: Prosopography of the Byzantine World (http://www.pbw.kcl.ac.uk), PASE: Prosopography of Anglo-Saxon England (http://www.pase.ac.uk), CCEd: Clergy of the Church of England Database (http://www.theclergydatabase.org.uk/index.html), PoMS: People of Medieval Scotland (http://www.poms.ac.uk), BoB: Breaking of Britain (http://www.breakingofbritain.ac.uk/).
2. http://www.emlot.kcl.ac.uk

Preliminaries: The Social Networks of Literary Production in the Spanish Empire During the Administration of the Duke of Lerma (1598-1618)

Brown, David Michael
dbrow52@uwo.ca
University of Western Ontario, Canada

Suárez, Juan Luis
jsuarez@uwo.ca
University of Western Ontario, Canada

The "preliminaries" section of a 17th-century book encompasses the pages appearing in the printed text before the beginning of the work itself. This information is divided into seven different types of documents: details of publication, documentation of censorship (both civil and ecclesiastical), licensing, selling price, dedications, letters, and errors. The importance of the preliminaries for this project lies in the information present in these sections: the names of the officials signing the documents, their governmental/institutional affiliation, dates, place of issue, and literary circles that appear in the form of dedications and poetry written by various authors and published in their friend's or associate's books. In a few pages, the preliminaries give a complete image of the formal process required for the publication of each work of literature. By compiling all this information into a graph database and performing queries specific to various research questions, we have at hand a valuable source of information about the historical networks that influenced the publication of Early Modern Spanish literature.

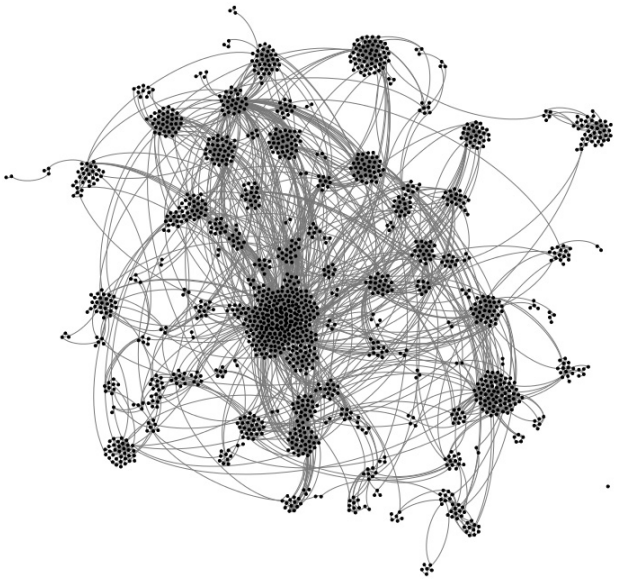
To get a comprehensive look at this information, we generated lists of every edition of what we consider literary texts (fiction in prose, theatre, poetry, chronicles) published during the 17th -century in the Spanish empire (Jiménez et al. 1980)(Calvo et al. 2003). As shown by the following screen shot, we have focused on acquiring every available edition of each literary work.

	RBL	PDF	Cervantes Saavedra, Miguel	El ingenioso hidalgo don Quixote de la Mancha	Bruselas	1611	Velpio
			Cervantes Saavedra, Miguel	El ingenioso hidalgo don Quixote de la Mancha	Valencia	1616	Crasbeeck
Sí	BNE	PDF	Cervantes Saavedra, Miguel	El ingenioso hidalgo don Quixote de la Mancha	Bruselas	1617	Antonio
Sí	BNE	PDF	Cervantes Saavedra, Miguel	El ingenioso hidalgo don Quixote de la Mancha	Barcelona	1617	Sorita
			Cervantes Saavedra, Miguel	Los trabajos de Persiles y Sigismunda	Madrid	1617	Cuesta
Sí	BNE-RBI	PDF	Cervantes Saavedra, Miguel	Los trabajos de Persiles y Sigismunda	Madrid	1617	Cuesta
Sí	BNE	PDF	Cervantes Saavedra, Miguel	Los trabajos de Persiles y Sigismunda	Pamplona	1617	Assiayn
Sí	BNE	PDF	Cervantes Saavedra, Miguel	Los trabajos de Persiles y Sigismunda	Barcelona	1617	Sorita
			Cervantes Saavedra, Miguel	Los trabajos de Persiles y Sigismunda	Valencia	1617	Mey
	BNE	PDF	Cervantes Saavedra, Miguel	Los trabajos de Persiles y Sigismunda	Paris	1617	Richer
	BNE	PDF	Cervantes Saavedra, Miguel	Los trabajos de Persiles y Sigismunda	Lisboa	1617	Rodríguez
Sí	BNE	PDF	Cervantes Saavedra, Miguel	Los trabajos de Persiles y Sigismunda	Bruselas	1618	
Sí	BNE	PDF	Cervantes Saavedra, Miguel	Novelas ejemplares	Madrid	1613	Cuesta

Sample of one of our acquisitions lists

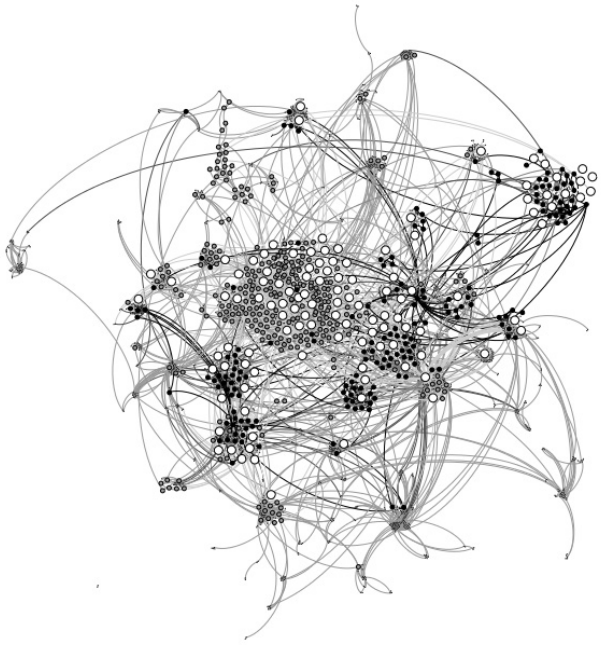
We then divided the 17th Century into periods corresponding to the different "validos" —royal favorites that served as head of government or "prime ministers" — of the various kings in order to address the changing power structures of the time and their influence in literary production (Hernán et al. 2002). Through interlibrary loans and, in some cases, trips to the libraries that hold the edition, we acquired copies of the pages of each book that make up the preliminaries section. Then, we manually built a graph database using sylvadb.com, an open source software and free graph database management service developed in the CulturePlex Lab. Within Sylva, data was stored and organized using a custom designed system of schemas based on a node/edge relationship system. Finally, we exported the database to Gephi (https://gephi.org/), a software package that allows for visualization and statistical/metric analysis of the network using built-in algorithms and Python based scripting (Bastian et al. 2009). This allows us to detect important communities within the network, key players, important objects, and hubs of production.

For this study, we have unearthed the social networks of publishing and literary creation in 17th-century Spanish literature, focusing particularly on the period during the rule of the Duke of Lerma (1598-1618). Currently the first of our *editions* lists (Duke of Lerma) consists of 330 editions, out of which we have successfully obtained 228 scanned copies of preliminaries sections: approximately 70% of the total number. Of these scans, 121 have been entered into the database, producing a graph with 1612 nodes and 3472 relationships. Rendered in Gephi using the built-in OpenOrd algorithm, the graph looks like this:



The Preliminaries graph rendered in Gephi

Using the algorithms, metric analysis tools, and filters built into Gephi we pinpointed the individuals, governmental and ecclesiastical bodies that influenced publication in this period. Also, by using the concept of "ego network" from social network analysis, we established what we call the "publication network" of some of the authors that interest us (Carrington et al. 2011). A publication network includes the editors, censors, and other individuals important in the formal process of publication, as well as any other individuals that are more directly connected to the author: friends, family, patrons, literary colleagues, etc. We determined the range of the publication network based on the internal data structure of the Preliminaries database as follows. Due to bibliographic concerns (Bowers et al. 1962) and organizational aspects of our data schema, in order to establish a connection between the author and those involved in the approval, licensing, and publication of an edition there are four steps e.g., Author->Work, Work->Edition, Edition->Approval, Approval->Censor. Therefore, to establish an author's publication network we needed to find neighbors for up to four degrees of separation. Although Gephi does not include ego network filters that extend to four degrees, using its Python based scripting console we were able to code functions that allowed us to isolate subsets within the graph to generate ego networks for any node to *n* degrees of separation. For instance, in the graph below we can see the publication networks of two authors associated with Mexico; Bernardo de Balbuena, author of *Grandeza Mexicana*; Juan de Torquemada, author of *Monarquía Indiana*; and the intersecting nodes in their publication networks:

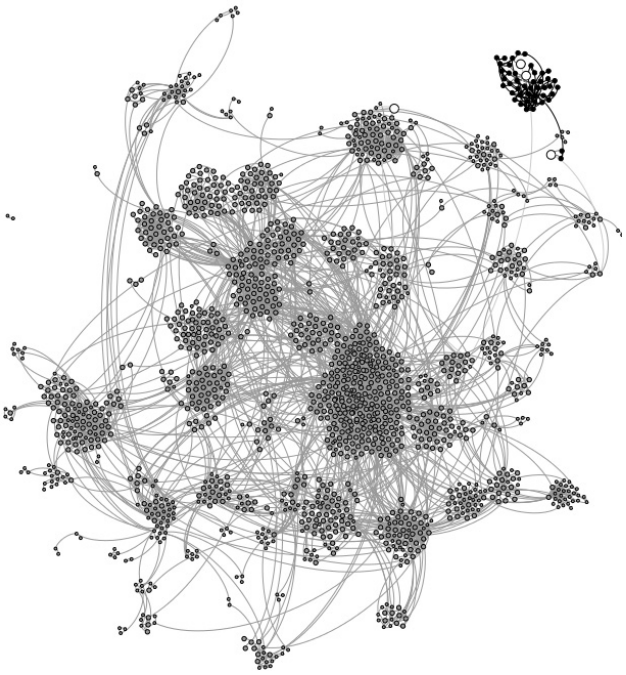


Publications Networks: Balbuena=Black, Torquemada=Grey, Intersecting Nodes=White

Using the above techniques, we set out to find and isolate the main nodes of this social network that made possible the creation and sustainability of a transatlantic network of cultural agents. The first thing that stands out in the graph is Lope de Vega and his powerful, Madrid based publication network (Martínez et al. 2011). Using the Python scripting console, we determined that Lope's publication network consists of 1083 nodes, or 67% of the nodes in the graph. This information is not new, based on the extremely prolific nature of his literary production we can assume that he was very well connected. However, we can also determine who *wasn't* in his publication network. Departing from Lope's publication network, we were able to locate the successful political and institutional connections that help us explain the central position of institutions such as the House of Zúñiga in the cultural fabric of the period.

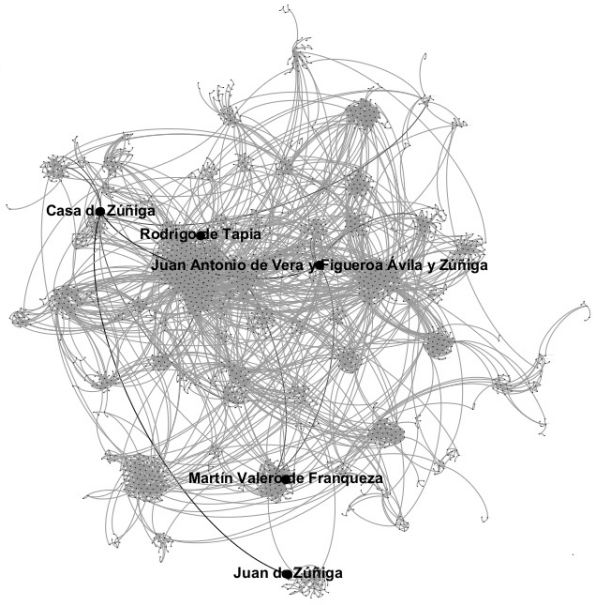
Publication network: Lope de Vega=Black

To do this we used the scripting console to remove the subset of nodes representing Lope's publication network from the other nodes that make up the graph, and returned a list of the names of all of the people who are not in Lope's publication network. A quick review of this list produced some interesting results: we found several authors based in Spain including Gonzalo de Céspedes y Meneses and El Inca Garcilaso; and two authors active in Peru, Diego Dávalos y Figueroa and Pedro de Oña. While a quick look at both Céspedes and El Inca produced interesting results, the two Lima based authors attracted our attention. In this period social circles were highly influenced by geography, and it is logical that these authors find themselves at the periphery of a network centered geographically in Madrid. However, despite geographic concerns both authors remain connected to Lope de Vega's network. We found that both Oña and Dávalos y Figueroa are connected to Lope's network at 3 degrees of separation through their dedications to the Viceroy of Peru, Luis de Velasco y Castilla; and at four degrees through Juan de Zúñiga, Diego de Ojeda, and the Order of Santiago:



Publication networks: Dávalos y Figueroa=Black, Lope de Vega=Grey, Intersecting nodes=White

In order to contextualize the Peruvian network we compared the aforementioned "Mexican" authors with the "Peruvian" authors. Combining the four social networks into two based on geographic constraints, we found that at 4 degrees of separation there was no direct overlap, so we upped the parameter to 5 degrees of separation and produced the following image:



Publication Networks: Intersection between Mexican and Peruvian Networks

As shown here, even at five degrees of separation there are few overlaps between the networks. However, in the above image we begin to notice the importance of the House of Zúñiga. It is well known that the House of Zúñiga was powerful in both Spain and the Americas, and also that certain members of this house were important patrons of the arts and literature (Cátedra 2003; Díez Fernández et al. 2005). Nonetheless, we don't think that their role in transatlantic literary production has been adequately explored. The political importance of this family in New Spain is obviously important (an Archbishop and a Viceroy); however, the Preliminaries graph illustrates not only the political role this house played in America, but also the importance of political figures/nobility in publication circles and how the members of one house can spread their cultural influence throughout geographic space. To take this concept one step further, we followed the Zúñigas back the Spain. Here we find the Duke of Béjar, Alonso López de Zúñiga y Pérez de Guzmán, and the first part of Don Quixote. It turns out that American authors were not the only artists soliciting support from the House of Zúñiga: Miguel de Cervantes dedicated part 1 of Don Quixote to the famous Duke of Béjar(Rico 2005).

The above samples show the potential of a research model that combines network-based analysis with quantitative and qualitative studies of cultural production, providing evidence of the interaction between political structures and cultural production in the Spanish Empire (Martínez et al 2008). By repurposing bibliographic data, the Preliminaries Project allows us to explore the concept of cultural networks within the framework of transatlantic studies and complexity theory (Wood 2010; Suárez 2007). Furthermore, this study demonstrates the effectiveness of digital humanities methods as a tool to locate previously overlooked areas for further study using a more traditional humanistic approach.

References

1. **Pedraza Jiménez, F. B., and M. R. Cáceres.** (1980). *Manual de literature española*. Pamploa: Cénlit.

2. **Huerta Calvo, J. (dir.)** (2003). *Historia del teatro español*. Madrid: Gredos.

3. **García Hernán, E.** (2002). *Políticos de la monarquía hispánica (1469-1700)*. Madrid: Fernández Ciudad.

4. **Bastian M., S. Heymann, and M. Jacomy** (2009). "Gephi: an open source software for exploring and manipulating networks." International AAAI Conference on Weblogs and Social Media.

5. **Carrington, P. J., and J. Scott** (2011). *The SAGE Handbook of Social Network Analysis*. Los Angeles: Sage.

6. **Bowers, Fredson.** (1962). *Principles of Bibliographic Description*. New York: Russell & Russell.

7. **Martínez, J. F.** (2011).*Biografía de Lope de Vega, 1562-1635: un friso literario del Siglo de Oro*. Barcelona, PPU.

8. **Cátedra, P. M.** (2003). *La "Historia de la Casa de Zúñig"a otrora atribuida a Mosén Diego de Valera*. Salamanca: Gráficas Cervantes.

9. **Díez Fernández, J.; I., and G. Santonja.** (2005). *El mecenazgo literario en la casa ducal de Béjar*. Burgos: Instituto Castellano y Leonés de la Lengua.

10. **Rico, F.** (2005). *El texto del "Quijote": preliminares a una ecdótica del Siglo de Oro*. Barcelona: Ediciónes Destino.

11. **Martínez Millán, J.;, and M. A. Visceglia (eds.)** (2008). *La monarquía de Felipe III*. >Madrid: Cyan, Proyectos y Producciones Editoriales. Print.

12. **Wood, A. T.** (2010). Fire, Water, Earth, and Sky: Global Systems History and the Human Prospect. *The Journal of the Historical Society*. X:3: 287-318.

13. **Suárez, J. L.** (2007). Hispanic Baroque: A Model for the Study of Cultural Complexity in the Atlantic World. *South Atlantic Review*. 72(1): 31-47.

Text Encoding, the Index, and the Dynamic Table of Contexts

Brown, Susan

sbrown@uoguelph.ca
School of English and Theatre Studies, University of Guelph, Canada; Department of English and Film Studies, University of Alberta

Adelaar, Nadine

adelaar@ualberta.ca
Department of English and Film Studies, University of Alberta

Ruecker, Stan

sruecker@ualberta.ca
Institute of Design, Illinois Institute of Technology

Sinclair, Stéfan

stefan.sinclair@mcgill.ca

Department of Languages, Literatures and Cultures, McGill University

Knechtel, Ruth
rknechte@ualberta.ca
Department of English and Film Studies, University of Alberta

Windsor, Jennifer
jwindsor@ualberta.ca
Humanities Computing, University of Alberta

Short Abstract

This paper investigates the nature of the index and its role within scholarly publishing by means of an experiment using the Dynamic Table of Contexts Browser to publish a scholarly essay collection that offers a full intellectual index combined with semantic encoding and free-text search functionality.

Extended Abstract

The index has long been a feature of printed text. The noun form of the word developed from various material pointers, whether literal fingers or portions of instruments, into the more abstract concept of a sign or token that emerged simultaneously with the lists placed at the backs of books in the late sixteenth century ("Index," def. n. 1, 2a, 4b, 5b).

The concept of the index is all over the digital world. Symptomatically, the most recent change to the meaning of the verb form emerged from computing in 1962 ("Index," def. n. 5d). The digital humanities themselves trace their origin back to Father Busa's concordance of the *Index Thomisticus*, a machine-generated index produced at IBM (Burton). Yet the conceptual index produced through the intellectual engagement of a human being with the meaning of the text, that is, the kind of index that we as scholars most value from print culture, is exceedingly rare in the context of digital texts, where instead the automated, machine-generated index abounds.

The immense gains resulting from the ability of computers to generate indefatigably exhaustive and unimaginably extensive indices or to deliver with lightning speed the portions of a text containing a word on which the user searches are undeniable. However, the fate of the intellectual index is uncertain given the rise of digital books generally and of semantic encoding within digital humanities publishing in particular. This paper reflects on the role of the index within scholarly publishing by means of an experiment using the Dynamic Table of Contexts Browser to publish a scholarly essay collection that offers a full professionally produced intellectual index, semantic encoding of recurrent features, and free text search.

The Dynamic Table of Contexts Browser was designed as a reading environment for digitally encoded texts that would combine the table of contents with tools drawing on embedded semantic markup to create new affordances for reading in a digital environment. The familiarity of the table of contents provides prospect and navigational assistance, while the browser leverages the embedded semantic markup to allow users, as a description of an earlier instance of the browser put it, "to add or subtract what are essentially index items in and out of the table of contents" (Ruecker et al., 180; Nelson et al., 12). The combination aims to provide an advance over the standard fixed-content, if expandable and contractable, digital table of contents, and thereby solve some key usability challenges related to lengthy digital texts.

What an index item is, "essentially," however, is far from simple. The investigation described here revolves in part around the tension between free text and controlled vocabulary as the basis for indexing. There has long been an unresolved, antagonistic relation between free-text terms and controlled vocabulary terms.[1] Both describe ways of representing and retrieving information in a digital context. Yet, long before the invention of the computer, scholars quibbled over the value and effectiveness of uncontrolled vocabularies and free-text searches (Svenonius, 333). This controversy falls into three stages: the nineteenth century "title-term or title-catchword indexing" of library catalogues; the invention of keyword in context (KWIC) indexing in 1959 by Hans Peter Luhn; and the rise of "instantaneous keyword searching" so familiar to us today (Svenonius, 333; Garret). These debates frequently focus on optimizing access to the vast collections of digital texts on and off the web, but at the level of the book the emphasis on generalized controlled vocabularies gives way to the question of the value of highly granular and customized intellectual indexes, produced by professional indexers, which seem to be struggling for survival against the generalized free-text search function offered by eBook interfaces.

At stake in the transition from the traditional print index to digital modes of retrieving and organizing knowledge is not only the profession of the indexer, but the broader socio-cultural signification and future of the 'index'. The first generation of eBooks often omitted indexes entirely, even when present in the printed book, or included them as passive page images lacking navigational features.[2] The American Society for Indexing (ASI) is working with the International Digital Publishing Forum (IDPF) to

"ensure inclusion of usable indexes in nonfiction digital book formats and e-books" ("Digital Trends Task Force") in the creation of the specifications for EPUB 3.0.[3] The new workflow required to "output in various formations (e.g. eBook, HTML, PDF for print, etc.)" is part of the challenge for the publishing industry, to which XML seems to be emerging as a leading solution, due to its separation of "function from layout" ("Moving to XML Workflow"). However, the differences between encoding and conventional indexing practices do not figure in considerations to date (MacGlashan).

Semantic encoding is, after all, indexical. It demarcates and hence allows an interface to point to a section of a text; it labels a span of text according to a controlled vocabulary of values that has been devised to elucidate the nature of the text being encoded and conceptually groups that particular span of text with all other spans that have been encoded likewise. The `<index>` tag in the Text Encoding Initiative tagset is defined as follows: "(index entry) marks a location to be indexed for whatever purpose" (TEI Consortium, "TEI element"). The TEI documentation provides an excellent summary of the tension between what it terms manual indexing and free-text search:

> The indexing of scholarly texts is a skilled activity, involving substantial amounts of human judgment and analysis. It should not therefore be assumed that simple searching and information retrieval software will be able to meet all the needs addressed by a well-crafted manual index, although it may complement them for example by providing free text search. The role of an index is to provide access via keywords and phrases which are not necessarily present in the text itself, but must be added by the skill of the indexer. ("TEI Consortium, "3 Elements")

This begs the question further, however, about the indexical function of markup, since most TEI indexing is still "manual" and much scholarly markup that uses the TEI involves tags, particularly named entities, that overlap substantially with a conventional index. Other semantically oriented schemas, such as the bespoke one developed by the Orlando Project to encode feminist literary history, provide numerous tags that overlap considerably with the terrain of the traditional index: a tag like `<education>` or `<relationsWithPublisher>` serves to index portions of the text collection that may not contain such keywords or phrases, but which have been identified as relevant to these concepts by the skilled encoders (Brown et al.). There are significant differences between markup and professional or manual indexing,[4] but the extent of overlap between the two is evident in the ways that such markup functions within interfaces which offer the user the ability to look up spans of text marked with those terms. Yet whereas markup generally is considered quite synonymous with indexing, broadly conceived, particularly where a controlled vocabulary is being employed, whatever the theoretical value of the intellectual index, it seems in practice to have been deemed as dispensable to online digital humanities projects as it has been to early eBooks.

Our survey of a range of online projects, as well as of systematic reviews of features of the digital edition, show little evidence that the backbone of scholarly print resource navigation, the semantic index, is deemed crucial to digital scholarly resources.[6] This probably has in part to do with cost balanced against perceived benefits.[7] User studies in information science show that the strengths of the manual index are rivaled or even outstripped by automated indexing and information retrieval: "users find them, on balance, more or less equally effective" (Anderson and Pérez-Carballo, 233; *cf.* Barnum et al.; *cf.* Fidel 575). However, the former is far more costly than the latter, particularly as the volume of digitized materials grows. Furthermore, some of the functionality of the back-of-the-book index is covered by standard entity markup, which is frequently combined with controlled-vocabulary markup associated with the particular domain of the resource. As John Walsh has argued of the image vocabulary employed by the Blake Archive, such controlled-vocabulary markup functions very much like an index (Walsh).

General usability studies do not, however, get at the value of the intellectual or subject index to the scholarly or expert user. Indeed, the Bureau of National Affairs, a Virginia, USA publisher of highly specialized news in such areas as law, employment, the environment, and health care, conducted a usability study at law schools comparing text searching and index-aided research, and found that index users had an 86 percent success rate while text searchers had only a 23 percent success rate, particularly for tasks that departed from specific facts ("Using Online Indexes"). This result suggests the power of intellectual indexing and the potential need for such indexing within online scholarly resources.

To evaluate the role of the intellectual index and its relationship to more common forms of indexing, via markup, in digital humanities publishing, we combined the two in an online edition within a new version of Dynamic Table of Contexts (DToC) interface to publish an online version of *Canadian Women Writers: Connecting Texts and Generations*, edited by Marie Carrière and Patricia Demers, in partnership with the University of Alberta Press and in conjunction with their print publication. The online version incorporates the same extensive intellectual index as the printed version, along with semantic markup for named entities and other recurrent features of the text, some of which (such as named entities) overlap with index terms. To accommodate the index, the interface has been revised

to incorporate a specialized panel for the index terms that operates as an alternative to the panel for tags; users can use the two panels to see both types of term embedded together in the table of contents. It incorporates free-text search, coinciding with the ASI proposal to EPub that user interfaces combine search and index functionality (Wright et al., "DTTF proposal to EPub," 2, 7), and draws on the utility of the concordance in the provision of KWIC-like snippets for index terms and tags as well as keyword searches.

This paper describes the markup strategy used to encode the collection, including the index, in TEI; summarizes the team's revisions to the interface; and demonstrates the function of the intellectual index within the interface in relation to table of contents, the markup of named entities, and other features of the text. The Dynamic Table of Contexts Browser retains the traditional content of the intellectual index prepared for the back of the print edition, but dramatically reorients its location in the digital interface by placing it at the "front" or within the persistent navigation features of the reading interface. We will present some preliminary results of combining markup with an intellectual index by reporting on an initial user study undertaken with scholars from the Canadian Writing Research Collaboratory community from which the collection emerged, and suggest future directions for investigating the function of the intellectual index within digital scholarly editions.

References

Anderson, J. D., and J. Pérez-Carballo. (2001). The Nature of Indexing: How Humans and Machines Analyze Messages and Texts for Retrieval. Part 1: Research, and the Nature of Human Indexing. *Information Processing and Management*. 37: 231-254.

Barnum, C., et al. (2004). Index Versus Full-text Search: A Usability Study of User Preference and Performance. *Technical Communication*. 51(2): 187-206.

Brown, S., S. Fisher, P. Clements, K. Binhammer, T. Butler, K. Carter, I. Grundy, and S. Hockey. (1998). SGML and the Orlando Project: Descriptive Markup for an Electronic History of Women's Writing. *Computers and the Humanities*. 31: 271-85.

Burnard, L. Technical Documentation. *TEI P5: Guidelines for Electronic Text* http://www.tei-c.org/Guidelines/Customization/Lite/U5-techdoc.html

Burton, D. (1981). Automated Concordances and Word Indexes: The Fifties. *CHum*. 15(1): 1-14.

Butler, T. J., S. Fisher, G. Coulombe, P. Clements, I. Grundy, S. Brown, J. Wood, and R. Cameron. (2000). Can a Team Tag Consistently? Experiences on the Orlando Project. *Markup Languages: Theory & Practice* 2(2): 111-125.

Digital Trends Task Force. *American Society for Indexing*. American Society for Indexing, http://www.asindexing.org/i4a/pages/index.cfm?pageid=3647 (accessed 15 October 2012).

Fidel, R. (1994). User-Centered Indexing. *Journal of the American Society for Information Science*. 45(8): 572-576.

Garret, J. (2006). KWIC and Dirty? Human Cognition and the Claims of Full-Text Searching. *Journal of Electronic Publishing*. 9(1).

OED Online. (2013). "Index, n." Def. n. 1, 2a, 4b, 5b, 5d. Oxford University Press. http://www.oed.com/view/Entry/94372?rskey=ZKOmMc&result=1&isAdvanced=false (accessed 28 October 2012).

Jansen, L., H. Luijten, and N. Bakker. (eds). (2010). *Vincent van Gogh — The Letters*. Amsterdam & The Hague: Van Gogh Museum & Huygens ING. http://vangoghletters.org (accessed 28 October 2012).

Lamb, J. (2011). Kindle and the Index. … *Turning it Off and On*. Wordpress.com. http://ccgi.jalamb.com/2011/05/kindle-and-the-index/ (accessed 1 May 2011).

MacGlashan, M. (ed). (2012). *The Indexer*. **30.1** Sheffield, UK: The Society of Indexers.

Moving to XML Workflow. *WordCo Indexing Services Inc.* WordCo, http://www.wordco.com/ebook/xml.shtml (accessed 25 October 2012).

Nelson, B., S. Ruecker, M. Radzikowska, S. Sinclair, S. Brown, M. Bieber, and the INKE Research Group. (2011). A Short History and Demonstration of the Dynamic Table of Contexts. *Research Foundations for Understanding Books and Reading in a Digital Age: Text and Beyond*. held 18 November 2011 at Ritsumeikan University. Kyoto: Japan.

Ruecker, S., S. Brown, M. Radzikowska, S. Sinclair, T.M. Nelson, P. Clements, I. Grundy, S. Balasz, and J. Antoniuk. (2009). The Table of Contexts: A Dynamic Browsing Tool for Digitally Encoded Texts. In Dolezalova, L. (ed). *The Charm of a List: From the Sumerians to Computerised Data Processing.* Cambridge: Cambridge Scholars Publishing. 177-187.

Svenonius, E. (1986). Unanswered Questions in the Design of Controlled Vocabularies. *Journal of the American Society for Information Science*. 37(5): 331-340.

TEI Consortium. (eds). (2012). 3 Elements Available in all TEI Documents. In *TEI P5: Guidelines for Electronic Text Encoding and Interchange*. 2.3.0. TEI Consortium. held 17 January 2013. http://www.tei-c.org/release/doc/tei-p5-doc/en/html/CO.html#CONOIX (accessed October 28, 2012).

TEI Consortium. (eds). TEI element index (index entry). *TEI P5: Guidelines for Electronic Text Encoding and Interchange*. 2.3.0. TEI Consortium, 17 Jan. 2013. Web. 28 Oct. 2012. http://www.tei-c.org/release/doc/tei-p5-doc/en/html/ref-index.html

The Bureau of National Affairs, Inc. (2008). Using Online Indexes. *BNA Law School Education Series*. http://www.levtechinc.com/pdf/Using%20BNA%20Indexes%20study.pdf (28 October 2012).

Van Vliet, H. T. M., and A. Kets-Vree (2000). Scholarly Editing in the Netherlands. Literary and Linguistic Computing. 15(1): 65-72.

Walsh, J. A. (2008). Multimedia and Multitasking: A Survey of Digital Resources for Nineteenth-Century Literary Studies. In Siemens, R., and S. Schreibman (eds.), *A Companion to Digital Literary Studies*. Oxford: Blackwell.

The Models Editions Partnership Consortium. (1996). *Why Markup is Important. A Prospectus for Electronic Historical Editions.* http://wyatt.elasticbeanstalk.com/mep/misc/prospectus.html#Section 4 (accessed 12 March 2013).

Wright, J. (2012). The devil is in the details: indexes versus Amazon's X-Ray. *The Indexer*. 30(1): 11-16.

Wright, J., D. Ream, and the Digital Trends Task Force (DTTF). (2012). DTTF proposal to EPub 3.0 Working Group. *American Society for Indexing's Digital Trends Task Force (DTTF)*. American Society for Indexing, http://www.asindexing.org/files/DTTF/DPFIndex_functionality_in_ePub.pdf (29 October 2012).

Notes

1. "Controlled vocabulary terms are variously called index terms, descriptors, subject headings and, somewhat erroneously and increasingly ambiguously, keywords. Terms not belonging to a controlled vocabulary are called free-text terms, natural language terms and, again, keywords. Free-text is also used as an adjective to describe a type of searching, viz., searching that can be performed without the constraint of having to translate one's own vocabulary into the vocabulary used by a particular system" (Svenonius, 332).

2. Blogger John Lamb notes that when the Amazon Kindle was released in 2007 it did not support an index. When publishers created an eBook for it they "often excluded the index, even when it existed in the print version" (Lamb). In November 2011, Amazon released a new search tool to replace the index, called the Kindle X-Ray: "a new search and information feature that allows you to find information about characters, events, and topics in books" (Wright, "Amazon", 11).

3. The DTTF's "proposal moved forward quickly and an Indexes Charter document was published for a vote. The IDPF approved the formation of the EPUB 3.0 Indexes Working Group in December 2011 […] When completed it will be added as a modular update" (Digital Trends Task Force).

4. These include the plurality of the encoders involved and the fact that the indexical terms were developed in advance and applied to multiple texts (Butler et al.). The plurality of the encoders also suggests an aspect of the potential of digital systems for novel forms of indexing that is beyond the scope of this paper: the crowdsourcing of keywords, and perhaps even relations among them, as a basis for a folksonomic navigational aid.

5. It is telling that one of the more extensive discussions of how to encode indexes in the TEI documentation relates to the encoding of "existing 'pre-electronic' documents" (Burnard). A section called "Why Markup is Important" from *A Prospectus for Electronic Historical Editions*, a methodological framework compiled by the Steering Committee of the Model Editions Partnership, indicates the extent to which early thinking about TEI considered indexing an intrinsic function of markup ("Why Markup is Important").

6. The Rossetti Archive and Orlando, for example, provide quite sophisticated search capabilities including Boolean functions or ones that draw on sub-elements or attributes in the markup. However, digital humanities projects seldom include anything resembling the intellectual index with its carefully organized hierarchy of terms systematically synthesizing the contents of the book. The exemplary edition of Vincent Van Gogh's letters project offers a case in point. Whereas the 6-volume hardback edition includes a "full index," the online edition provides access to the letters by period, correspondent, place, or other features that could easily be flagged via structural and entity markup (Jansen et al.). Portions of the scholarly apparatus are available via a table of contents or through the same searches (Van Vliet and Kets-Vree).

7. It is noteworthy that even in the embattled context of print scholarship, the intellectual index seems to be losing ground: indexes are often absent from scholarly collections of essays altogether, or limited to named entity listings.

Developing a virtual research environment for scholarly editing. Arthur Schnitzler: Digitale Historisch-Kritische Edition

Buedenbender, Stefan
bued2101@uni-trier.de
Universität Trier, Germany

Burch, Thomas
burch@uni-trier.de
Universität Trier, Germany

Fink, Kristina
kfink@uni-wuppertal.de
Bergische Universität Wuppertal, Germany

Lukas, Wolfgang
wlukas@uni-wuppertal.de
Bergische Universität Wuppertal, Germany

Queens, Frank
queens@uni-trier.de
Universität Trier, Germany

Sirajzade, Joshgun
sirajzad@uni-trier.de
Universität Trier, Germany

Introduction

Arthur Schnitzler: Digitale historisch-kritische Edition. Werke 1904-1931 is a long-term project founded by the *Nordrhein-Westfälische Akademie der Wissenschaften* with a runtime of 18 years. The aim is to create a digital critical edition of both the majority of works published during Schnitzler's lifetime and his literary estate, providing scholars with an up-to-date, philologically dependable textual basis.

The first project phase (2012-14) is primarily about conceptional groundwork and focuses on two texts, *Fräulein Else* and *Komödie der Verführung* which were chosen both for their literary value and their usefulness as a testbed. Editing them will go hand in hand with finalizing the editorial principles and establishing the digital workflow.

To this end, the team of editors at *Bergische Universität Wuppertal* closely cooperates with computer scientists at *Trier Center for Digital Humanities* who are developing a customized virtual research environment. Based upon a unified data model, it will support all editorial steps from entering the meta data through transcribing the textual witnesses to preparing the online publication and printouts. This paper shall outline the philological challenges of the project and show how the projected platform is to meet them.

Initial Situation and Objectives

For scholars working on Schnitzler's writings, the availability of a philologically well founded textual basis is currently a key desideratum. Arthur Schnitzler is probably the only German-language author of literary world rank of whom there are to this day neither a commented student edition nor a (historical) critical edition. The project seeks to remedy this situation by:

- Providing the first edition of most of the works published during Arthur Schnitzler‘s lifetime with an authentic, critically constituted text.
- Providing a genetic edition of the major part of the literary estate (belonging to both published and unpublished work).
- Making the whole edited material accessible by text-critical comments and contextualizing its literary-historical background

The underlying material is just as comprehensive as manifold. Beyond the printed work, there are extensive preliminary works and sketches, as well as numerous unpublished dramatic, narrative, lyrical, aphoristic and essayistic texts. In addition to this literary estate material (which is complemented by film scripts and screenplays) there is a sizable number of non-literary writings. These can be either scientific (e.g. medical) treatises, literary criticisms and other comments, self-criticisms and records about the history of his own works — or, on the other hand, “ego-documents”: autobiography, diaries, dream journaling, correspondences. Thus, the catalogue of the *Literaturarchiv Marbach* lists about 28,000 letters and the estate material archived at *Cambridge University Library* comprises approximately 40,000 sheets.

Given this constellation, the integral approach of the projected edition has to be emphasized: The edited material

in its diversity is to be united within a homogeneous central data pool, allowing for different types of output. Following the principle of single-source publishing, the projected public online edition could be completed in the future by an annotated student edition or even a critical edition.

Challenges in Scholarly Editing

The new edition of Schnitzler's — published as unpublished — oeuvre is rich in philological challenges. It presents itself as a dynamic yet coherent system of most diverse transformation processes, branchings and cross-links. The individual literary "work" proves to be an unreliable unit; Schnitzler used to work for several years or even decades on a text. In this process, not only the genre would occasionally change, but there also may occur genetic bifurcations (one draft becoming multiple works) or fusions (several drafts being merged into one work).The project thus chooses the text witness in its materiality as a basis, prioritizing the categories of topography, genesis and intertextuality. Hence, the individual witness has to be described not only in its semantics, but also in its physical form and its processuality (writing stages). Moreover, the cross-links and correspondences of sequences and fragments between text witnesses are of particular interest.

Literary Computing and Software Development

In terms of literary computing, two aspects are to be emphasized: Firstly, the various requirements in scholarly editing (meta data capture, transcription, annotation, linking text and facsimile, linking text sequences with each other, collation, preparation of indices and critical apparatus) can only be met with a series of appropriate tools.

Secondly, these tools must be coordinated. In practical terms, this means above all that they have to fit into a unified philological workflow. This brings to the fore a facet that is often considered of minor interest in digital editorial projects: the aspect of the data format.

It is essential to determine a data model which can serve as a basis for the whole philological process, possibly without laborious intermediate conversions. From meta data collection through transcription and annotation to topography, networking and publication, all editorial stages need to be supported. This is complicated all the more by a set of overlapping hierarchies:

Firstly, there is the documentary view (with reference to the material witness, i.e. the single sheet), secondly the semantic-textual view (the text in its philological structure), thirdly sequences of any size have to be linked. This is not a trivial constellation for the XML-format, and the project currently evaluates how different standards deal with this problem. The benchmark are both the guidelines of the TEI and their concrete implementation in current projects (such as the digital *Faust*-edition). However, some of the proposed strategies, such as a splitting up into multiple files with different markup or the breaking up of subsidiary hierarchy levels, may add great complexity to the editor's work. So we currently test if the workflow could alternatively build on a database system, out of which the current processing status can be exported or archived any time into an international, platform-independent standard.

Virtual Research Environment

The tools needed for the project are currently being developed and made interoperable at the *Trier Center for Digital Humanities*. As for the meta data management, an adapted module of the *Forschungsnetzwerk und Datenbanksystem* FuD is used which allows the physical description of textual witnesses and their often complex tradition via specially customized input masks. For some other tasks (see below), currently existing solutions will be evaluated and tested in terms of usability. However, the core functionality is to be covered largely by *Transcribo*, an entirely new transcription, topography and edition software.

Transcribo

Transcribo is developed in close collaboration between computer scientists and philologists of the project partners. The graphical user interface is centered around the digital facsimile, i.e. usually the scanned witness. Units of varying extent (e.g. words, lines or paragraphs) can be marked by means of a rectangle-or polygon-tool, and then be transcribed and annotated. Each image file is displayed twice: the original can be examined on the left while the view on the right is used as a work space, where the transcribed text is placed topographically exact on top of the slightly attenuated facsimile. Where the spatial arrangement does not match the textual word order, words can be combined into sequences to preserve the semantic relationships in the transcript. As a central feature, the program further allows to mark genetic and philologically relevant phenomena in each transcribed unit and add annotations to them. This is done by a context menu with a selection of project-specific options. These include so far different variants of corrections (such as immediate corrections or late corrections with single, double or multiple strikethroughs and overwritings), the marking of highlightings, uncertain readings and non-identifiable

graphs. This selection, however, can be extended and will be adjusted over the entire course of the project to meet the requirements of the underlying material.

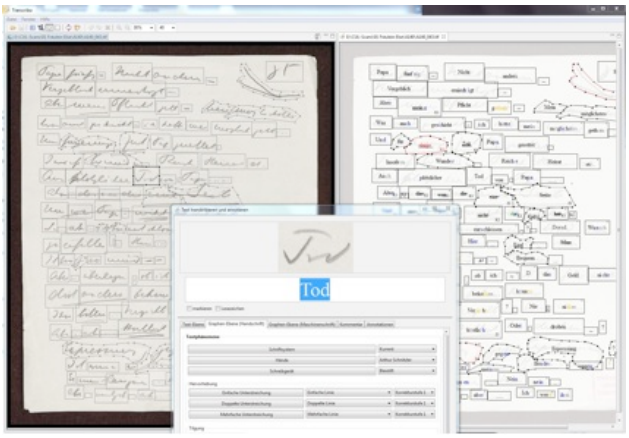


Fig. 1:
Transcription and annotation in Transcribo.

Further Extensions

FuD and *Transcribo* lay the basis for scholarly editing in an online environment by covering the key requirements: meta data management, transcription, topography description and basic annotations. This nucleus will be expanded over the next years to a virtual research environment that supports the entire project-specific workflow in a unified interface. This implies the integration of a number of additional features:

- Collation; here it has to be stressed that we will not only have to compare fixed texts, but also to identify similar sequences in a large data pool.
- The linking of freely definable sequences within or between text witnesses.
- The identification of dependencies between witnesses, in particular the creation of genetic paths.
- A publishing program for printouts.

While in the domain of publishing and collation, there are programs that can be adapted (we evaluate at present: XML-Print and CollateX, Juxta, TUSTEP), the remaining tools are to be developed. The result will be a research environment that — due to its platform-independent and modular design — can be adapted to meet the needs of numerous other editorial projects. Concrete interest in using this environment or individual modules of it has already been expressed by several projects such as:

- *August Wilhelm Schlegel "Kritische Ausgabe der Vorlesungen“ Band IV–VI*
- *Digitalisierung und elektronische Edition der Korrespondenz August Wilhelm Schlegels*
- *Wolfgang Koeppens „Jugend“*

References

Arthur Schnitzler: Digitale Historisch-Kritische Edition. http://www.buw-output.uni-wuppertal.de/ausgabe4/lukas/.

Trier Center for Digital Humanities. http://kompetenzzentrum.uni-trier.de.

Digitale Faustedition. https://faustedition.uni-wuerzburg.de/dev/project/about/.

TEI guidelines. http://www.tei-c.org/Guidelines/P5/.

FuD. http://fud.uni-trier.de/.

XML-Print. http://kompetenzzentrum.uni-trier.de/de/projekte/projekte/xml-print/.

Schlegel, A. W. *Kritische Ausgabe der Vorlesungen“ Band IV–V:I.* http://www.uni-tuebingen.de/fakultaeten/philosophische-fakultaet/fachbereiche/neuphilologie/deutsches-seminar/abteilungen/neuere-deutsche-literatur/mitarbeitende/lehrstuhl-braungart/projekte/august-wilhelm-schlegel.html.

Digitalisierung und elektronische Edition der Korrespondenz August Wilhelm Schlegels. http://www.slub-dresden.de/ueber-uns/projekte/erschliessung-und-digitalisierung/edition-schlegel/.

Koeppens, W. *„Jugend“*. http://www.phil.uni-greifswald.de/philologien/deutsch/forschung-kooperation/wka/dfg-projekt-jugend.html.

A national virtual laboratory for the humanities in Australia: the HuNI (Humanities Networked Infrastructure) project

Burrows, Toby Nicolas
toby.burrows@uwa.edu.au
University of Western Australia, Australia

Verhoeven, Deb

deb.verhoeven@deakin.edu.au
Deakin University, Australia

Summary

The Humanities Networked Infrastructure (HuNI) is a national "Virtual Laboratory" which is being developed as part of the Australian government's NeCTAR (National e-Research Collaboration Tools and Resources) programme. HuNI is using Linked Data to combine a range of different Australian datasets and deploying a suite of software tools to enable researchers to work with the data in various ways.

The HuNI project began in May 2012 and is scheduled to be completed at the end of 2013. It is being developed through a partnership between thirteen public institutions, led by Deakin University in Melbourne. At the Digital Humanities 2012 conference we presented a short paper on the overall design and proposed architecture for HuNI. We now propose a long paper reporting on HuNI's progress in its first twelve months and demonstrating the initial version of the Virtual Laboratory.

A national virtual laboratory for the humanities in Australia: the HuNI (Humanities Networked Infrastructure) project

The Humanities Networked Infrastructure (HuNI) is a national "Virtual Laboratory" which is being developed as part of the Australian government's NeCTAR (National e-Research Collaboration Tools and Resources) programme. The aims of NecTAR's Virtual Laboratories are to integrate existing capabilities (tools, data and resources), support data-centred research workflows, and build virtual research communities to address existing well-defined research problems. HuNI is addressing these aims across the whole of the humanities and creative arts, using Linked Data to combine a range of different Australian datasets and deploying a suite of software tools to enable researchers to work with the data in various ways.

The HuNI project began in May 2012 and is scheduled to be completed at the end of 2013. It is being developed through a partnership between thirteen public institutions, led by Deakin University in Melbourne. At the Digital Humanities 2012 conference we presented a short paper on the overall design and proposed architecture for HuNI (Burrows 2012). We now propose a long paper reporting on HuNI's progress in its first twelve months and demonstrating the initial version of the Virtual Laboratory. We plan to address the following issues.

Data harvesting from heterogeneous sources

The datasets being combined into HuNI come from a range of disciplines, including literature, performing arts, film and media, history, art and design. Some originate from an academic community, others from the curatorial sector. Various standard and customized metadata schemas are used, and software environments vary considerably. HuNI has tested and deployed several methods for harvesting data directly. Harvesting data indirectly through a third-party pre-aggregator has also been employed, though the use of the National Library of Australia's Trove service and especially its PeopleAustralia component (Dewhurst 2008).

Transformation to Linked Data

Transforming the harvested data into RDF triples is an essential first step in building HuNI's Linked Data store. This process has raised a variety of issues around data quality and validation, including spelling, misnaming, semantic ambiguities, incorrect coding and so on. Various methods have been used to address these issues, accompanied by considerable discussion around the respective roles of dataset custodians and HuNI itself in this process. We will also report on the use of tools like Google Refine for this kind of data cleaning (Van Hooland 2012).

Mapping and aligning vocabularies and ontologies

One of the key elements of HuNI is its semantic mapping and alignment service, which connects similar semantic entities in the different datasets — at the level of both classes and instances. A range of ontologies are being used in this mapping and alignment process, with CIDOC-CRM serving as the key ontology (Doerr, Hunter, Lagoze 2003; Gill 2004). Aspects of the Europeana Data Model have also been evaluated and incorporated (Isaac, Clayphan, Haslhofer 2012). User needs and requirements have been fed into the design of this process, in order to identify how extensive the mapping and alignment need to be before they start adding measurable value for researchers.

Strategies for software tools

The Australian datasets which have been combined into HuNI also offer a range of software tools for working with their data. These tools have mostly been developed

in-house, and are written in a variety of languages. They include LORE, AusStage, Heurist and OCCAMS (Bollen et al. 2009; Gerber, Hyland, Hunter 2010). HuNI has been assisting the enhancement of these tools to work with Linked Data, enabling users of these datasets to incorporate HuNI's data into their working environment.

HuNI also incorporates generic tools for working with Linked Data within the Virtual Laboratory itself, grouped around three main stages in the research workflow:

- Discovery (search and browse services);
- Analysis (annotation, collecting, visualization and mapping);
- Sharing (collaborating, publishing, citing and referencing).

Engaging and involving the research community

HuNI aims to reach and involve researchers from disciplines across the whole of the humanities and creative arts. It is critical that this broad research community is engaged effectively in the governance of the project as well as in the identification and implementation of user requirements and in order to address these aspirations a formal stakeholder management plan has been adopted. External validation is being provided through an international Expert Advisory Group. A range of communication channels have been established and used, ranging from "user story" requirements workshops to social media dissemination. An Agile software development framework has been adopted, enabling detailed and frequent user input. Considerable time was devoted to establishing these processes in the first stage of the project.

Project Governance

One of the key challenges for large-scale data interoperability projects such as HuNI is achieving common goals, problem definition and processes amongst participating organisations (Pagano 2010). HuNI involves 13 partner institutions including universities, development agencies and cultural institutions. Acknowledging that collaboration and data sharing involve various organisational complexities, a steering committee and several advisory groups have been set up as part of the formal governance structure for HuNI.

Summary

The paper will discuss HuNI's approach to these issues and examine the lessons learnt from the project to date. We will compare HuNI's approach with that taken by other digital environments for the humanities, particularly those involving the aggregation of digital resources and the creation of collaborative environments (Svensson 2010). We will assess HuNI's significance as a new model for integrating and sharing knowledge and enabling research in the humanities and creative arts in the future. Particular focus will be given to the way in which HuNI embodies a theoretical framework in which semantic entities and assertions about their relationships, expressed as Linked Data, serve as the fundamental building-blocks. Instead of digital objects, primary sources or metadata records, these semantic assertions form the "data" on which a service like HuNI is built (Borgman 2007:215-217; Burrows 2011).

HuNI will be undertaking a formal evaluation process as part of the last stage of the project, during the last quarter of 2013. In this paper, we will be able to present informal feedback to date from researchers and other users on the pilot version of the service.

References

Bollen, J., N. Harvey, J. Holledge, and G. McGillivray. (2009) AusStage: e-Research in the performing arts, *Australasian Drama Studies*. 54: 178-194.

Borgman, C. L. (2007). *Scholarship in the Digital Age: Information, Infrastructure, and the Internet*. Cambridge, MA: MIT Press.

Burrows, T. (2012). Designing a national 'Virtual Laboratory' for the humanities: the Australian HuNI project. In Meyer, J. C. (ed), *Digital Humanities 2012: Conference Abstracts*. held July 16-22 at University of Hamburg. 139-141

Burrows, T. (2011). Sharing humanities data for e-research: conceptual and technical issues. In Thieberger, N. et al. (eds), *Sustainable Data from Digital Research*. Melbourne: PARADISEC, 177-192. http://hdl.handle.net/2123/7938

Dewhurst, B. (2008). People Australia: a Topic-Based Approach to Resource Discovery. *VALA2008 Conference proceedings*. held in Melbourne: VALA. [http://www.valaconf.org.au/vala2008/papers2008/116_Dewhurst_Final.pdf]

Doerr, M., J. Hunter, and C. Lagoze (2003). Towards a Core Ontology for Information Integration, *Journal of Digital information* 4.1 http://jodi.ecs.soton.ac.uk/Articles/v04/i01/Doerr/

Gerber, A., A. Hyland, and J. Hunter (2010). A collaborative scholarly annotation system for dynamic Web documents — a literary case study. In *The Role of Digital*

Libraries in a Time of Global Change (Lecture Notes in Computer Science 6102). Berlin: Springer-Verlag. 29-39.

Gill, T. (2004) Building semantic bridges between museums, libraries and archives: The CIDOC Conceptual Reference Model, *First Monday* 9 (5). http://www.firstmonday.org/htbin/cgiwrap/bin/ojs/index.php/fm/article/viewArticle/1145/1065

Isaac, A., R. Clayphan, and B. Haslhofer (2012). Europeana: moving to Linked Open Data, Information Standards Quarterly. 24.2-3. 34-40.

Pagano, P. (2010). Data interoperability. In: **GRDI2020 Consortium.** *Technological and Organisational Aspects of a Global Research Data Infrastructure: a View from the Experts.* 25-33. http://www.grdi2020.eu/Repository/FileScaricati/9a85ca56-c548-47e4-8b0e-86c3534ad21d.pdf

Svensson, P. (2010). The Landscape of Digital Humanities, *DHQ: Digital Humanities Quarterly* 4.1 http://digitalhumanities.org/dhq/vol/4/1/000080/000080.html

Van Hooland, S., R. Verborgh, and R. Van de Walle (2012). Joining the Linked Data Cloud in a cost-effective manner. *Information Standards Quarterly*. 24.2-3. 24-28.

Bindings of Uncertainty. Visualizing Uncertain and Imprecise Data in Automatically Generated Bookbinding Structure Diagrams

Campagnolo, Alberto
a.campagnolo1@arts.ac.uk
Ligatus Research Centre, University of the Arts, London, UK

Velios, Athanasios
a.velios@arts.ac.uk
Ligatus Research Centre, University of the Arts, London, UK

1. Introduction

'If it is drawn it must be true.' Most people have the tendency to treat visualized data as facts and are much less prone to question visualizations than written words. Incorporating ways of conveying the notions of uncertainty and imprecision within visualizations is therefore a critical issue. This presentation will address this problem and consider what possible steps can be taken to visualize uncertainty and imprecision within the constraints of diagrams whose shapes and appearance are necessarily dictated by those of the objects that they represent.

Amongst the disciplines devoted to the study of books, bookbinding history is young, lacking a well-established vocabulary and a stable description system (Szirmai 1999). Bookbinding descriptions are often inaccurate, and difficult to communicate especially to international audiences.

Verbal descriptions are inadequate and are usually accompanied by drawings, which do not comply with any standard. Different authors employ different styles and conventions, and sometimes a number of conventions are used within one publication.

Recent work at the Ligatus Research Centre addresses the bookbinding vocabulary and classification problem. We have developed a descriptive schema and glossary for bookbinding structures utilizing eXtensible Markup Language (XML) technologies. This has been used to survey the bookbinding structures of the books from the St. Catherine's Monastery Library, Mount Sinai, Egypt. The records are combined with freehand drawings of the binding structures (Velios 2008; Velios and Pickwoad 2005). More recently, we have been working towards a methodology to automatically transform the XML descriptions of bookbinding structures into Scalable Vector Graphics (SVG) diagrams. The advantages of the automated transformations are: (i) standardized output; (ii) production speed as they save significant time during the survey, and (iii) better accuracy as they function as verification of the surveyed data during the survey. However, the data produced during the survey is often uncertain (when features are obscured), imprecise (when surveyors have limited capacity to interpret evidence), and incomplete (when the available time is limited). Uncertain data is so flagged within the XML records, and also the imprecision and the absence of certain elements can be inferred from the records. Therefore, it is possible and appropriate to communicate imprecise and uncertain data within the visualizations. These should be drawn clearly and without affecting the overall meaning of the diagram.

2. Uncertainty and Imprecision of Data

When dealing with historical bookbinding structures, one has to accept that a degree of uncertainty, imprecision and incompleteness is inevitable for the following reasons.

2.1 Uncertainty

As in the case of quantitative and geographical data visualizations (Wilkinson 2005; Wainer 2009; Gethin Powell 2012), factors of uncertainty can be found in (i) inherent problems with the definition of the object of study, i.e. the limitations of current knowledge regarding bookbinding structures and their description; (ii) the sources of information and their interpretation, i.e. the books being described and the interpretation of their binding structures; (iii) issues with the categorization and representation of certain complex features, i.e. undetermined complex shapes for highly decorative and non-functional elements. Uncertainty can arise in all of these areas. One of the main problems faced by bookbinding historians lays in the young age of the discipline. Goldschmidt, in 1928, pointed out that 'far fewer people [could] give a reasoned opinion on the country of origin and the approximate date of an old bookbinding, than a piece of pottery or furniture' (Goldschmidt 1928). Unfortunately, almost a century later, the situation has not changed and often bookbinding descriptions are still inaccurate. Moreover, a book is sometimes too damaged to show clear evidence of its origin, or else, elements of the binding, especially decorative ones like metal furniture, are so complex that it is difficult to categorize and describe them.

2.2 Imprecision

The automated visualizations of binding structures, based upon their XML verbal descriptions, do not aim at a naturalistic representation of the item being described. But rather at conveying its general form in a highly prototypified representation, thus making it possible and easier for it to be compared with similar structures found in other books. Similarity of shapes depends on the preservation of certain particular features, referred to as *prägnant* features, e.g. symmetry, orientation in space, etc., while *non-prägnant* features, e.g. precise measurements, can be substantially changed without psychologically affecting the overall impression for the observer (Goldmeier 1972). Precise measurements are therefore not essential for the prototypification of shape, and for this reason only few precise measurements are required from the bookbinding record. Nonetheless, we feel that imprecisions in measurement-dependent elements should still be identifiable because they indicate scale, which may be a critical aspect of the description.

2.3 Human Error

When inaccuracy is objectively determinable, it can be expressed as error (MacEachren, et al. 2005). As it is often the case with human generated data, input problems and errors do occur. However, these are not easily identifiable through automated means and, considering that our transformations are also thought of as a possible step in ensuring data correctness by part of the surveyors, ambiguities due to human error are not being considered here.

3. Representation of Uncertainty and Imprecision

Uncertainty and its representation have been studied in the scientific and geographical visualization communities (Wilkinson 2005). Bertin's (Bertin 1981; Bertin 1983) basic graphic variables — location/position, size, value, texture, colour, orientation, shape/form — with the addition of colour saturation, transparency/opacity, and sharpness/blurring (MacEachren 1992) have been used to build graphic representations and depict uncertainty. Interestingly, all these graphical features correspond to the basic feature channels in our primary visual cortex (V1), being thus perceptually distinct features (Ware 2012).

Cultural heritage visualizations are constrained by the shapes, spatial relationships, and relative sizes of the objects being portrayed. Therefore, only a limited set of graphic variables can be used to depict uncertainty. While uncertainty depiction needs to be perceptually salient, understanding the overall meaning of the diagram should not be precluded. Transparency has been used in archaeological reconstructions (Zuk and Carpendale 2006); however, lacking a distinct background, the use of transparency on its own can be problematic. Further, automated visualizations need a flexible system to apply a higher or lower degree of uncertainty to its elements.

3.1 Uncertainty

Out of the possible graphic variables, we have identified blurring as the best option to express uncertainty in our diagrams. People can easily interpret less sharp details as uncertainty (MacEachren 1992), while blurring can be applied to single lines without affecting the diagram's overall appearance. SVG provides a Gaussian-blur filter applicable to any shape and allows for different standard deviation values for the *x* and *y* axes. This grants the preservation of the overall shape of uncertain elements (see fig. 1) and permits to visualize different degrees of uncertainty through a flexible and universally applicable system. Uncertain elements are drawn in the configuration

deemed most probable, while the degree of blurring can be determined by the number of possible variations.

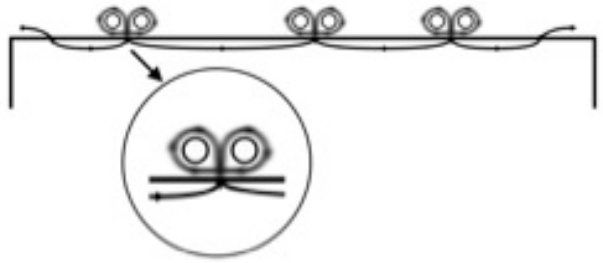


Figure 1
Example of uncertainty of a sewing pattern

Once identified, uncertainty can be resolved by providing the user with a set of 'small multiples' (Tufte 1990), i.e. alternative possible visualizations, allowing for immediate comparison within the scope of the eye span (see fig. 2).

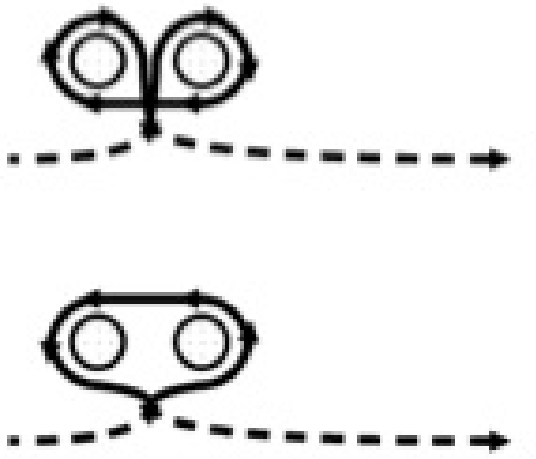


Figure 2
Example of small multiples for a sewing pattern

3.2 Imprecision

To indicate imprecision we need a graphical variable that could make the element identifiable without affecting the overall outline of the diagram, thus allowing for uninterrupted contour realization by part of the user. While colour labelling could be an option, we need a V1 feature channel that could make the element perceptually distinct leaving the contour unaffected. We find that this could be achieved with a surround colour (see fig. 3).

Figure 3
Figure 3 Examples of a board thickness diagram with and without imprecision highlight

A coloured halo is a perceptual distinction feature (Ware 2012) that can be universally applied to graphics through SVG filters. The reduced contrast with the background makes the feature perceptually less salient, while the shape contour is left unaltered.

4. Conclusions

Visualizations within scholarly research projects should convey uncertainty and imprecision where needed. Depicting uncertainty has been the object of study of many scholars, however, its application to diagrams whose shapes are dictated by those of the objects being represented, like in the case of cultural heritage objects, poses particular issues and constraints. Even more so in the case of automatically generated diagrams within a large-scale project as our bookbinding structure visualizations.

This presentation is intended to serve as an example of the kind of perceptually salient graphical variables that can be used to successfully depict uncertainty and imprecision within cultural heritage prototypical diagrams.

References

Bertin, J. (1981). *Graphics and Graphic Information Processing.* Berlin: Walter de Gruyter & Co.

Bertin, J. (1983). *Semiology of Graphics.* Madison: University of Wisconsin Press.

Gethin Powell, R. (2012). Uncertain Date, Uncertain Place: Interpreting the History of Jewish Communities in the Byzantine Empire using GIS. In *Digital Humanities 2012 Conference Abstracts.* Hamburg: Hamburg University Press. 329-331.

Goldmeier, E. (1972). *Similarity in Visually Perceived Forms.* Madison, CT: International Universities Press.

Goldschmidt, E. P. (1928). *Gothic & Renaissance Bookbindings, Exemplified and Illustrated from the Author's Collection.* London: Ernest Benn, Houghton Mifflin Co.

MacEachren, A. M. (1992). Visualizing Uncertain Information. *Cartographic Perspective.* 13: 10-19.

MacEachren, A. M., et al. (2005). Geospatial Information Uncertainty, What We Know and What We Need to Know. *Cartography and Geographic Information Science* 32(3): 139-160.

Szirmai, J. A. (1999). *The Archaeology of Medieval Bookbinding.* Brookfield, VT: Ashgate Publishing.

Tufte, E. R. (1990). *Envisioning Information.* Cheshire, CT: Graphic Press.

Velios, A. (2008). Hierarchical Recording of Binding Structures. *The Book and Paper Group Annual* 27: 97.

Velios, A., and N. Pickwoad. (2005). Current Use and Future Development of the Database of the St. Catherine's Library Conservation Project. *The Paper Conservator* 29: 39-53.

Wainer, H. (2009). *Picturing the Uncertain World, How to Understand, Communicate, and Control Uncertainty through Graphical Display.* Princeton, NJ: Princeton University Press.

Ware, C. (2012). *Information Visualization, Perception for Design.* 3rd edn. Waltham, MA: Morgan Kaufmann.

Wilkinson, L. (2005). *The Grammar of Graphics.* 2nd edn. New York: Springer.

Zuk, T., and S. Carpendale (2006). Theoretical Analysis of Uncertainty Visualizations. in *Visualization and Data Analysis. Proceedings of SPIE-IS&T Electronic Imaging 2006;* SPIE. held in Bellingham, WA. 66-79.

Versioning Texts and Concepts

Carter, Daniel
carter.daniel.w@gmail.com
University of Texas at Austin

Ross, Stephen
saross@uvic.ca
University of Victoria, Canada

Sayers, Jentery
jentery@uvic.ca
University of Victoria, Canada

Schreibman, Susan
schreibs@tcd.ie
Trinity College Dublin

Introduction: This paper presents the results of the Modernist Versions Project's (MVP) survey of existing tools for digital collation, comparison, and versioning. The MVP's primary mission is to enable interpretations of modernist texts that are difficult without computational approaches. We understand versioning as the process through which scholars examine multiple witnesses of a text in order to gain critical insight into its creation and transmission[1] and then make those witnesses available for critical engagement with the work. Collation is the act of identifying different versions of a text and noting changes between texts. Versioning is the editorial practice of presenting a critical apparatus for those changes. To this end, the MVP requires tools that: (1) identify variants in TXT and XML files, (2) export those results in a format or formats conducive to visualization, (3) visualize them in ways that allow readers to identify critically meaningful variations, and (4) aid in the visual presentation of versions.

The MVP surveyed and assessed an array of tools created specifically for aiding scholars in collating texts, versioning them, and visualizing changes between them. These tools include: (1) JuxtaCommons, (2) DV Coll, (3) TEI Comparator, (4) Text::TEI::Collate, (5) Collate2, (6) TUSTEP, (7) TXSTEP, (8) CollateX, (9) SimpleTCT, (10) Versioning Machine, and (11) HRIT Tools (nMerge). We also examined version control systems such as Git and Subversion in order to better understand how they might inform our understanding of collation in textual scholarship. This paper presents the methodologies of the survey and assessment as well as the MVP's initial findings.

Problem: Part of the MVP's mandate is to find new ways of harnessing computers to find differences between witnesses and to then identify the differences that make a difference (Bateson). In modernist studies, the most famous example of computer-assisted collation is Hans Walter Gabler's use of the Tübingen System of Text Processing tools (TUSTEP) to collate and print James Joyce's *Ulysses* in the 1970s and 1980s. Yet some constraints, such as those identified by Wilhelm Ott in 2000, still remain in the field of textual scholarship, especially where collation and versioning applications are concerned. Ott writes, "scholars whose profession is not computing but a more or less specialized field in the humanities have to be provided with tools that allow them to solve any problems that occur without having to write programs themselves. This leads to a concept of relatively independent programs for the elementary functions of text handling" (97). Indeed the number of programs available for collation work have proliferated since 2000, including additions to TUSTEP (TXSTEP) as well as the newest web-based collation program, JuxtaCommons.

Accordingly, the MVP has reviewed tools currently available for collation work in order to provide an overview

of the field and to identify software that might be further developed in order to create a collating, versioning, and visualization environment. Most of these tools were developed for specific projects, and thus do what they were designed to do quite well. Our question is whether we can modify existing tools to fit the needs of our project or whether a suite of collation and visualization tools needs to be developed from scratch. This survey is thus an attempt to chart the tools that may be useful for the kinds of collation and versioning workflows our team is developing specifically for modernist studies, so we can then test methods based on previous tools and envision future developments to meet emerging needs. Our initial research with Versioning Machine and JuxtaCommons suggests that there is potential for bringing tools together to create a more robust versioning system. Tools such as the Versioning Machine work well if one is working with TEI P5 documents; however, we are equally interested in developing workflows that do not rely upon the TEI, or do not require substantive markup. Finally, we are examining whether the development of version control systems such as Git present viable alternatives to versioning methods now prevalent in textual studies.

Method: Our method adapts the rubric Hans Walter Gabler devised for surveying collation and versioning tools in his 2008 white paper, "Remarks on Collation." We first assessed the code and algorithms underlying each tool on our list, and we then tested each tool using a literary text. In this particular case, we used two text files and two TEI XML files from chapter three of Joseph Conrad's *Nostromo*, which we have in OCR-corrected and TEI-Lite marked-up states from the 1904 serial edition and the 1904 first book edition. During each test, we used a tool assessment rubric (available upon request) to maintain consistent results across each instance. All tests were accompanied by research logs for additional commentary and observations made by our research team.

Preliminary Findings: Our preliminary findings suggest that:

- Many existing collation tools are anchored in obsolete technologies (e.g., TUSTEP, which was originally written in Fortran, despite having undergone major upgrades, still relies on its own scripting language and front end to operate; also, DV-Coll was written for DOS, but has been updated for use with Windows 7).
- Many of the tools present accessibility obstacles because they are desktop-only entities, making large-scale collaborative work on shared materials difficult and prone to duplication and/or loss of work. Of the tools that offer web-based options, JuxtaCommons is the most robust.
- The "commons" approach to scholarly collaboration is among the most promising direction for future development. We suggest the metaphor of the commons is useful for tool development in versioning and collation as well as for building scholarly community (e.g., MLA Commons). We note the particular usefulness in this regard of the Juxta Commons collation tool and the Modernist Commons environment for preparing digital texts for processing. The latter, under development by Editing Modernism in Canada, is currently working to integrate collation and versioning functions into its environment.
- Version control alternatives to traditional textual studies-based versioning and visualization presents an exciting set of possibilities. Although the use of Git, Github, and Gist for collating, versioning, and visualizing literary texts has not gained much traction, we see great potential in this line of inquiry.
- Developers and projects should have APIs in mind when designing tools for agility and robustness across time. Web-based frameworks allow for this type of collaborative development, and we are pleased to see that Juxta has released a web service API for its users.
- During tool development, greater attention must be given to extensibility, interoperability, and flexibility of functionality. Because many projects are purposebuilt, they are often difficult to adapt to non-native corpora and divergent workflows.

References

Bateson, G. (1972). Steps to an Ecology of Mind. New York: Ballantine.

Gabler, H. (2013). Remarks on Collation. *Academia.edu*. Academia.edu, 2008. 14 Mar. 2013. http://www.academia.edu/167070/_Remarks_on_Collation_ .

Gabler, H. (2000). "Towards an Electronic Edition of James Joyce's Ulysses." Literary and Linguistic Computing 15(1): 115-20.

McGann, J. J. (1991). The Textual Condition. Princeton, NJ: Princeton University Press.

Ott, W. (2000). "Strategies and Tools for Textual Scholarship: The Tubingen System of Text Processing Programs (TUSTEP)." Literary and Linguistic Computing 15(1): 93-108.

Reiman, D. H. (1987). "Versioning." Romantic Texts and Contexts. Columbia: University of Missouri. 167-179.

Notes

1. For a definition of the “social” life of texts, see Jerome McGann’s *The Textual Condition*. For a definition of “versioning,” see Donald Reiman, who writes, “In those cases where the basic problem facing the scholar or reader involves two or more radically differing versions that exhibit quite distinct ideologies, aesthetic perspectives, or rhetorical strategies, the alternative to ‘editing,’ as conventionally understood, may be what I call ‘versioning’” (169).

Pure Transcriptional Markup

Caton, Paul
pncaton@gmail.com
King's College London, United Kingdom

Renear (2000) introduces the term "transcriptional markup" with a loose sense of pertaining to the reproduction of an existing text, and he does not attempt to define transcription *per se*. More recently Huitfeldt and Sperberg-McQueen (2008), and Huitfeldt, Marcoux, and Sperberg-McQueen (2009, 2010) have developed a model of transcription (hereafter the HMS model) that they believe "provides a sort of greatest common denominator for markup systems" (2010, 15). But an awkward gap exists between the HMS model's abstract components and the realities of, for example, the TEI markup language and its typical usage. Here I make the case for a kind of markup — perfectly possible though in practice improbable — which bridges that gap. I call it *pure transcriptional markup* because it refines the sense of Renear's term and grounds it in an actual, formally specified model of transcription.[1]

Transcription and text encoding are clearly connected, especially where markup replaces the work of presentation.[2] However, not everything that typical scholarly encoding practice might assert regarding the logical domain of a pre-existing text can be considered proper to transcription (if "transcription" is to have any specific meaning of its own). In the initial HMS model Huitfeldt and Sperberg-McQueen (2008) avoid the trap of trying to decide what visible details of an exemplar document E should be reproduced in a transcription T by asking instead what overall relation must hold between the two such that document T is a successful transcription of document E (expressed as *t_similar*). They say that under a set of reading conditions R, marks of E can be seen as a sequence of tokens each instantiating a type. So, abstractly, a document E is a sequence of types, and if under the same set of reading conditions R a document T can be seen as a token sequence whose corresponding type sequence matches E's type sequence, then E and T are t_similar.

The initial HMS model considers a simple sequence of basic, indivisible tokens at grapheme level; later versions (2009, 2010) introduce components that allow the modeling of a complex structure of token groups (and related types). These groups, known as compound tokens, may have as constituents either compound tokens or basic (atomic) tokens. The resulting view of a document comprising a *structured*, *multi-level* token sequence immediately suggests similarities with the well-known 'ordered hierarchy of content objects' (OHCO). As noted above, its creators explicitly connect the HMS model to markup, saying some aspects "serve purposes analogous to the generic identifiers and attribute-value pairs of SGML and related markup languages" (2010, 11), that "element types are types", and "element instances are tokens" (2010, 15).

However, some features of common encoding practice are problematic for the model. I briefly outline two of them here; a single example will illustrate both.[3] Suppose an original printed text contains the sentence "Joe *stinks*!" Describing this sentence in terms of the HMS model we have:

- 11 basic tokens at character level — 'J', 'o', 'e', ' ', '*s*', '*t*', '*i*', '*n*', '*k*', '*s*','!'
- 2 compound tokens at word level — 'Joe' and '*stinks*'
- 1 compound token at sentence level — 'Joe *stinks*!'

A typical TEI encoding of our example might be:

```
<persName>Joe</persName> <emph
   rend="italics">stinks</emph>!
```

For the first of the awkward features, recall that for HMS element types are types and element instances are tokens. In markup theory an instance of an element with #PCDATA content would be start tag + content + end tag. So `<persName>Joe</persName>` is a single element instance and hence a single token. But what type does it instantiate — type 'persName', or type 'Joe'? The problem here is that the single element instance appears to be two tokens, one wrapped inside another. Furthermore, they appear to be operating at different levels: as a specific lexical item, and as a characterization of the lexical item.

The second awkward feature can be seen in `<emph rend="italics">stinks</emph>`. If element types are types, then we must assume an `<emph>` element instantiates an 'emphasis' type. Yet there are strong grounds for arguing that emphasis is *not* a type in the sense used by the model. Wetzel (2009, xii-xiii), citing arguments from Wollheim 1968, believes that while types may well be

considered universals, they differ from other universals such as properties. Speaking of words as types she says "they are *objects* according to the common sense and scientific theories we have about them—values of the first-order variables and referents of singular terms—rather than properties" (124). The emphasis associated with the token "stinks" by the use of italic typeface is at the type level a property associated with the word type 'stinks', not a type itself. Viewing emphasis as a property does accord with the model, which allows for relations between properties and types.

Commonly in encoding schemes there are elements that certainly seem to be associated with types as per the HMS model; but there are also elements we would associate with properties, a point noted by Dubin (2003). The TEI scheme has many type-like elements, but no constraining principle that elements should be restricted to Peirceian types. Rather the 'targets' seem to be textual *features* — that may or may not be objects, may or may not be visible properties. Markup is directed at organizational features made *manifest* by the work of presentation (paragraphs, lists, etc.), and non-organizational features that stand out in the regular text flow by virtue of either visible difference or semantic nature.

The challenge, then, is firstly to describe a general view of encoding that does fit the HMS model, and secondly to see if this view helps account for actual encoding in terms of the model. We noted that in the HMS model transcription is defined by the mediation of a type sequence between E and T. The conceptual movement is not E_token_structure -> T_token_structure but rather E_token_structure -> E_type_structure -> T_token_structure.

The middle part of this progression — establishing the type sequence under a reading — may happen in the transcriber's head, but it is the core of the model nevertheless. It is also fundamental to the HMS model that at any one level, a token instantiates a single type. From these givens, some criteria for pure transcriptional encoding emerge.

* if elements are tokens they must instantiate a single type. We therefore move #PCDATA content into attribute values and make all basic-level elements empty. Our example "Joe *stinks*!" might be represented as follows:

```
<sentence> <word designation="persName">
<character type="j" form="majuscule"/
> <character type="o"/> <character
type="e"/> </word> <whitespace/> <word
communicative_intent="emphasis"> <character
type="s"/> <character type="t"/> <character
type="i"/> <character type="n"/> <character
type="k"/> <character type="s"/> </word>
<punctuation type="exclamation"/> </sentence>
```

* elements, attributes, and element structure must either supply overtly or make available through logical inference the information necessary for t_similarity; we must therefore view pure transcriptional encoding as applying to E_type_structure rather than to E_token_structure.[4] * under a reading R, at whatever level a token is considered basic, so must an element at that level be considered basic and therefore atomic; ie. if character level is considered basic, then "e" and "`<character type='e'/>`" are equally indivisible. * the status of elements as tokens must be somewhat unusual. Wetzel (2012), following Wollheim (1968), notes that types usually resemble their tokens. Even if we allow for minor differences of form (eg. token "E" instantiating type 'e'), we cannot claim "`<character type='e'/>`" resembles 'e'. Pure transcriptional markup represents an intermediary stage not normally intended for human consumption — in effect a kind of *precipitation out* of the E_type_structure into something half-way towards a normal token sequence.

The elements of pure transcriptional markup are analogous to the posited sentences that Zellig Harris in his mathematical theory of grammar says "go beyond what is normally said in English and are characterized as grammatically possible rather than as actual sentences" (1982, p. 15, my emphasis). The crucial point for Harris is that these sentences, while never encountered 'in the wild', can through a series of rule-governed transformations be reduced to — and so account for — the sentence forms we *do* encounter in everyday English. These operations "have the common property . . . of reducing high-likelihood, low information entries [or words into sentences]" (p. 8). I suggest a similar relation holds with respect to the difference between pure transcriptional markup and everyday markup, and that we could describe a series of non-random transformations that would reduce the high-likelihood, low-information encoding such as we see in the example above to the kind of markup we normally encounter. In digital humanities we commonly 'zero out' most transcriptional encoding leaving a majority of tokens in #PCDATA form - so "`<character type='e'/>`" becomes just "e". Where we want to retain information about a type property without using a specific form of #PCDATA token, we apply transformations that zero out the *nature* of the type (ie. we remove the low-information markup that says "stinks" is a word) leaving just the *property* to be expressed by the markup — hence TEI's `<emph>`.

References

Caton, Paul (2004). *Text Encoding, Theory, and English: A Critical Relation*. Ph.D dissertation. Brown University.

Caton, Paul (2009). Lost in Transcription: Types, Tokens, and Modality in Document Representation. In *Digital Humanities* held June 2009 at College Park, University of Maryland.

Dubin, D. (2003). Object mapping for markup semantics. In Usdin, B. T. (ed). *Proceedings of Extreme Markup Languages.* Montreal, Quebec.

Harris, Z. (1982). *A Grammar of English on Mathematical Principles*. New York: Wiley-Interscience.

Huifeldt, C., and C. M. Sperberg-McQueen (2008). What is transcription? *Literary and Linguistic Computing.* 23 (3). 295-310. doi:10.1093/llc/fqn013

Huitfeldt, C., Y. Marcoux, and C. M. Sperberg-McQueen (2009). "What is transcription? (Part 2)." In *Digital Humanities* held June 2009 at College Park, University of Maryland.

Huitfeldt, Claus, Yves Marcoux and C. M. Sperberg-McQueen (2010). "Extension of the type/token distinction to document structure." In *Balisage: The Markup Conference 2010*. held August 3-6, 2010 in Montréal, Canada. In *Proceedings of Balisage: The Markup Conference 2010. Balisage Series on Markup Technologies*.5. doi:10.4242/BalisageVol5.Huitfeldt01.

Renear, A. (2000). The descriptive/procedural distinction is flawed. *Markup Languages* 2 (4). 411-420.

Renear, A., and D. Dubin (2003). Towards identity conditions for digital documents. In S. Sutton (ed.) *Proceedings of the 2003 Dublin Core Conference*. University of Washington, Seattle, WA.

Wetzel, L. (2009). *Types and Tokens: On Abstract Objects*. Cambridge, MA.: MIT Press.

Wollheim, R. (1968). *Art and Its Objects*. New York: Harper and Row.

Notes

1. This paper presents work from a much larger, ongoing project prompted by ideas first presented in Caton 2009. I no longer defend the conclusions of that earlier work, and the current project represents a substantial rethinking and development of my initial ideas.
2. On the work of presentation see Caton 2004. Another part of the larger project from which this work is drawn examines in detail how presentation mediates transcription and encoding.
3. Necessarily, the outline given here is greatly condensed from an extensive discussion in the larger project, with some consequent loss of continuity and supporting evidence from the argument.
4. Note that we would still have to associate a semantics with the pure transcriptional markup if we wanted to establish formal identity between documents in the manner suggested by Renear and Dubin (2003).

A New Ecological Model for Learning

Cenkl, Pavel Thomas
pcenkl@sterlingcollege.edu
Sterling College, United States of America

This paper introduces a new ecology of learning and innovative connections between ecological and a humanities curriculum. Drawing on points of intersection between experiential liberal arts education, digital humanities, biomimicry, and ecopsychology, this session will engage instructors and administrators in course development strategies and in helping students plan their own learning by using a DH-supported systems approach to curriculum design.

The presentation will take as its case study Sterling College in Craftsbury Common, Vermont, the smallest four-year residential liberal arts college in the United States, and a community-centered, ecologically focused institution built upon a foundation of experiential education.

The College empowers the development learning community in which students and faculty engage in meaningful experiences each day as part of an integrative environmental liberal arts curriculum. The experiential scholarship that undergirds the College's curriculum enables countless moments of engagement between student, faculty, and place, but it is only with effective reflection that these moments coalesce and find their way into the larger dialogues and stories that make up the fabric of our community.

The presentation will explore the following areas:

Community

Concept

From individual experience to community narrative.

Application

Community learning, broadly, is enabled by tools that (1) create collaborative spaces and empower collaboration

and (2) track 'wear' by layering metadata for administrators and faculty to shape and innovate teaching strategies, which further supports migration from instructor-based to co-creative learning.

Discussion

We are currently beginning to explore how technology can help to facilitate reflection on experience to both create a community timeline of layered narratives and help to support learning experiences campus-wide by sharing these boundary objects — or boundary events — and extending discourse across the College to support our integrative curricular model. The resultant archival wear can contribute to metadata that, itself, becomes the community story and creates a feedback loop to help faculty implement more effective classroom experiences.

Work

Concept

Learners are more active and engaged when participating in meaningful and consequential work as both an individual and shared experience.

Application

Work — whether remote or local — can be supported by mobile, accessible text and image sharing to serve as both individual archive and group collaboration can bridge between work experience and 'traditional' classroom scholarship.

Discussion

> "Can there be any greater reproach than an idle learning?
>
> Learn to split wood, at least.
>
> If one has worked hard from morning till night, though he may have grieved that he could not be watching the train of his thoughts during that time, yet the few hasty lines which at evening record his day's experience will be more musical and true than his freest but idle fancy could have furnished.... The scholar may be sure that he writes the tougher truth for the calluses on his palms. They give firmness to the sentence.
>
> Indeed, the mind never makes a great and successful effort, without a corresponding energy of the body."
>
> — Henry David Thoreau, A Week on the Concord and Merrimack Rivers (1849)

Work, whether daily practice of chores or performance of an extraordinary task, always represents the engagement of an individual with the materiality of the world. In a learning environment, the enrichment of that work by empowering space for dialogue among practitioners develops community. The geographer Yi Fu Tuan distinguishes between space — as a region empty of human memory or associations — and place — spaces re-created through individual or community stories, which suggests that inhabitation and work in a locale provides it with meaning.

If threads of narrative are similarly built through interactions between the individual and the world, the ensuing story would consist of an archive of field notes and observations layered with reflections that give meaning to experience and underscore field or classroom experience as learning.

Hyperreflection and Storytelling

Concept

Reflecting from within the heart of experience.

Application

Non-text-based, 'good-enough' video, audio, image, and soundscape capture. Geotagging media can foreground the relationship between place and story as well as place and community.

Discussion

"Reflection from the midst." — Maurice Merleau-Ponty

Tools can complement real experience. Reflection is a critical aspect of experience as learning; however, technology can aid in the creation of boundary events during experience to empower later reflection. If we grant that reflection is itself a form of experience, then the reflective exercise becomes a graduated event, mirroring a process writing curriculum, from field observation and description to narrative and synthesis. Hyperreflection can help generate individual narratives of experience that can coalesce and create a larger community story.

"The real Logos," asserts philosopher and environmentalist David Abram "is EcoLogos." Abram explores Maurice Merleau-Ponty's ideas, expressed in The

Phenomenology of Perception and elsewhere, about the dialogic nature of the body and the world: “language is everything, since it is the voice of no one, since it is the voice of the things, the waves, the forests.”

Capturing Reality

Concept

Is language an obstacle to reflection on experience?

Application

Mapping indelible, dynamic trajectories across space regardless of media.

Discussion

> “What is crucial for us here is the place from which this real erupts: the very borderline separating the outside from the inside, materialized in this case by the windowpane” — Slavoj Žižek, Looking Awry, 1995

Michel de Certeau's suggestion in that individuals' "trajectories form unforeseeable sentences, partly unreadable paths across a space" — although he is writing particularly about consumer culture — resonates with David Abram's premise that language can be far more sensuous and experiential than our cultural scaffolding of mere graphemes enables us to be. Both de Certeau and Abram point both to the physicality of the Real and the challenges of representation.

If language can, in essence become an obstacle to effective reflection on individual or community experiences, what are the possible solutions? what tools can support reflection in media res?

Participation

Concept

Effective learning environments are systems that elicit and empower user contribution and collaboration.

Application

Students can shape their engagement of physical learning environments by scaffolding their experience with collaborative environments where they define their place and expectations of experience.

Discussion

If the work/learning community depends upon learning opportunities, facilitation, experience, and effective reflective practices, it can be aligned with Tim O'Reilly's definition of what he calls a Participatory Architecture, which points to "systems that are designed for user contribution." Whether explicitly or not, working communities both create and are created by the architectures within which they function. Thus, the spaces working communities inhabit simultaneously define and are defined by the place and the work itself. As Tim Cresswell writes, place is "something producing and produced by ideology" and "meanings of place are produced through practice."

Pedagogy and Reflection

Concept

Experience without reflection is merely activity.

Application

Pools of data streams should be aggregated and filtered by students.

Discussion

The engagement, collaboration, and experience that current and emerging technologies purport to provide are only metaphors for the learning/teaching in which many of us are already engaged. Getting one’s hands dirty in the performance of literal, actual, meaningful work can be the scaffold for community, collaboration, and engagement that technology can potentially help facilitate. It is this very interface of ‘high touch’ engagement with students in experiential learning and the ‘high tech’ of collaborative technologies that has been challenging my thinking about technology lately — how to be sure that effective technology supports rather than replaces the meaning of experience.

In the high-touch environment of an experiential work/learning curriculum, experience becomes part of the learning experience only through reflection. High tech tools that support such reflection can enhance reflective practice by creating a pace layered archive of experience from field notes, photos, and video streams that pool in collaborative

archives and workspaces, which can then be tapped for continued co-creative learning through more refined and reflective applications such as weblogs, wikis, and pooled again by feed aggregation tools for summary and synthesis.

Environment

Concept

To engage in the environmental crisis through education, we must make space in learning for engagement with the environment.

Application

Global shared forms. Networked blogs to effect real social and political change. Data mining of weblogs.

Discussion

The ecosystem has become as much a metaphor for collaborative technologies as it presents a framework within which to contemplate its development; however, as much as ecology may be an apt metaphor for digital community — in its dynamic development and organic integration of ideas in (often serendipitous) boundary objects, there continues to be a tension between the ubiquity of software and the reality of experience, a tension which is ignored by many.

Self, society, and environment always inhabit the same space — thus creating a layered topography of individuals and their context.

If knowing is defined as being "constructed through the engagement between bodies and machines within the world...this knowledge can be arrived at through a range of methodologies and voices" (Susan Kozel), how do we engage the palpable existence of the world we are trying to "save" without knowing it?

Bibliopedia, Linked Open Data, and the Web of Scholarly Citations

Cenkl, Pavel Thomas
pcenkl@sterlingcollege.edu
Sterling College, United States of America

Overview

Bibliopedia, which recently completed an NEH Digital Humanities Start-Up Grant, performs data-mining and cross-referencing of scholarly literature to create a humanities-centered collaboratory. Currently a working prototype, Bibliopedia can search resources including JSTOR and Library of Congress for metadata about scholarly articles and books, examine the articles and books for citations, then present the results in a publicly accessible database. Bibliopedia is designed to work with all humanities scholarship. It will also allow users to create browsable and customizable bibliographies of all the works cited by each article and book. Most importantly, it uses semantic web technology to enable automated textual analysis, data extraction, cross-referencing, and visualizations of the relationships among texts and authors. Using existing open source software, it extracts citation data from existing plain text resources and transforms them into linked open data. This process makes the information easily accessible to the wider scholarly and linked data communities, enables network visualizations of the scholarly landscape. This presentation will cover the details of the Bibliopedia system to show others how they can replicate it. We will also offer to all interested academic parties our existing installation and hosting platform for their experimentation. In particular, we will present our Drupal-based semantic wiki, which features a full web services API, and our custom citation crawler.

Linked open data, one of the core technologies of the semantic web, promotes open sharing of digital scholarly research while it encourages further, potentially unexpected uses. Bibliopedia's method for incorporating linked open data (via RDFa) requires only minimal technical expertise to reproduce. One of the central components of Bibliopedia is the Drupal content management sytem (CMS), which as of version 7 exposes data via RDF/RDFa as part of its core functionality. This functionality, moreover, is not limited to Drupal. For example, Omeka, another CMS developed at the Roy Rosenzweig Center for History and New Media, George Mason University, has some limited support for linked data through its DublinCoreExtended plugin. Bibliopedia demonstrates the power and flexibility of Drupal's approach to linked data while providing more general lessons for digital humanists who seek to incorporate this technology into their projects.

Project Details

Bibliopedia will aid humanities researchers of all levels of expertise by making simple the currently difficult tasks of discovering new scholarly works and the relationships

among them. It will create an a scholarly community to verify and elaborate cross-referenced, linked bibliographic data through easy-to-use wiki pages. Scholarly literature will become browsable not only backwards in time, but also forwards, something that is currently impossible.

The semantic web is transforming the Internet from a collection of pages and data readable only by humans to one that machines can understand and process. Semantic web technology promises the ability automatically to determine meaning and then infer connections among different elements, thereby vastly improving search capabilities, discovery of new information, and the overall usefulness of the Internet. Just as information accessible only to humans comprises the great majority of the general Internet, so too is data about scholarly literature locked away in text that computers cannot process without great difficulty. At best, search engines for repositories such as JSTOR permit researchers to query author name, journal titles, and keywords, but once a work is found, the search stops. No connections among works are found precisely because machines cannot currently read that data. Although Google Scholar attempts to show citations of articles, its usefulness is highly limited because it does not make clear the relationships among articles, present very limited metadata about each article (if any), fails to provide for community elaboration or correction, and includes only works that are publicly available. Yet despite its limitations, Google Scholar stands as a significant technological advance beyond keyword-based search engines such as those provided by JSTOR and Project Muse.

Bibliopedia will, by aggregating data from as many sources as possible, converting citations into semantic web format, and then cross-referencing an ever-growing database of scholarly works, be able not only to overcome many of the limitations of Google Scholar and become a powerful research tool in its own right, but also to make a valuable contribution to the growing semantic web. Introducing high quality metadata about humanities scholarship to the semantic web will enable others in the semantic web/linked data world to process that data in new, unexpected ways that will accrue further benefits to the scholarly community. For example, the standards underlying the semantic web make data visualization and automated inferences about relationships trivially easy rather than the complex problems such tasks currently present. Bibliopedia will, then, through the innovation of placing metadata about scholarly literature into a linked data format, open up a vast range of possible future innovations and analyses based on that data, which is currently locked away and readable only by select humans.

Another virtue of a linked data format is that it will help resolve many of the challenges inherent in metadata, some will inevitably remain. Rather than attempt to solve this incredibly complex problem through automation alone, then, Bibliopedia will, in the process of displaying its results for human consumption, also provide for human feedback in the form of correction and elaboration. A common disadvantage of fully automated text analysis and data extraction tools such as Google Books, Google Scholar, and other digital research tools is that their automatic parsers have errors in their metadata that they do not allow subject matter experts to repair. Bibliopedia will pursue the goal of unifying that information into an environment that not only displays the information efficiently, but actively encourages crowd-sourcing metadata on books, articles, and publications of all kinds. In thus opening data up to revision by the scholarly community, Bibliopedia can build on the strong work of mature data silos, improve overall data quality, and provide the academic community at large a continuously evolving research tool.

There currently exists a multitude of projects and tools designed to work with book metadata, cross-reference scholarly articles (localized to the sciences), or create user communities around a chosen interest. Further, some of the most important trends currently revising the ways we use technology are social media, collaboration, and data aggregation. By incorporating the benefits realizable from each of these trends, Bibliopedia will create a powerful tool for scholarly research at all levels. None of the existing tools, however, focus on scholarship for the humanities, nor do they present the information in the linked data format necessary to the semantic web.

Linked Open Data & the OpenEmblem Portal

Cole, Timothy W.
t-cole3@illinois.edu
University of Illinois at UC, United States of America

Han, Myung-Ja K.
mhan3@illinois.edu
University of Illinois at UC, United States of America

Wade, Mara R.
mwade@illinois.edu
University of Illinois at UC, United States of America

Stäcker, Thomas
staecker@hab.de
Herzog August Bibliothek at Wolfenbüttel, Germany

1. Introduction

Supported by the National Endowment for the Humanities (NEH) and the Deutsche Forschungsgemeinschaft (DFG), the University of Illinois at Urbana-Champaign (Illinois) and the Herzog August Bibliothek, Wolfenbüttel (HAB), have digitized 728 Renaissance emblem books, thereby substantially expanding the digitized corpus (Wade et al., 2012; Daly, 2002).
Each book contains tens, even hundreds, of individual emblems. All together Illinois and HAB have digitized approximately 70,000 individual emblems, creating detailed descriptions(emblem-level metadata)for more than 17,000 of these. Each emblem is identified with a globally unique URI (Uniform Resource Identifier) maintained in a shared emblem registry. The *OpenEmblem Portal* prototype[1] was collaboratively designed and built to provide access to these materials and to demonstrate the feasibility of international repository interoperability.

Experimentation with Linked Open Data (LOD) services and RDF-based annotation tools, described below, is now underway to demonstrate how Semantic Web technologies can facilitate both discovery and the use of digitized emblem resources. While the *OpenEmblem Portal* focuses currently on emblems digitized from print, emblematic modes of expression permeate the fine and applied arts as well. Through adherence to LOD best practices and emerging annotation standards, the *OpenEmblem Portal* eventually will allow scholars to link an emblem design found in a Bavarian church or a Swedish manor house to a printed emblem. This paper reports on our progress with LOD technologies and describes planned next steps to leverage LOD to facilitate emblem research and pedagogy.

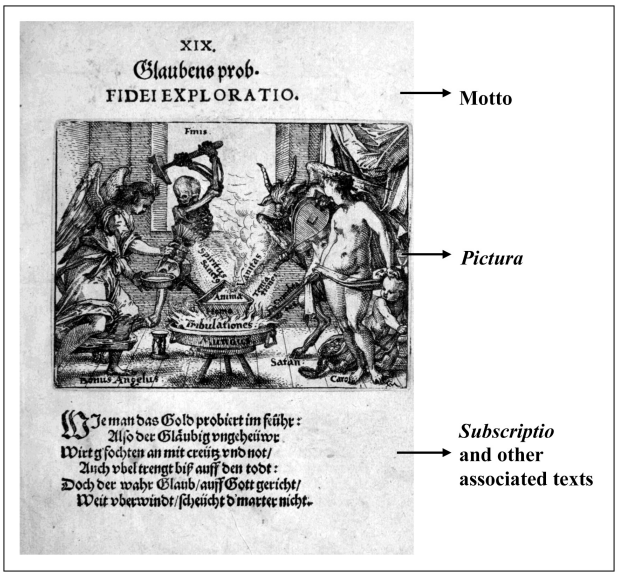


Figure 1:
An early modern emblem (Murer 1622)

2. Motivation

Emblems are hybrid, integrating texts and graphics(see Figure 1). Emblem texts were written in European vernacular languages and Latin. As a genre of adaptation, emblem components — *picturae*, mottos and other associated texts — were constantly (re)used in new constellations to create new meaning, and circulated widely in many different interpretive contexts. For example, Classical and Biblical themes regularly appear across the genre, in emblem images and texts. Emblem-based scholarly inquiry focuses on tracing themes and motifs across authors, over time, through many languages, and in specific religious, political, and social contexts. The purview of emblems includes all areas of Renaissance knowledge from natural history and technology, to secular and divine love.

Discoverability by theme and topic, preferably through a hierarchical, multi-lingual thesaurus, is critical to emblem research and pedagogy. The ready availability of contextual information concerning emblems facilitates both emblem discovery and author-based scholarly investigations and suggests good potential for LOD approaches. The use of annotation to enrich emblem metadata with newly discovered relationships has the potential to stimulate scholarly discourse. The hypothesis motivating the current study postulates that the use of LOD approaches and data sets will allow us to better leverage existing ontologies(e.g. Iconclass[2]), name authority resources(e.g. the Virtual International Authority File [VIAF][3]), and compatible annotation models (e.g. Open Annotation[4]). Though LOD technologies are now well understood generically, their practical application to emblem studies is novel and represents an opportunity to apply LOD approaches to a mixed text-image corpus of interest to a well-established international community.

3. Enhancing Discoverability through Linked Open Data

Iconclass, a multilingual classification system for cultural heritage resources, is a hierarchical thesaurus consisting of more than 28,000 controlled vocabulary terms describing objects, people, events and abstract ideas (e.g. Figure 2).Iconclass has proven well-matched to emblem literature. In populating our emblem-level metadata records, we utilize more than 15% of Iconclass headings. Iconclass has been published as a LOD data set, and iconclass.org provides LOD services. In our Portal these services enable multi-lingual browsing of digital emblematica. Users

browse the Iconclass hierarchy in their preferred language to discover emblem content. Because we employ Iconclass LOD services, this functionality is enabled without a local copy of the thesaurus. Figure 2 shows the Iconclass Notation "25F33(EAGLE)" in Resource Description Framework(RDF)[5] XML as retrieved dynamically from iconclass.org.

```
<rdf:RDF  xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
          xmlns:skos="http://www.w3.org/2004/02/skos/core#">
  <rdf:Description rdf:about="http://iconclass.org/25F33%28Eagle%29">
    <rdf:type rdf:resource="http://www.w3.org/2004/02/skos/core#Concept"/>
    <skos:prefLabel xml:lang="fr">oiseaux de proie, rapaces (avec NOM)</skos:prefLabel>
    <skos:prefLabel xml:lang="en">predatory birds (with NAME)</skos:prefLabel>
    <skos:prefLabel xml:lang="de">Greifvögel (mit NAMEN)</skos:prefLabel>
    <skos:prefLabel xml:lang="it">uccelli rapaci (col NOME)</skos:prefLabel>
    <skos:prefLabel xml:lang="fi">petolinnut (NIMEN kanssa)</skos:prefLabel>
    <skos:inScheme rdf:resource="http://iconclass.org/rdf/2011/09/"/>
    <skos:notation>25F33(Eagle)</skos:notation>
    <skos:narrower rdf:resource="http://iconclass.org/25F33%28...%29%28%2B0%29"/>
    <skos:narrower rdf:resource="http://iconclass.org/25F33%28...%29%28%2B1%29"/>
    <skos:narrower rdf:resource="http://iconclass.org/25F33%28...%29%28%2B2%29"/>
    <skos:narrower rdf:resource="http://iconclass.org/25F33%28...%29%28%2B3%29"/>
    <skos:narrower rdf:resource="http://iconclass.org/25F33%28...%29%28%2B4%29"/>
    <skos:narrower rdf:resource="http://iconclass.org/25F33%28...%29%28%2B5%29"/>
    <skos:narrower rdf:resource="http://iconclass.org/25F33%28...%29%28%2B6%29"/>
    <skos:narrower rdf:resource="http://iconclass.org/25F33%28...%29%28%2B7%29"/>
    <skos:narrower rdf:resource="http://iconclass.org/25F33%28...%29%28%2B8%29"/>
    <skos:narrower rdf:resource="http://iconclass.org/25F33%28...%29%28%2B9%29"/>
    <skos:narrower rdf:resource="http://iconclass.org/25F33%28EAGLE%29"/>
    <skos:narrower rdf:resource="http://iconclass.org/25F33%28FALCON%29"/>
    <skos:narrower rdf:resource="http://iconclass.org/25F33%28FISH-HAWK%29"/>
    <skos:narrower rdf:resource="http://iconclass.org/25F33%28HAWK%29"/>
    <skos:narrower rdf:resource="http://iconclass.org/25F33%28KITE%29"/>
    <skos:narrower rdf:resource="http://iconclass.org/25F33%28SPARROW-HAWK%29"/>
    <skos:narrower rdf:resource="http://iconclass.org/25F33%28VULTURE%29"/>
    <skos:narrower rdf:resource="http://iconclass.org/25FF33%28...%29"/>
    <skos:broader rdf:resource="http://iconclass.org/25F33"/>
  </rdf:Description>
</rdf:RDF>
```

Figure 2:
Iconclass notation in RDF XML

This RDF XML is used in our Portal to support Iconclass browsing for emblem discovery. In the scenario illustrated in Figure 3, a user has found an emblem from Johann Vogel's *Meditationes Emblematicae de Restaurata Pace Germaniae.* One of the Iconclass headings assigned to this emblem is 25F33(EAGLE)(+5245), "predatory birds: eagle (+ animal(s) holding something)." The user can move from the display of emblem metadata to a view of the Iconclass hierarchy, entering at this heading, and then browse up or down the hierarchy in English, French, German, or Italian, to retrieve other emblem *picturae* having related Iconclass headings.

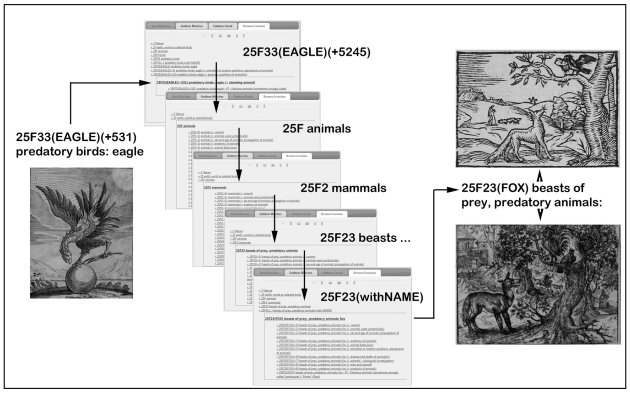


Figure 3:
From an emblem metadata display,scholars can browse the Iconclass hierarchy to find other emblems with related headings

We are now linking to name authority services to facilitate emblem discovery by name. Many national libraries have released author name data sets as LOD. These records include variant name forms and links to Websites having contextual information about individual authors. By using LOD services and associating name strings in our indices with canonical identifiers — e.g. http://viaf.org/viaf/54384883/ , the *OpenEmblem Portal* can offer expanded discovery capabilities. VIAF RDF records(e.g. Figure 4)integrate links to and summaries of information about authors maintained by multiple national libraries, including by the *Deutsche Nationalbibliothek* (DNB).[6] So far, we have been able to automate the linking of 90% of our emblem book metadata records to VIAF name entries.

```
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
         xmlns:owl="http://www.w3.org/2002/07/owl#"
         xmlns:foaf="http://xmlns.com/foaf/0.1/"
         xmlns:skos="http://www.w3.org/2004/02/skos/core#"
         xmlns:rdaGr2="http://rdvocab.info/ElementsGr2/" >
  <rdf:Description rdf:about="http://viaf.org/viaf/54384883">
    <rdf:type rdf:resource="http://xmlns.com/foaf/0.1/Person"/>
    <rdf:type rdf:resource="http://rdvocab.info/uri/schema/FRBRentitiesRDA/Person"/>
    <foaf:name>Vogel, Johann, 1589-1663</foaf:name>
    <foaf:name>Vogelius, Johannes, 1589-1663</foaf:name>
    <foaf:name>Vogel, Johannes 1589-1663</foaf:name>
    <foaf:name>Vogelius Joannes 1589-1663</foaf:name>
    <rdaGr2:dateOfBirth>1589</rdaGr2:dateOfBirth>
    <rdaGr2:dateOfDeath>1663</rdaGr2:dateOfDeath>
    <owl:sameAs rdf:resource="http://d-nb.info/gnd/1014349664"/>
  </rdf:Description>
  <skos:Concept rdf:about="http://viaf.org/viaf/sourceID/DNB%7C1014349664#skos:Concept">
    <skos:inScheme rdf:resource="http://viaf.org/authorityScheme/DNB"/>
    <skos:prefLabel>Vogel, Johann, 1589-1663</skos:prefLabel>
    <skos:altLabel>Vogelius, Johannes, 1589-1663</skos:altLabel>
    <skos:altLabel>Vogel, Johannes 1589-1663</skos:altLabel>
    <foaf:focus rdf:resource="http://viaf.org/viaf/54384883"/>
  </skos:Concept>
  <skos:Concept rdf:about="http://viaf.org/viaf/sourceID/BNF%7C14536690#skos:Concept">
    <skos:inScheme rdf:resource="http://viaf.org/authorityScheme/BNF"/>
    <skos:prefLabel>Vogel, Johann, 1589-1663</skos:prefLabel>
    <skos:altLabel>Vogelius, Johannes, 1589-1663</skos:altLabel>
    <skos:altLabel>Vogel, Johannes 1589-1663</skos:altLabel>
    <skos:altLabel>Vogelius Joannes 1589-1663</skos:altLabel>
    <skos:exactMatch rdf:resource="http://data.bnf.fr/ark:/12148/cb145366903"/>
    <rdfs:seeAlso rdf:resource="http://catalogue.bnf.fr/ark:/12148/cb145366903"/>
    <foaf:focus rdf:resource="http://viaf.org/viaf/54384883"/>
  </skos:Concept>
  ...
</rdf:RDF>
```

Figure 4:
Fragment of VIAF RDF record for Johann Vogel

Many records in VIAF list alternate name forms, i.e., name variants by which authors also are known. Name variants are indexed and made searchable through the VIAF OpenSearch service[7] . Scholars often know an author by one name and search for resources accordingly. Thus, though not in the Portal's own indexes, a search for "Joannes Vogelius" using the VIAF service returns a list of author identifiers, including one for "Johann Vogel (1589-1663)," who is the author of two volumes available through the *OpenEmblem Portal.* Rather than independently maintaining all name variants in our local indices, the Portal in its next stage of development will leverage canonical URIs and remote LOD services to enhance discoverability.

4. Providing Context through Linked Open Data

VIAF LOD services also can provide users contextual information about emblem book authors. Enriching metadata with canonical URIs for named entities can enable users to link to contextual information pertaining to an author. Currently, VIAF identifiers can be used to link to OCLC's *WorldCat Identities*[8] , *Wikipedia* entries[9] , and multiple national library authority entries. As illustrated in Figures 5 and 6, users can see author gender and nationality, names of co-authors, lists of other publications, and descriptions of the individual's religion and life history.(We are also investigating linking to work-level bibliographic authorities such as VD17[10] and VD18[11] .)

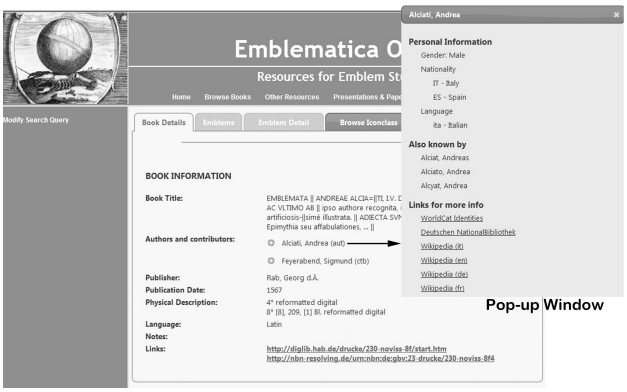


Figure 5:
Information distilled from VIAF records tells users more about authors

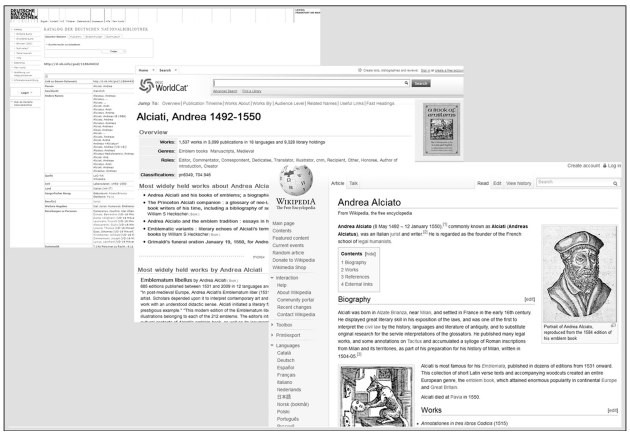


Figure 6:
Links from the pop-up window also connect users to rich contextual information

5. Support for Annotation & Scholarly Discourse

Annotation is a pervasive element of scholarly practice, employed both to organize knowledge and to facilitate the creation and sharing of new knowledge. As a tool of scholarly discourse, Web-based annotation of digital resources has the potential to facilitate research, collaboration, and pedagogy. The W3C Open Annotation Community Group[12] has proposed a data model for scholarly annotation that aligns well with the principles of LOD. According to this view, a student's annotation of *picturae* from two possibly related emblems might be represented graphically in Figure 7.

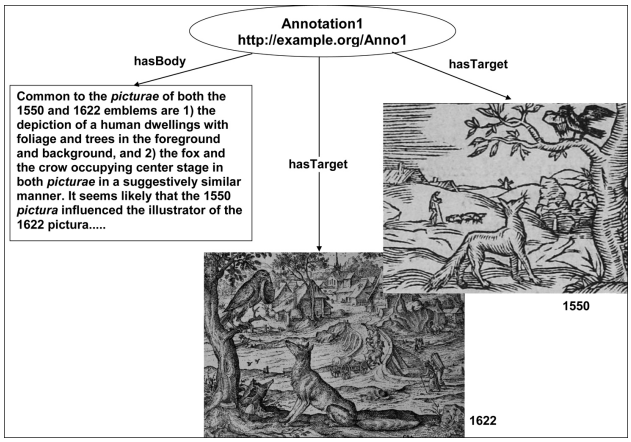


Figure 7:
Graphical representation of an annotation involving two picturae

Persistent emblem and annotation identifiers are essential to expressing this annotation as a RDF graph aligned with LOD best practice. The Open Annotation model is extensible and recursive. For example, an instructor might add a subsequent annotation referencing the first student's original annotation and including additional analysis (Figure 8).

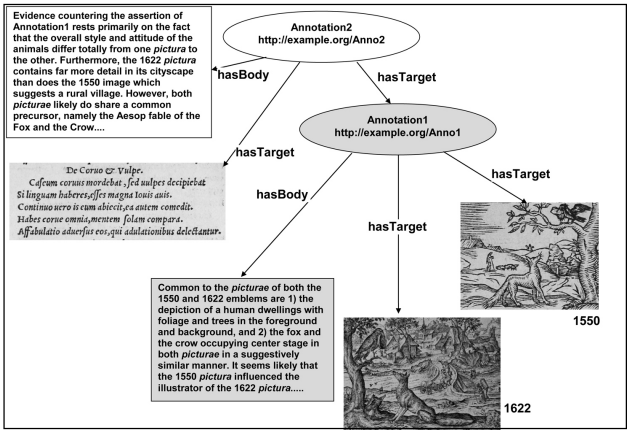


Figure 8:
Start of a chain of annotations (i. e. a netchain [Sukovic, 2008; 2011])

6. Conclusion

Initial experimentation using LOD techniques in conjunction with the metadata describing resources indexed in the *OpenEmblem Portal* suggests a potential to enhance the discoverability and usability of digitized emblem resources. Faster discovery of and linking to context about authors, places, and other attributes of emblematica has significant potential to propel emblem scholarship and pedagogy. Contextualization that formerly required multiple library visits can now be made available instantly. In tension with this is the risk of information overload and the hazard of diminishing performance as the network of services involved in any transaction grows; connections to LOD data sets must be implemented carefully and used judiciously to realize maximum benefit. LOD technologies also have potential to enable more robust scholarly annotation tools, thereby facilitating asynchronous scholarly discourse. Just as the emblem-level indexing of Henkel and Schöne (1967) spurred new emblem studies research then, early experimentation suggests the potential of LOD technologies to facilitate new emblem studies research initiatives today.

References

Daly, P. M. (2002). *Digitizing the European Emblem: Issues and Prospects.* New York: AMS Press.

Gueroult, G. (1550). *Premier livre des emblemes. Composé par Guillaume Gueroult.* Lyons: Balthazar Arnoullet.

Henkel, A., and A. Schöne (1967). *Emblemata.* Stuttgart: Metzler.

Murer, C. (1622). *XL [i.e. Quadraginta] emblemata miscella nova: Das ist: XL. underschiedliche Auszerlesene newradierte Kunststuck.* Zurich: Wolf.

Sukovic, S. (2008). Convergent Flows: Humanities Scholars and Their Interactions with Electronic Texts. *The Library Quarterly* 78(3): 263-284.

Sukovic, S.(2011). E-Texts in Research Projects in the Humanities. *Advances in Librarianship* 33: 131-202.

Wade, M., T. Stäcker, R. Stein, D. Graham, and H. Brandhorst (2012). Digital Emblematics — Enabling Humanities Research of a Popular Early Modern Genre. *Poster Session, Digital Humanities Annual Conference* held July 2012 in Hamburg, Germany.

Notes

1. http://emblematica.grainger.illinois.edu/OEBP/UI/SearchForm
2. http://iconclass.org/
3. http://viaf.org/
4. http://www.openannotation.org/spec/beta/
5. http://www.w3.org/RDF/
6. http://www.dnb.de/DE/Home/home_node.html
7. http://www.oclc.org/developer/documentation/virtual-international-authority-file-viaf/request-types#opensearch
8. http://www.oclc.org/developer/documentation/worldcat-identities/using-api
9. http://en.wikipedia.org, http://de.wikipedia.org, http://fr.wikipedia.org , etc.
10. http://www.vd17.de
11. http://www.vd18.de
12. http://www.w3.org/community/openannotation/

Solitary Mind, Collaborative Mind: Close Reading and Interdisciplinary Research

Coles, Katherine
k.coles@english.utah.edu
University of Utah, United States of America

Lein, Julie Gonnering
jkglein@gmail.com
University of Utah, United States of America

We have discussed in some detail elsewhere, including in a longer paper at this conference, our ongoing development of software that visualizes poems as complex dynamic systems (Abdul-Rahman, et al. 2013a; Abdul-Rahman, et al. 2013b). In this presentation, we will discuss how working within this interdisciplinary context, which forced us to think about how poems work both as large, complex systems and on their most granular levels, has already led us to new insights about poetry, its features, and its operations. Our focus here will be not on the technology itself, nor on the literary perspectives we have been bringing to bear in its design. Rather, we will discuss how working collaboratively with visualization scientists, whose research practices and values could not be more different from our own, has caused us to imagine poems differently, and on how this adjustment in our thinking, which would not otherwise have happened, has *in itself, independently* of the technology, led us to new insights about poetry. Interdisciplinary dialogues are also revising our ideas about broader subjects like aesthetics and information as well as the specific ways our different fields perceive them acting on each other. For this reason, we argue, collaboration across the boundaries of very different disciplines has its own inherent value, even to a discipline like ours in which writing, reading, and the 'freedom to explore' has most traditionally been associated with solitude.

Even as interest in distant reading (Moretti 2007) rises in the digital humanities, close reading remains central to the practices at the heart of literary scholarship. As helpful as it can be in opening new directions of criticism, distant reading — with its reliance on abstract models based on data mining and quantification — removes human readers from the center of the reading process, bringing them in after the 'reading' phase is finished to interpret not texts but data. Thus, these approaches fail to support the most important practice in the study of poetry, which is valued precisely because of its experiential richness. In contrast, close readers engage texts directly, intimately and in detail; they trace the finest interactions among such literary features as rhyme and meter, sound, figures, and syntax, noting how even the subtlest movements and operations (a comma, a repeated vowel, etc.) influence a reader's interpretation(s) and experience(s) of a particular poem. While the words in a given poem in a given version remain the same, different skilled readers will interact with those words in different ways by choosing moment by moment what to engage and what to ignore. Thus, there is no single 'correct' solution to any close reading problem, though there may be many incorrect solutions. A reader's choices are led both by what is happening on the page and by the reader's preferences and interests. Close reading, as both expression and experience, thus manifests interactions between poems and human minds.

While close reading is sometimes positioned as belonging to the set of practices not amenable to digitization, we have found in our own work and in a review of the work of others that this is not necessarily the case. In fact, there is no shortage of software designed and applied with the *intention* to aid close reading practice, whether or not it is yet effective in doing so (Chaturvedi 2011; Chaturvedi, et al., 2012; Clement 2012; Plamondon 2006; Ruecker, et al., 2008; Unsworth and Mueller 2009). This abundance indicates, in our view, an urgent desire in the literary community to embrace and explore the power of computation while at the same time prioritizing and protecting the relationship between literature and human readers. As we explain in our longer paper, one distinguishing feature of our project is the strong emphasis we are placing on poetry's experiential quality, which we locate in its radical multidimensionality — especially its relationship to time. But we are also, through our interdisciplinary discussions, endeavoring to theorize how to position data visualization and computers as potential — and potentially potent—tools to aid close reading.

Our goal of using visualization to heighten poetic experience has been challenging to describe, understand, and pursue for all the members of our team. In fact, we have struggled to integrate into this overarching objective two distinct and not always obviously compatible elements: intensified aesthetic experience and revelation of new information. Early in our research, we looked to music visualizations for possible analogies that might inform our work. But sometimes impassioned discussions have shown us, for instance, that while visualizers like MilkDrop (Geiss 2012) and the 'visual music' of multimedia artists Abstract Birds (2012) may enhance aesthetic pleasure by combining visual and aural elements, they do not necessarily lead viewers to new intellectual perceptions about the music. The computer scientists in our group prioritize information clarity and communication through timesaving tools and techniques. Accustomed to collaborating with biologists, engineers, physicists and physicians to visualize phenomena like combustion, the brain and its electrical impulses, magnetic fields, etc., they have emphasized that their work is more than 'just pretty pictures.' At the same time, we poets have consistently defended the aesthetic as meaningful: in poet Robert Creeley's words, "FORM IS […] AN EXTENSION OF CONTENT" (Olson 1966, p. 16, emphasis original). More is at stake than superficial display, we have argued; especially when visualizing aesthetic objects, events, or systems (like poetry), aesthetic choices are necessarily core design considerations. The aesthetics of individual visualizations should somehow reflect the uniqueness of individual poems.

Similarly, our visualization scientist colleagues are used to identifying specific problems to solve, data to isolate,

hypotheses to test: explicit aims that help them define programming strategies and evaluation techniques for very complicated software. Our stance on verifiability and accuracy is more ambivalent and ambiguous. Initially we resisted the very idea of literary hypotheses-making and testing, until we saw it as describing the many questions and choices a close reader constantly makes on the fly in her encounter with a poem. We continue to contemplate what 'poetic data' might mean, and what its relationship to literary forms and devices like lineation and metaphor might be. Even in our initial focus on poetic sound, the issue of accuracy has proven difficult to treat. How should we mathematically define rhyme? What constitutes the smallest measurable repeating sonic cluster? Of course we want our program to be able to correctly identify single repeating phonemes to show assonance and consonance. But accuracy becomes categorically more uncertain and tenuous as soon as we move beyond single phonemes. Should only full end rhymes qualify as rhymes? What about internal rhymes? Slant rhymes? As we build this software in hopes of its being useful for as many readers as possible, we want to consider not only how we, but how others whose literary perspectives differ from our own, might answer such questions. We want our software to allow individual reader-users to choose how to define some of these parameters in order to best support their unique close readings.

In our program, then, visualizations are not authoritative arbiters to reduce poetic complexity, give definitive answers, or merely save time; rather, they are *aesthetic agents*, using visual constructs and perception to reveal poetic features, patterns, and qualities readers might not otherwise have noticed by prompting them toward fresh experiences, insights, and questions. Thus, the visualizations do not replace close reading; they suggest rich avenues for initial and subsequent explorations, cuing the reader into which operations of the poem may merit further investigation. This orientation has also encouraged us to consider how the human-computer interaction developing through this software may prove more nuanced than the word 'tool' superficially suggests. It has provoked us to ask whether and in what sense the computer generating these visualizations might fruitfully be considered a fellow literary entity and even collaborator in the close reading process—questions we have differed on, as we will explain in our presentation.

Given our usual scholarly practices — based in the essentially experiential and qualitative nature of poetry, not to mention in the space of determined solitude where poetry is usually experienced—we were not natural targets for a project in visualizing poetry or even for a project that is highly collaborative. Of course we recognize that collaboration can generate programs and poems that otherwise would not have been possible. And this interdisciplinary project is indeed accomplishing those results. But it is also challenging us to examine fundamental issues, including our expectations and understanding of collaboration itself. We were brought into the project by our collaborators' commitment to respond to our existing values and practices. Among the surprises for us has been the usefulness of collaboration itself to the practices and priorities we already embrace. At the same time, when one of the computer scientists in our group recently invoked Gregory Bateson's notion of information as *'the difference which makes a difference,'* (1972, p. 453, emphasis original) we recognized it as an apt description of our interdisciplinary research: exploring our various differences is not only leading us to new territory, it is helping us see familiar ground anew.

Funding

This work was supported by a Digging Into Data Challenge grant: in the US, by the National Endowment for the Humanities; and in the UK by the Arts and Humanities Research Council, Economic and Social Research Council, and JISC. We would also like to acknowledge and thank our collaborators: Alfie Abdul-Rahman, Min Chen, Christopher Johnson, Eamonn Maguire, Miriah Meyer, and Martin Wynne.

References

Abdul-Rahman, A., K. Coles, J. Lein, and M. Wynne (2013a). Freedom and Flow: A New Approach to Visualizing Poetry. Paper to be presented at Digital Humanities 2013, University of Nebraska-Lincoln. July 2013.

Abdul-Rahman, A., J. Lein, K. Coles, E. Maguire, M. Meyer, M. Wynne, A. E. Trefethen, C. Johnson, and M. Chen (2013). Rule-based Visual Mappings — with a Case Study on Poetry Visualization. Paper conditionally accepted to EuroVis, Leipzig, Germany.

Abstract Birds. (2012). http://www.abstractbirds.com (accessed 19 February 2013).

Bateson, G. (1972). Form, Substance, and Difference. In *Steps to an Ecology of Mind.* New York: Ballantine.

Chaturvedi, M. (2011). *Visualization of TEI Encoded Texts in Support of Close Reading.* M.S. thesis. Ohio: Miami University of Ohio. http://etd.ohiolink.edu/sendpdf.cgi/Chaturvedi%20Manish.pdf?miami1323623830 (accessed 6 September 2012).

Chaturvedi, M., G. Gannod, L. Mandell, H. Armstrong, and E. Hodgson (2012). Myopia: A Visualization Tool in Support of Close Reading. *Digital*

Humanities 2012. held July 2012 at University of Hamburg. http://lecture2go.uni-hamburg.de/konferenzen/-/k/13930 (accessed 6 September 2012).

Clement, T. (2012). Methodologies in the Digital Humanities for Analyzing Aural Patterns in Texts. *Proceedings of the 2012 iConference.* Toronto, Canada, February 2012. ACM Digital Library: http://dl.acm.org/citation.cfm?id=2132213 (accessed 6 September 2012).

Geiss, R. (2012). *MilkDrop.* http://www.geisswerks.com/milkdrop/ (accessed 19 February 2013).

Moretti, F. (2007). *Maps, Graphs, and Trees: Abstract Models for Literary History.* London: Verso.

Plamondon, M. (2006). Virtual Verse Analysis: Analysing Patterns in Poetry. *Literary and Linguistic Computing.* 21(Suppl), 127-141. http://llc.oxfordjournals.org (accessed 3 December 2012).

Ruecker, S., M. Radzikowska, P. Michura, C. Fiorentino, and T. Clement. (2008). Visualizing Repetition in Text. *CHWP* A.46 http://projects.chass.utoronto.ca/chwp/CHC2007/Ruecker_etal/Ruecker_etal.htm (accessed 6 September 2012).

Unsworth, J., and M. Mueller. (2009). *The MONK Project Final Report.* http://www.monkproject.org/MONKProjectFinalReport.pdf (accessed 6 September 2012).

A 3D Common Ground: Bringing Humanities Data Together Inside Online Game Engines

Coltrain, James Joel

james.coltrain@gmail.com
University of Nebraska, United States of America

Over the past decade humanities scholars have begun to seriously explore the potential of using 3D modeling software to reconstruct historic spaces and architecture. However, because of technological and design limitations, their works have tended either to feature visually rich environments that were closed and static, or more interactive and data driven spatial projects that failed to seize upon the potential of realistic 3D imaging. But now as 3D gaming engines are making it easier to display visually sophisticated virtual environments in browser windows in real-time, and to access and interface with large data sets, new possibilities exist for 3D reconstructions. Using both demos of historical reconstructions running live in the Unity 3D engine and mockups of future plug-ins, this presentation imagines future 3D historical reconstructions that bring a variety of data together, and argues for projects that are persistent, historical, collaborative, and curated.

The use of online 3D engines for the display of historical reconstructions first allows for persistent environments that can constantly be explored by any number of users. These reconstructions might be a single town, building, or other reconstructed site that would exist perpetually on the web, constantly accessible to the public. Previously projects aiming for a high level of detail and realism would have to store their 3D scenes offline, and could only share their work by rendering image stills or animation clips that would show prerecorded views. However, significant improvements in the graphical quality of 3D engine browser plug-ins mean that now, these scenes can be uploaded to the web, and users can walk and fly though historical spaces freely, and even interact with the environment.

The sophistication of these 3D engines also opens up new possibilities for depicting historical change over time. The customization of game engines means that scene elements like individual buildings can easily be assigned time ranges, allowing for an environment that can show growth and changes in the architecture and landscape as users move backwards and forwards through virtual time. These innovations are made possible because 3D gaming engines handle complex, visually realistic 3D models more efficiently than platforms like Google Earth or ArcGIS, meaning scholars no longer must choose between a realistic, but static reconstruction, or a historical GIS scene that showed change over time but could not support advanced imaging.

Because browser engines are more adept at dynamically displaying change within a 3D scene the possibilities for scholarly and community collaboration have enlarged considerably. First, with only a little additional scripting, 3D engines can be customized to access large outside data sets, like those used with Google Maps API or standalone GIS programs, but they also can allow any visitor to annotate virtual space in real-time. In a persistent online reconstruction of historic church for instance, users could walk down the aisles and through the balconies, while pausing to mark spots, leaving virtual footnotes in 3D space, which could then link to outside data, text, or images. Additional scripting could store these user annotations in a database, and allow for the community accumulation of data from different scholars and the public, within the same persistent environment. With more advanced plug-ins, there is even the future possibility of users manipulating the environment itself from inside a browser, adding or modifying architectural elements as part of their own interpretations.

Because these persistent, historical, and collaborative 3D reconstructions could quickly become cluttered with multiple interpretations, dedicated editors who could curate the virtual space could provide a balance between open contributions and consistent quality. First, individual contributions or annotations could be grouped into different layers, which viewers could choose to show or hide, and certain of these layers could be ranked both by user voting, as well as by editor endorsement. Especially in the case of contributions that involved significant changes to the 3D models in the scene, site editors could even act as or coordinate with peer reviewers, allowing for scholars to gain publication credit for adding especially valuable 3D content or other data.

This presentation features as an example, a real time reconstruction of an 18th century North American imperial fort running in a customized version of the Unity 3D game engine. This reconstruction shows the development of fort's architecture over time through various stages of construction, disrepair and remodeling, and also features multiple interpretations of particular architectural elements based on conflicting pieces of evidence. Within this evolving 3D environment, the presentation also show sets of annotations in 3D space that link to copies of archival documents, maps, plans, and archeological artifacts. In addition to plotting historical events or the locations of artifacts, these points may act as footnotes linking to particular pieces of evidence that informed the look or design of a given part of the interpretation. A custom feature programmed into the 3D engine could also allow individual users to tour the space and plot their own original points, each of which can link to multimedia content. Finally, this demonstration will explore possibilities for the importation of Google Earth points, GIS data, and topographic maps into the example 3D reconstruction as distinct viewable layers.

Surrogacy and Image Error: Transformations in the Value of Digitized Books

Conway, Paul
pconway@umich.edu
University of Michigan, United States of America

The large-scale digitization of books is generating extraordinary collections of visual and textual surrogates, whose preservation is premised partly upon expected transformations in teaching and scholarship in the humanities. Questions have been and continue to be raised about the quality and usefulness of digital surrogates produced by third-party vendors and deposited in digital repositories for preservation and access (Cohen 2010). If the surrogacy of published materials that serve as primary sources is to find wide acceptance within humanities scholarship, then those who build and manage preservation repositories must be able to make claims about, validate the quality of, and certify the fitness of use of these preserved digital surrogates. Understanding the relationship between digital surrogacy and the presence or absence of evidence regarding digitization processes is thus a substantial challenge for scholars and preservation archivists alike.

The purpose of the paper is to synthesize and extend the findings and implications of a major ongoing research project at the University of Michigan School of Information. The research explores the relationship between quality (or its absence in the form of unacceptable error) and usefulness of digitized books at scale. The HathiTrust Digital Library is the test bed for the project. HathiTrust is an international repository collaborative that is preserving and providing access to the output of large-scale book digitization projects, including those by Google, the Internet Archive, and a host of localized digitization programs (York 2010). The research reported here is designed to produce some foundation of statistical truth, accompanied by a transparent methodology, so that follow-up user validation studies can explore how digital image error impacts the acceptance of digital surrogates for scholarly inquiry and the management of physical collections in libraries.

This research into the quality and usefulness of large-scale digitization is built on a synthesis of scholarship in multiple fields that typically do not intersect: information quality (Knight 2008), digital image analysis (Lin 2006), relevance clues (Saracevic, 2007), and humanities scholars' use of digital collections (Henry 2010). The research is grounded on a model of error that specifies the gap between the digitization ideal, represented by digitization best practices and standards, and the realities of repositories' acceptance of digitized content produced by third parties. The design of the research (Conway 2011), data gathering and analysis procedures (Conway and Bronicki 2012), and summary findings (Conway 2012) are reported separately. This work was supported by the US Institute of Museum and Library Services [grant number LG-06-10-0144-10]; and The Andrew W. Mellon Foundation.

The emphasis of the research is on the visual representation of books as digitally bound bitmap sequences, derived from sometimes deeply flawed source volumes and produced through a complex set of manual scanning processes and automated post-scan image processing procedures. The transformation of published books to digital code and algorithm is mitigated by the terms of digitization technologies. "The aesthetic

transformations that make digital objects so eloquent are themselves always subject to the functional constraints imposed by the material variables of computation. Understood at this level, digital surrogates are just as 'real' (and tangible) as their analog counterparts." (Eaves 2003, 164) The relationship between source and digital surrogate conforms to the "law of contact" proposed by Taussig (1993): "things which have once been in contact with each other continue to act on each other at a distance after the physical contact has been severed." (52-53) Significantly, digital surrogates produced through high-volume digitization carry with them traces of the terms of their creation. Such traces may inevitably affect the trust that is essential the acceptance of digital surrogates as sources of scholarship. "If we cannot trust our means of reproduction of images of texts, can we trust the readings from them? How do scholars acknowledge the quality of digitized images of texts?" (Terras 2011, 1)

This paper is an explicit effort to foster a conversation on the impact of digital imaging at scale on humanities scholarship by marshaling empirical research data on digitization error to characterize the strengths and limitations of digital surrogacy. The implications for the use of surrogates are derived from data gathered from four 1,000-volume random samples of digital surrogates covering the full range of source volumes digitized by Google and the Internet Archive from more than 20 research libraries. Proportional and systematic sampling of page-images within each volume in the samples produced a study set of over 350,000 page images, which have been evaluated visually by highly trained coders working in two university libraries in the United States. Using a web-enabled database system, coders assigned error severity scores for eleven page-level errors and five book-level errors specified in the carefully tested model.

Statistical analysis of the datasets produces a stark portrait of the visual properties of book surrogates in a 10 million volume collection, in which nearly a third of all volumes that exhibit a low level of text-oriented degradation coexist with volumes where severe error cascades through inter-related digitization processes. Minor error that does not limit the readability of digitized text might be accepted as a part of the price of enhanced access. Only a minority of the volumes in HathiTrust are error free at very low levels of severity. The four most common errors (thick text, broken text, warped pages, and obscured content) are easily and reliably detectable and so common as to be part of the fabric of digital surrogacy. With the exception of Asian language text digitized by Google, near fatal errors largely exist randomly and in very small proportions in the corpus of HathiTrust volumes digitized by Google and the Internet Archive. Extremely severe error, however, compromises the integrity of large-scale digitization and threatens the long-term trustworthiness of repositories that preserve digital surrogates. The findings from one aspect of a multi-faceted investigation into the quality of the digital surrogates suggest that the imperfection of digital surrogates is a transparent and nearly ubiquitous attribute, one that reflects the flaws of the source and introduces new complexity in preservation repositories.

Additional details about the project, including its metrics and progress reports, may be found on the project's website: http://hathitrust-quality.projects.si.umich.edu/ For information on HathiTrust Digital Library, see: http://www.hathitrust.org

References

Cohen, D. (2010). Is Google Good for History? Dan Cohen's Digital Humanities Blog. Posting on 12 Jan. 2010. http://www.dancohen.org/2010/01/07/is-google-good-for-history/ (accessed 8 March 2013).

Conway, P. (2011). Archival Quality and Long-term Preservation: A Research Framework for Validating the Usefulness of Digital Surrogates, *Archival Science*, 11, 3. Open access online.

Conway, P. (2012). Validating Quality in Large-Scale Digitization: Selective Findings on the Distribution of Imaging Error. In *Proceedings of UNESCO Memory of the World in the Digital Age*, September 26-28, 2012, Vancouver, BC Canada.

Conway, P and J. Bronicki (2012). Error Metrics for Large-Scale Digitization. In *Curating Quality: Ensuring Data Quality to Enable New Science*: An invitational workshop sponsored by the National Science Foundation, September 10-11, 2012, Arlington, VA USA.

Eaves, M. (2003). Graphicality: Multimedia Fables for 'Textual' Critics. In Bergmann-Loizeaux E., and N. Fraistat (eds.) *Reimagining Textuality: Textual Studies in the Late Age of Print*, Madison: University of Wisconsin Press, 99-122.

Henry, C. (2010). *The Idea of Order: Transforming Research Collections for 21st Century Scholarship*. Washington, DC: Council on Library and Information Resources. http://www.clir.org/pubs/abstract/reports/pub147 (accessed 8 March 2013).

Knight, S. (2008). *User Perceptions of Information Quality in World Wide Web Information Retrieval Behaviour*. Ph.D. thesis, Edith Cowan University.

Lin, X. (2006). Quality Assurance in High Volume Document Digitization: A Survey. In *Proceedings of the Second International Conference on Document Image Analysis for Libraries* (DIAL'06), 27-28 April, Lyon, France, 319-326.

Saracevic, T. (2007). Relevance: A Review of the Literature and a Framework for Thinking on the Notion

in Information Science. Part III: Behavior and Effects of Relevance, *Journal of the American Society for Information Science and Technology*, 58, 13: 2126-2144.

Taussig, M. (1993). *Mimesis and Alterity: A Particular History of the Senses*. Routledge, London.

Terras, M. (2011). Artefacts and Errors: Acknowledging Issues of Representation in the Digital Imaging of Ancient Texts. In Fischer, F., Fritze, C. and Vogeler, G. (eds), *Kodikologie und Paläographie im digitalen Zeitalter 2 / Codicology and Palaeography in the Digital Age 2*. Norderstedt, Germany: Books on Demand, 43 - 61. http://discovery.ucl.ac.uk/171362/ (accessed 8 March 2013.

York, J. J. (2010). Building a Future by Preserving Our Past: The Preservation Infrastructure of HathiTrust Digital Library. In *Proceedings of 76th IFLA General Congress and Assembly*, 10-15 August 2010, Gothenburg, Sweden.

Uncovering Reprinting Networks in Nineteenth-Century American Newspapers

Cordell, Ryan
r.cordell@neu.edu
Northeastern University, United States of America

Maddock Dillon, Elizabeth
E.Dillon@neu.edu
Northeastern University, United States of America

Smith, David
dasmiq@gmail.com
Northeastern University, United States of America

Our paper describes *Uncovering Reprinting Networks in Nineteenth-Century American Newspapers*, a project to develop theoretical models to help scholars better understand what qualities—both textual and thematic—helped particular news stories, short fiction, and poetry "go viral" in nineteenth-century newspapers and magazines. Prior to copyright legislation and enforcement, literary texts as well as other non-fiction prose texts circulated promiscuously among newspapers as editors freely reprinted materials borrowed from other venues. What texts were reprinted and why? How did ideas—literary, political, scientific, economic—circulate in the public sphere and achieve critical force among audiences? By employing and developing computational linguistics tools to analyze the large textual databases of nineteenth-century newspapers newly available to scholars, this project will generate new knowledge of the nineteenth-century print public sphere.

Although digital databases of nineteenth-century U.S. newspapers are currently keyword searchable, such searches do not allow one to discover frequently reprinted texts of which one is not already aware. To find reprinted texts with currently available online tools, a scholar must first know that a text was reprinted, and then laboriously search for witnesses using a battery of search terms across a range of nineteenth-century periodicals archives. Scholars can easily find more witnesses of texts they already know were popular during the nineteenth-century, but the vast majority of viral texts, since lost to scholarship, remain undiscovered, buried amid millions of scanned pages. This limitation is especially dangerous because it tends to reinforce a scholar's existing suppositions—on, say, the dominance of a text that is canonical today—while leaving undiscovered the more popular text that is no longer on our radar and that might reveal to us precisely what we have failed to understand about popular opinion, reading habits, and public debate in the period. In other words, the sheer time it takes to trawl through digital archives can reify ideas of canonicity rather than expand our ideas about which texts should be central to academic inquiry.

Our paper will describe how we seek to redress the limitations of keyword searching to uncover previously-lost reprinted texts. We are developing models and algorithms to compare repetitions of words and n-grams (contiguous sequences of n words) between documents in a large corpus. We employ space-efficient n-gram indexing techniques to identify candidate newspaper issues and then exploit local models of alignment (in contrast to whole-sequence models used in much duplicate detection work) to identify the boundaries of reprinted passages, which are not known a priori. Probabilistic alignment techniques also provide greater robustness in the presence of optical character recognition errors. We then group pairs of matching passages into larger clusters of text reuse. This all-to-all comparison automatically uncovers duplicated passages: both shorter quotations and fully reprinted texts.

We have already seen significant success applying these techniques to the Library of Congress' Chronicling America archive (http://chroniclingamerica.loc.gov/). In our initial tests, the archive yielded tens of thousands of potential reprinted texts among its publications. Most of these discovered texts are unfamiliar to us, indicating that the archive does include many, many popular texts that have been for years outside the purview of period scholars. This early success points to enormous potential for the project for scholars both in the digital humanities and nineteenth-century American studies. This project also capitalizes on

the Library of Congress' investment in creating an open, accessible record of American print culture, using the data it compiled to develop new knowledge about an important moment in both print and national history. Once honed, these approaches also could be applied to other digitized sources—e.g., materials from the American Memory and Internet Archive magazine and book collections—to uncover patterns of textual reuse and reappropriation in other periods and places.

Our paper will also describe new models we are developing to characterize reprinted texts using both internal and external evidence. We augment models of the linguistic features of reprinted texts with features of the political, social, religious, and geographic affinities of the venues where they appeared, and evaluate the effectiveness of both these components by manually annotating collections of reprinted texts. Using GIS, for instance, we have begun to map our discovered print histories onto historical spatial and social data in order to outline the communities through which particular texts traveled. Historical census data can illuminate who may have read any given publication based on where it was reprinted. By analyzing community data around reprinted texts, we are testing what social variables may have affected textual virality during the antebellum period.

Our ultimate goal for *Uncovering Reprinting Networks* is to construct models to describe nineteenth-century virality: what textual, social, or other factors contributed to the success of a text in the antebellum newspaper, magazine, or book? By recovering a corpus of antebellum-period, viral texts—many of them previously lost to scholarly view—from within the vast Chronicling America collection, we will give scholars of the period a new window into the values and priorities of antebellum editors and readers. Importantly, this granular data about the nature of frequently circulated texts and the paths of their circulation will enable us to understand the shape and constraints of the public sphere, the development of which was key to nineteenth-century U.S. history, including the democratic extension of the franchise, antebellum sectionalism, the abolitionist movement, and the westward growth of the nation.

Scholarly Open Access Research in Philosophy: Limits and Horizons of a European Innovative Project.

Cristina, Marras
cristina.marras@cnr.it
National Research Council: Institute for European Intellectual Lexicon and History of Ideas (CNR-ILIESI)

Antonio, Lamarra
agora.coordinator@iliesi.cnr.it
National Research Council: Institute for European Intellectual Lexicon and History of Ideas (CNR-ILIESI)

Abstract

The paper concerns the development and achievement of the European project *Agora: Scholarly Open Access Research in European Philosophy*. This project aims at improving the spread of European research results in the field of philosophy, while advancing new paradigms of Open Access (OA) publishing, peer-reviewing, and rendering thanks to the interlinking, connecting, and commenting of digital versions of primary sources (manuscripts, original printed editions, critical editions) and secondary literature (articles, monographs, audio and video contributions). In particular, the paper presents and discusses the experiments on semantic linking and the open peer review experiment, their results and some critical aspects and open issues.

1 Introduction

At the present time, in the Humanities and the Social Sciences it is still unusual to have access to networks of interrelated infrastructures offering large datasets so arranged and encoded to allow easy comparative research as well as to encourage the online publication of its outcomes for scholarly and learning purposes.[1] This paper deals with developments and achievements of the European project *Agora: Scholarly Open Access Research in European Philosophy*,[2] which aims at improving the spread of European research results in the field of philosophy, while advancing new paradigms of Open Access (OA) publishing, peerreviewing, and rendering thanks to the interlinking, connecting, and commenting of digital versions of primary sources (manuscripts, original printed editions, critical editions) and secondary literature (articles, monographs, audio and video contributions).

In this presentation we will first offer an overview of the semantically structured digital libraries set up by the Agora consortium of partners, which include extensive and coherent collections of high quality OA contents from classical to contemporary European philosophy and a large

selection of the related critical literature. We will then concentrate on some innovative aspects of the OA archiving and publishing which characterize the Agora project as regards the rendering of texts (in fac-simile as well as in transcription) and the need of ensuring the requested scholarly credibility of their digital editions.

After a description of the federation of textual archives, which constitute the Agora portal, we will briefly present the five experiments envisaged in the project, namely: 1. Semantic linking; 2. Linked Open Data (LOD); 3. Advanced Scholarly linking and rendering; 4. Open peer review; and 5. OA business models in the field of European philosophy. We will focus, in particular, on the experiments on semantic linking and on the open peer review experiment, on their results and some of their critical aspects.[3]

2 The Portal Federation

The digital libraries (Portal Federation) created thanks to the Agora project (in continuity with the previous project *Discovery*)[4] represent a coherent collection of philosophical datasets both complementary and fully interoperable with general-purpose European-wide digital libraries and other content aggregators. These datasets offer the basis for the scholarly output and include, among others, testimonies about Socrates and Socratics, texts by Pre-Socratics, Diogenes Laertius, and Sextus Empiricus, Giordano Bruno, René Descartes, John Locke, Gottfried Wilhelm Leibniz, Giambattista Vico, Alexander Gottlieb Baumgarten, Immanuel Kant, Friederich Nietzsche, and Ludwig Wittgenstein. These philosophers are mostly studied in the Philosophy departments, although many of them are also widely studied by researchers working on history of science and epistemology.

The collections of texts are organized in several platforms managed by open source software; four of them are dedicated to primary sources namely: *Ancient Philosophy* and *Modern Philosophy* (both available within the portal *Daphnet*, i.e. *Digital Archives of PHilosophical Texts on the NET*), *NietzscheSource* and *WittgensteinSource*.[5] Moreover, large collections of selected secondary literature concerning these primary sources are stored in additional OJS (Open Journal System) platforms, and finally two online journals for newly produced scholarly literature are also included in the Portal: *Lexicon Philosophicum: International Journal for the History of Texts and Ideas* and *The Nordic Wittgenstein Review* (see *infra* Section 4).[6]

The Agora collection can really be seen as one of the most extensive and freely available collection of primary and secondary source for scholars and students in the field of philosophy.

The software infrastructure, built up according to the semantic web standards, coherently integrates the functions of an *archive* for the consultation of primary sources (which in the Humanities is equivalent to raw data), a for the consultation of research output, and a for peer review and publication of new research in dedicated digital journals and monographs.

All scholarly information made accessible by the Agora project is first marked up in order to establish links to the underlying data sources, which are available in the federation, and subsequently published in the three new online OA journals created by the Agora consortium partners. The above mentioned functions (i.e. that of archive, library, and publisher) together provide the critical mass of content which is required to propose an innovative set of solutions for the Humanities and also to carry out experiments to test these solutions.

3 The semantic linking experiment

In general, the goal of semantic linking experiment of the project Agora[7] is to test the possibility of a novel way of building up, querying, and browsing a knowledge network in relation to a given set of primary and secondary scholarly sources and to assess its suitability as a collaborative research tool as well as a learning device. More specifically, this experiment aims at analysing and testing procedures, protocols and tools to enrich with semantic information the textual corpora (or collections of texts) included in the various platforms of the portal. To this aim, any individual document can be put in relation either with another document (internal or external to a given platform), or with an element belonging to a class as defined in the reference ontology.[8]

The experiment will focus on two main cases: the relations between primary and secondary sources, and the tagging of primary sources with reference to a selected list of the most relevant philosophical subjects. In the first case, a subset of the critical literature (scholarly articles, monographs) included in a specific OJS platform is semantically linked to a related dataset (manuscripts or editions of primary sources), which is already available online (and can be either internal or external to the portal). We will call this kind of relationship a 'Text-to-Text interlinking', whereas we will call a 'Text-to-Subject interlinking' the case of a semantic connection between a document and a specific subject included in one of the subclasses which constitute the class of philosophical subjects in the reference ontology of the portal.

As a matter of fact, such a vast research area will be covered in different but complementary ways by three partners of the project. Indeed, while the CNRS-ITEM team in Paris will work only on Text-to-Text relations, the UIB-WAB team in Bergen as well as the CNR-ILIESI one in Rome will consider both the Text-to-Text and the Text-to-Subject interlinking. In the specific perspective of the CNRS-ITEM team, the relationship between secondary and primary sources will be envisaged and a dedicated software will be developed — called *Contexta* — aiming at establishing and managing bi-univocal correspondences between them, so that, given a document (primary source), it will be possible to know at the same time not only the list of the quoted external sources but also the list of the external sources quoting it. As mentioned above, the Norwegian and the Italian teams will also consider the Text-to-Subject interlinking but, in their turn, they will use different methodological approaches. Whereas WAB will go on with making use of *SWickyNotes* (a software developed in the framework of the previous European project Discovery), CNRILIESI will test a novel semantic annotation tool, called *Pundit*.[9]

We will now focus on the CNR-ILIESI contribution to the semantic linking experiment.[10] A result of another European project, called *SemLib*,[11] *Pundit* is a newly produced software specifically dedicated to establishing relations between texts and semantic information with reference to a domain ontology. It will be tested on a selected corpus of philosophical writings covering both the ancient and the early modern philosophy and belonging to the *Daphnet* portal, which includes works by the Greek sceptical philosopher Sextus Empiricus and the German philosopher and mathematician G. W. Leibniz. In this case, the experiment will aim at semantically interlinking these texts with a number of scholarly contributions stored in the same portal as well as at semantically enriching them with reference to a number of philosophical subjects (or themes) offered by the domain ontology. The basic strategy of this tool consists in creating a triple for each individual information that has to be recorded and stored, each triple consisting in its turn of two terms (a subject and an object) connected by a relation. In the case of a Text-to-Text relation, both subject and object of the relation are texts (or better, portions of texts), while in the case of a Text-to-Subject relation its object is an element of the list of philosophical subject included in the appropriate class of the portal ontology.

Having set up our sample of texts, the second step has been to define a number of relations, which would be useful for the creation of the two types of triples we were looking for, i.e. the triples expressing relations between texts and those classifying texts according to the most relevant philosophical subjects. Clearly, in the meantime two lists of philosophical subjects were also needed: one including the most relevant subjects for tagging Sextus' writings, another one for tagging Leibniz's writings. At the present moment, a list of eleven relations (and their inverse) has been selected and defined to classify Text-to-Text relationships, which includes: 'quotes', 'paraphrases', 'refutes', 'argues for', 'explains', 'criticizes', 'agrees with', 'interprets', 'is similar to', 'makes a reference to', 'makes an internal reference to'.

By means of the *Pundit* software it is then possible to annotate, for instance, a given passage of a monograph dealing with a specific paragraph of Leibniz's *Monadology*, simply by selecting on the screen both the texts to be related and choosing the most appropriate relation in the list. In this way, the user might for instance create a triple of this kind: '<lmonograph A, p. n> *explains* <Monadology, § n>'. This and its inverse relation ('<Monadology, § n > *is explained by* < monograph A, p. n >') would be stored among the annotations connecting the primary source *Monadology* and a given monograph dealing with it.

As regards philosophical subjects, two lists were necessary, since we decided that it would be preferable to express them in the same language of the primary sources. Consequently, we selected a preliminary list of ancient Greek subjects for the works of Sextus (including approximately 200 subjects), and a preliminary list of French subjects for Leibniz's works (including approximately 650 subjects). On these basis, two small teams will be at work in order respectively to annotate both Sextus' and Leibniz's texts and to edit the final lists of relations and subjects; in this respect in particular, the merging is expected of the two lists of subjects into only one, which would contain pairs of Greek-French subjects, when a semantic equivalence can be stated. In addition, a list of relations possibly connecting texts and subjects has been set up, which includes the following five relations (and their inverse): 'defines', 'indirectly defines', 'extensional instantiation', 'intensional instantiation', 'dealings'.
Also in the case of the Text-to-Subject interlinking, semantic annotation would require the creations of triples connecting a given text with a subject by means of the appropriate relation. Thus, for instance, the user might select a paragraph of Leibniz's *Monadology* and connect it to the subject 'action' thanks to the triple: '<Monadology, § n > *defines* < action >', which is equivalent to the inverse triple '<action > *is defined by* < Monadology, § n >'.

As a result, once the texts annotated, it will be possible: (a) to move from primary sources to the relevant secondary sources (and vice-versa) according to the different relations connecting a philosophical texts to their relevant scholarly literature, and (b) to locate and select passages from the annotated texts according to their relevance for the main philosophical subjects contained in the reference domain ontology either by using a Greek or a French key-

subject indifferently. Insofar as the experiment will provide satisfactory results, we will enlarge both the textual dataset and the list of subjects, our final goal being to set up a multilingual structured thesaurus of philosophical subjects in seven languages, namely ancient Greek, Latin, Italian, French, Spanish, English, and German.

4 The peer review experiment

Especially in Europe, peer reviewing in the *narrow* sense is not largely used in the Human and Social Sciences (SSH), because the book (monograph, edition) is, with some exceptions in local and specialized communities, the main publishing tool, while journals represent a minimal fraction of the whole amount of scientific publications, since articles account for only 20% to 35% in the Humanities, depending on the discipline (Simba Information, 2010). When eatablishing editorial series, the initiators (usually the publisher) establish scientific committees, whose members carry out an initial quality control; most of the time it is the author, who generally pays for the printing of the book with funds granted by his/her institution or other funds. Journals are mainly funded through subscriptions; usually, in philosophy the Impact Factor is not calculated, and bibliometric techniques are not used; niche journals in foreign languages (academic communities being national) are very numerous, and there is a wide discretion in assessing their relevance.

Scientific journals have been in existence for over 400 years and since the appearance of the first learned journals, as for example the *Journal des Sçavans* which first issue is dated 5 January 1665, or the *Philosophical Transactions of the Royal Society*, first issue dated 6 March 1665, some level of peer review were set in place (Harcourt, 1972; Biagioli, 2002; Johns, 1998). Nevertheless, some journals do not use peer reviewing in the true sense of the expression, but rather directly request articles from authors. In many cases, peer review in the Humanities is realized ex post (after publication), and often through reviews.

Various forms of "open review" has been advocated and experimented by scholarly journals most of them in medicine, biomedicine, hard science. In the last decade the benefits and the advantages as well as the limits and defects of the open peer review as been largely discussed.[12]

The discussion on the changing nature and role of peer review in SSH is almost absent and only recently started to be an issue (British Academy, 2007). The debate grows along with the impact and the influence of electronic journals and the use of internet in conducting peer review. The electronic procedures respond to the necessity of more flexibility, transparency, and non-expensive, fast, and more reliable and objective evaluation (Salomon, 2007).

The "Open Collaborative Peer Review Experiment"[13] conducted by the project Agora wants to contribute to this debate and investigate whether the quality and the credibility of a journal and its authors can be assured also by an articulated peer review process under the "responsibility" of the scientific community and not by quantitative evaluation criteria.

The experiment combines traditional double-blind peer review with different modes of open peer review and post-review comments by other scholars. The goal of the experiment is to enhance and determine standards for open peer-reviewing in Humanities and Social Sciences. The experiment is designed to contribute to the current discussion concerning the value of the peer review and the different ways to approach it. The experiment is conducted on two new OA international Journals: *Lexicon Philosophicum: International Journal for the History of Texts and Ideas* and *The Nordic Wittgenstein Review*.[14] The first, published by the CNR-ILIESI, focuses on the history of philosophy and the history of ideas, with a special attention to textual and lexical data (the journal also includes a specific section devoted to Digital Humanities). The second one, is published by the Nordic Wittgenstein Society (NWS) and is dedicated to all aspects of Wittgenstein's thought and work.

The peer review procedure followed by the two journals consists in two main phases: *refereeing* and *reviewing*, and is based on a four-stages process. During the first stage, the submitted manuscripts go to the general eligibility and quality check made by the editors. The editorial procedure is designed to follow best practice concerning excellence, impartiality, transparency, purposefulness, efficiency, confidentiality and integrity. The general eligibility is followed by the acknowledgement of the results to the authors, and in case of ineligibility sufficient information describing the decision is communicated. In the same case resubmission is encouraged.

During the second stage, eligible articles go through a double-blind peer review process in which the identity of reviewers and authors is kept confidential from each other. During the third stage, accepted articles for publication go in pre-print form for one-month of open peer-reviewing or commentary, during which registered users are asked to comment on and to discuss the accepted papers. Discussions are moderated by editors and editors-inchief. Authors are therefore able to take benefit from suggestions, comments and discussions from the open peer review in addition to the comments received by the reviewers when finalizing their paper for publication.

During the fourth stage (the post-peer review), published articles are available in the appropriate Agora platform for ranking and commenting. Such a process is designed to improve the quality of the review and the peer review

process but it also wishes to contribute as much as possible useful suggestions to the authors in the final preparation of their publications.

This procedure would probably present less elements of originality if it were not accompanied by a constant activity of monitoring and analysis. The Agora project in fact includes a Work Package (WP2) specifically dedicated to "Monitoring and evaluating Agora's experiments".[15] An accurate analysis and evaluation of the open peer review for the first two issues of the two journals, including the double-blind phase, is carried out and reported to the Consortium and the EU officers and reviewers[16] by a dedicated team of experts who closely works with the editorial boards of the journals. One aspect of this evaluation involves the various parties who plays a role in the reviewing process such as authors, reviewers, editorial boards, and editors-in-chief via interviews and a survey, which aims not at detecting only positive comments but at identifying critical aspects to reflect upon in order to undertake remedial actions.

The results of the first phase of the evaluation stressed the following points: editors had realised that the peer reviewers needed more guidance concerning how to conduct and complete the review process and to remedy the situation; in addition, they agreed that instructions for reviewers and review forms needed to be clear and very detailed. The scholarly community is somehow reticent to comment openly and in most of the cases preferred to do it via email directly addressing the authors.

The Editorial Boards of the two journals considered also implementing a triple-blind review, in which even the editors would not know the identity of the authors, but in practice, it was not possible to test such an hypothesis. Very often the authors contacted the editors in advance with questions; on the other hand, a triple-blind review would require an editorial apparatus heavier than the one a journal of small dimensions can handle.

The Agora Open Collaborative Peer Review Experiment is to some extent still problematic, because of some cultural resistances still active in the field, not to mention the intrinsic limits inherent to the search for an objective evaluation. On the other hand, it is a possible answer to some of the questions emerging in the European (and international) debate concerning peer review and research evaluation to be transferred in the SSH, and especially an answer to the need of integrating policies *and* practices *into* coherent procedures by involving in a transparent and constructive way the scientific community in the process of building up knowledge (Marras, Ranjbaran, 2011).

The open collaborative peer review experiment is a way of engaging communities in reflecting on and refining the peer review process, even if it does not always result into open peer review. Nevertheless, it contributes to a better understanding of what a good peer review can be in philosophy and, more broadly, in the Humanities.

Moreover, the open collaborative peer review experiment and its evaluation want to contribute to frame peer review as an intellectual subject and to activate research on it by promoting empirical analysis (practices) as well as philosophical reflection about the conditions of possibility of the academic knowledge (Biagioli, 2002, cit.)

5 Conclusion

In drawing the conclusions we would like to stress a few points and some open issues.

1. The synergy between the solutions developed for the Agora project (based on OA and interoperability) as well as the variety of perspectives and approaches it supports are key for ensuring that platforms are not simply content aggregators but mostly rich and attractive collaborative environments capable of managing exchanges between languages and traditions.

2. The multilingual Agora portal will serve as a unique access point to the philosophical content of the federation and will expose its metadata to the LOD cloud. This integrated pilot will serve as a model for establishing and accessing a growing open research archive for scholarly publications in Humanities and Social Sciences as a whole.

3. The integrate infrastructure would enable scholars, university students and teachers to improve knowledge transfer and sharing, to promote remote learning and distance collaboration, thus fostering the development of critical thinking skills and autonomous production of scientific contributions. In the meantime, the associated research publication system would enable a de-localized community of specialists to work in a cooperative way, to publish the results of this work on the Internet, and to establish collaborative initiatives with colleagues based in other universities and research centres.

It will therefore be useful to investigate limits and potentiality of the model proposed by Agora, and see in which way it can still improve the spread of research results in the field of Humanities and Social Sciences and advance new paradigms of knowledge and knowledge organization.

Basing on the results achieved in the project, we can summarize a few issues, which are still open:

1. The sustainability of particular publishing practices in philosophical research,

2. The economic dimension of philosophical journals and publications in Humanities *vis-ávis* that in the hard sciences,

3. While peer review and edition costs are extremely high in the hard sciences, there remains a strong tradition of voluntary work in the Humanities, which can be easily

transferred to the digital world in order to contribute to a viable economic model.

In conclusion, we would like to stress that what inspired the construction of the Agora federation portal is not simply the creation of a digital library but the desire to set up and offer to the scholarly community a digital environment designed for research. Users should access not only classical philosophical texts (primary and secondary sources, journals) but a very large set of the available features and tools (tags, annotations, interlinking, advanced search, *etc*.), which are needed to foster improvements and advancements in the analysis and interpretation of philosophical problems.

References

Biagioli, M. (2002). From Book Censorship to Academic Peer Review. *Emergences: Journal for the study of media & composite culture*. 12(1): 11–45. DOI:10.1080/1045722022000003435.

British Academy (2007). *Peer Review: The Challenges for the Humanities and Social Sciences*. http://www.britac.ac.uk/policy/peer-review.cfm (accessed 13 March 2013).

ESF (2011). *Research Infrastructures in the Digital Humanities. Science Policy Briefing*, n. 42. European Science Foundation, September 2011. Available at: http://www.esf.org/research-areas/humanities/publications.html (accessed 13 March 2013).

ESFRI (2010). *Road Map*. http://ec.europa.eu/research/infrastructures/pdf/esfristrategy_ report_and_roadmap.pdf (accessed 13 March 2013).

Garshol, L. M. (2004). *Metadata? Thesauri? Taxonomies? Topic Maps! Making sense of it all*. http://www.ontopia.net/topicmaps/materials/tm-vs-thesauri.html#N773 (accessed 14 March 2013).

Grenon, P., Smith, B. (2009). Foundations of an ontology of philosophy, *Synthese* 182, 2011:185–204.

Harcourt, B. (1972). History and the Learned Journal. *Journal of the History of Ideas*, 33(3): 365-378.

Johns, A. (1998). *The Nature of the Book. Print and Knowledge in the Making*. Chicago: Chicago University Press.

Marras, C., Ranjbaran, F. (2011). ESF Member Organisation Forum on Peer Review, *European Peer Review Guide – Integrating Policies and Practices into Coherent Procedures*, Strasbourg. http://www.esf.org/activities/mo-fora/peer-review.html (accessed 13 March 2013)

Salomon, D. (2007). The role of peer review for scholarly journals in the information age. *Journal of electronic publishing*, 10:1, Winter 2007: http://dx.doi.org/10.3998/3336451.0010.107 (accessed 13 March 2013).

Simba Information (2010). *Social Science and Humanities Publishing 2009–2010*, Market Research, Simba Information.

Su X., Gulla J. A. (2004). Semantic Enrichment for Ontology Mapping. *Natural Language Processing and Information System*. Lecture Notes in Computer Science. Ed. by Meziane and M´etais, vol. 3136: 217-228, Springer Verlag, Berlin-Heidelberg. http://link.springer.com/chapter/10.1007%2F978-3-540-27779-8_19 (accessed 14 March 2013).

Notes

1. An interesting document reflecting on the centrality of Research Infrastructures to the Humanities and drawing on a number of case studies has been published by the European Strategy Forum on Research Infrastructures (ESF, 2011). See also: ESFRI, 2010.
2. The project is a CIP-project co-funded by the European Commission under the “Information and Communication Technologies Policy Support Programme” (ICT PSP); on this regard see: http://ec.europa.eu/information_society/activities/ict_psp/index_en.htm . The Project reference number is 270904, CIP-ICT-PSP.2010.2.5 - Open access to scientific information; see: http://ec.europa.eu/information_society/apps/projects/factsheet/index.cfm?project_ref=270904 . The ICT PSP program aims at stimulating a wider uptake of innovative ICT-based services and the exploitation of digital content across Europe by citizens, governments and businesses. Details of the project are available in the project web site: http://www.project-agora.org . The Agora’s consortium involves six European countries, and some of their major research institutions and leading software companies including the Institut des Textes et Manuscripts Modernes of the Centre National de la Recherche Scientifique (ITEM-CNRS), France; The Wittgenstein Archives at the University of Bergen (UIB-WAB), Norway; the Åbo Akademi University, Finland; the University of Copenhagen (KU), Danemark; the Net7 company, Italy; the Istituto per il Lessico Intellettuale Europeo e Storia delle Idee of the Consiglio Nazionale delle Ricerche (CNR-ILIESI), Italy, which also coordinates the project. Antonio Lamarra is the project coordinator.
3. In writing this paper we took benefit from consulting a number of documents and reports prepared by the partners of the consortium. We are grateful to all of them and in particular to Anna Maria Carusi (KU), Paolo D’Iorio (CNRS-ITEM), Yrsa Neuman (Åbo Akademi University), Alessio Piccioli (Net7), Alois Pichler (UIBWAB).
4. For more details conserning the *Discovery* project see: http://www.discovery-project.eu/home.html .

5. See: www.daphnet.org (which includes two platforms of primary sources: http://modernsource.daphnet.org/ , under the responsibility of Roberto Palaia (ILIESI-CNR), and http://ancientsource.daphnet.org/ , under the responsibility of Emidio Spinelli (University of Rome, *Sapienza*); www.nietzschesource.org , under the responsibility of Paolo D'Iorio (CNRS-ITEM) and www.wittgensteinsource.org, under the responsibility of Alois Pichler (UIB-WAB). A platform dedicated to the French philosopher J.J. Rousseau (*Rousseausource*) is currently under preparation. The platforms are built under open source software: *Daphnet* and *Wittgesteinsource* are managed by MURUCA (http://muruca.netseven.it), a framework for building digital libraries that allows to edit and enhance digital material and it can be used also as a research tool. The "Nietzsche facsimile edition" (www.nietzschesource.org/DFGA) is managed by TALIA (http://net7sviluppo.com/trac/talia/wiki/TaliaTutorial), and the "Nietzsche critical edition" is managed (www.nietzschesource.org/eKGWB) by a combination of MySQL and PHP.
6. A third journal, *Studia Nietzscheana*, is currently in preparation.
7. The Semantic Linking Experiment is carried out within the activities of the Work Package 5: "Scholarly interlinking & rendering experiment", leaded by Paolo D'Iorio (CNRS-ITEM). This experiment, semantically linking primary and secondary source takes into account the specific needs of researchers working in the field of philosophy and it is conducted by CNR-ILIESI, and by UIB-WAB, in collaboration with ONTOS Verlag.
8. On semantic enrichment and discussion about ontologies you can see: Garshol, 2004; Grenon, Barry, 2009; Su, Gulla, 2004.
9. For more details and demos on Pundit see: http://thepund.it .
10. The experiment is carried out in collaboration with Andrea Costa, Michela Tardella and Francesco Verde.
11. More details on *SemLib* are available at: http://www.semlibproject.eu
12. See for example the debate carried out in *Nature*: http://www.nature.com/nature/peerreview/debate/ (accessed 13 March 2013).
13. The Open Peer Review Experiment is the focus of the Work Package 6, which is carried out by ILIESI and UIB-WAB and leaded by Cristina Marras (CNR-ILIESI).
14. The *Nordic Wittgenstein Review* (editors-in-chief: Yrsa Neuman of Åbo Akademi University, Simo Säätelä and Alois Pichler of the University of Bergen) and *Lexicon Philosophicum* (editors-in-chief Antonio Lamarra and Roberto Palaia, (CNR-ILIESI) are built in OJS and semantically linked to the primary sources published in the federation portal (see above, Sections 2 and 3). Both journals are open access and share the same policy: the individual contributions are made available under the Creative Commons General Public License Attribution, Non-Commercial, Share-Alike version 3 (CCPL BY-NC-SA). The first issue of the *Nordic Wittgenstein Review* is available in print (published by the ONTOS Verlag) and online, whereas *Lexicon Philosophicum* is published only online. See: http://www.lexicon.cnr.it ; http://www.nordicwittgensteinreview.com .
15. The WP2 is leaded by Anna Maria Carusi (KU).
16. See: Agora Project Deliverable D2.3 *Interim Evaluation Report* (June 2012) compiled by Anna Maria Carusi and Giovanni De Grandis (KU), document for internal use of the Consortium available upon request.

On Our Own Authority: Crafting Personographic Records for Canadian Gay and Lesbian Liberation Activists

Crompton, Constance

constance.crompton@ubc.ca
University of British Columbia, Okanagan Campus, Canada

Schwartz, Michelle

michelle.schwartz@ryerson.ca
Ryerson University

Reflecting on her life, Jane Rule suspected that so far as the media was concerned in she was "the only lesbian in Canada" (Martin) the year her novel, *Desert of the Heart*, Canada's first English-language literary lesbian novel was published. Online she is one of the best-represented figures in the Canadian gay and lesbian liberation movement, a movement captured by *Lesbian and Gay Liberation in Canada* (LGLC), an infrastructure pilot project of the Canadian Research Writing Collaboratory (CWRC) at the University of Alberta. After introducing the LGLC project, the issues that underpin our TEI encoding, and the CWRC project's extension of existing personographic records, we, the LGLC project's co-directors, demonstrate the stakes in producing authority records for the online recuperation and discoverability of underrepresented gay and lesbian activism.

Starting Out: A Single and Singular Source Text

The LGLC project reconfigures Donald McLeod's remarkable monograph, *Lesbian and Gay Liberation In Canada: A Selected Annotated Chronology, 1964-1975*, as a TEI-encoded resource within CWRC. The chronology is organized by date and then by location, with each entry neatly summarizing a small moment in history, followed by a bibliography of sources. The book focuses primarily on "self-declared lesbians and gay men and their activities in regard to the forging of lesbian and gay communities and liberation in Canada," and therefore, dedicates most attention to demonstrations, political actions, lobbying, and legal reforms (McLeod viii). As a secondary and supplementary focus, the book notes "artistic and cultural contributions with significant lesbian or gay content" as part of the chronology, and includes three appendices listing lesbian and gay organizations, periodicals, bars, and clubs (McLeod viii-ix). Heterosexuals who were instrumental either in supporting or opposing the gay liberation movement are included, and foreign events are noted if they either had a direct impact on the Canadian gay liberation movement, or featured prominent involvement by gay Canadians. The start date of the chronology corresponds with the formation of the Vancouver-based Association for Social Knowledge (ASK), the first large-scale homophile organization in Canada. The end date coincides with the founding of the National Gay Rights Coalition/Coalition national pour les droits des homosexuels (NGRC/CNDH), the "first truly national coalition of Canadian lesbian and gay groups" (McLeod viii).

We originally planned the LGLC project as a straightforward and independent representation of Donald McLeod's book in HTML and KML, both underpinned by TEI, to be housed on the Canadian Lesbian and Gay Archives website. The LGLC's inclusion within CWRC allows us to capitalize on their infrastructure: a repository and suite of tools dedicated to preserving and showcasing Canadian history and writing. The CWRC infrastructure will cast the LGLC project as an online research and social space — visitors will be able to plot the events of the Canadian gay liberation movement on maps and timelines; read and annotate the text; converse with one another; and contribute their own images and reflections.

The Problem of Representation: Modeling Personhood in TEI and RDF

The LGLC project works within CWRC not solely to recover gay history, but to respond, at the level of code, to the debates that have shaped that history. The LGLC project's initial personographic encoding has been motivated by the movement to recuperate lost lesbian and trans histories. The debates may be familiar: in the 1980s there was a rush to claim biological females who lived as men, like surgeon James Barry, music hall performer Annie Hindle, and jazz pianist Billy Tipton, as lesbians who had not had access to "lesbian" as an identity marker. In the following decade the transgender community rejoined with competing claims that these historical actors are part of trans history (Halberstam). What ought the conscientious encoder do in the face of such temporal specificity? In building a reliable and ethically responsible code, we are mindful of Alan Galey and Stan Ruecker's call to make "a virtue of the entanglement of past and future intentions in any artifact"(421); however, we have still experienced anxiety about TEI's modes of expression. The standards, which for example, only allow four designations for categorizing sex [1], facilitate interoperability, but limit our encoding's expressiveness.

Judicious customization that meets the needs of our host project helps us balance our twin goals of ethical representation of the queer community and conformance to digital humanities community practice. Our mid-term objective is to model the history of the queer community's nonce naming practices through TEI, with the long-term goal of using the CWRC space to allow for user-added folksonomic tagging (Ornelas 231). In the meantime, our aim is to extend our event encoding without resorting to schema customization. CWRC is in the process of defining a wide range of RDF predicates to express the relationship between people in its aggregate data sets. Currently, within an event, encoders can tag time, places, agents of action, and recipients of action. CWRC draws on the TEI encoding to write out RDF triples representing the relationship between incidents, time, and people. CWRC allows encoders to draw from CWRC's local authority records, authority records on the web, to offer alternate identifiers complete with name and URI, or to write a new personographic records. This opportunity for strategic customization certainly helps the LGLC project model personhood with more nuance, and, looking forward, will allow for a more fine-grained articulation of the relationships between all the people CWRC's aggregate datasets.

Authoritative Solutions: Dissemination and Discoverability Online

In addition to our desire to integrate LGLC events with other events in Canadian literary history, we are anxious to make the LGLC records widely discoverable. The Canadian lesbian and gay liberation movement is a much-neglected part of Canadian history, one often overshadowed by Stonewall narratives. We aim to rectify the scarcity and relative invisibility of Canada's gay and lesbian liberation movement online. It is challenging, even when doing traditional research, to draw out all the connections between the organizations, events, and people that make up Canadian LGBT history. The activists in the movement are underrepresented in Wikipedia, and often have only partial records in the Virtual International Authority Files (VIAF), Katalog der Deutsche National Bibliothek (DNB), Freebase, and the Library of Congress Authorities.

CWRC houses its own authority records, which, using linked open data, are connected to VIAF, the DNB, and the Library of Congress, among other authorities. Encoders within CWRC can take their identification of a person further, and create a CWRC-specific authority record for that person, supplementing or even contesting the information housed in other authority records. The actors in LGLC dataset give us an opportunity to code for self-identification over time, correcting VIAF, LC Authorities, Freebase and DBpedia's failure to account for activists' changing identities and multiple names, outside of the prose descriptions that they link to or draw from Wikipedia. As legitimate authority records, CWRC's personographies can serve as the basis for more accurate Wikipedia articles, thus feeding back into VIAF, DBpedia, and Freebase, all of which link to Wikipedia. Encoding practices within the CWRC framework offer a new way to comprehend the relationship between the gender performance, sexual practice, and cultural context embedded in identity naming practices, and provide a way to make those relationships discoverable online.

Our paper shows how the *Lesbian and Gay Liberation in Canada* project ensures the representation of historical queer Canadian experience in online resources. Ultimately, as co-directors of the LGLC project, we offer our code and our data modeling principles as methods to overcome the limitations of extant authority records. In this way, LGLC contributes to the existing feedback, citation, and authority models in a manner that supports the LGBT community and improves online access to often overlooked histories — after all, Jane Rule was not the only lesbian in Canada in the 1960s.

References

Galey, A., and S. Ruecker (2010). How a Prototype Argues. *Literary and Linguistic Computing* 25.4: 405–424.

Ornelas, A. (2010). Queer as Folksonomies. In Greenblatt, E. (ed). *Serving LGBTIQ Library and Archives Users: Essays on Outreach, Service, Collections and Access.* Jefferson, NC: McFarland. 229–239.

Halberstam, J. J. (1998). Transgender Butch: Butch/FTM Border Wars and the Masculine Continuum. *GQL: A Journal of Lesbian and Gay Studies* 4.2: 287–310.

Martin, S. (2007). 'B.C. Novelist Wrote a Cult Classic and Became a Lesbian Role Model'. *The Globe and Mail*.

McLeod, D. W. (1996). *Lesbian and Gay Liberation in Canada: A Selected Annotated Chronology, 1964-1975*. Toronto: ECW Press/Homewood Books.

Notes

1. The four values permitted on the sex's value attribute are *male, female, unknown,* and *not applicable.*

What ever happened to Project Bamboo?

Dombrowski, Quinn
quinnd@berkeley.edu
University of California, Berkeley, United States of America

This paper provides a description and analysis of the trajectory of Project Bamboo, a major humanities cyberinfrastructure initiative in the United States. While previous work has addressed the opportunities, consequences and form of humanities cyberinfrastructure (Svensson 2011; Blackwell and Crane 2009), and the development trajectory of cyberinfrastructure initiatives in the sciences (Ribes and Finholt 2009), very little has been written about Project Bamboo (2008-2012), and how it did and did not reflect visions put forth for humanities cyberinfrastructure. In addition, there has been little discussion of the precipitous drop in community enthusiasm and even awareness of Bamboo, despite its considerable amount of funding.

During its 2008-2010 planning phase, approximately 600 scholars, librarians, and IT professionals from 115 institutions participated in conversations about "enhanc[ing] arts and humanities research through the development of shared technology services". While this broad engagement contributed to the development of a rich picture of the technology needs of humanists and librarians at a wide range of institutions, it resulted in a multiplicity of visions for what Project Bamboo should specifically do to address those needs. By the time the Bamboo phase one technology

implementation proposal was funded in fall 2010, there was already a great deal of unclarity among the planning project participants about whether and how Bamboo was translating the enthusiasm for community building and technical interoperability that emerged in the planning project into concrete deliverables. Two years later, most of the digital humanities community had largely written off Project Bamboo for having failed to produce anything of value, or even make clear what it was attempting to deliver. This paper will elucidate the evolution of Project Bamboo's goals and scope from its inception to its demise, the factors that contributed to its increasingly negative public perception, including an assessment of which form inherent problems for cyberinfrastructure initiatives vs. avoidable tactical missteps.

The Bamboo Planning Project received a $1.4 million dollar grant from the Andrew W. Mellon foundation in March 2008 (towards a total budget of $2.4 million), to carry out a series of five workshops that would lay the foundation for a follow-up proposal geared towards the technical development of infrastructure. The planning project proposal anticipated a total of 200 participants from 65 institutions (Bamboo 2008, 27). Instead, nearly twice that number of institutions applied to participate in the first set of workshops, and participation remained significantly higher than expected throughout the entirety of the planning project. The first workshop—held in four different locations, with a process refined after each iteration—encouraged teams of participants to discuss open-ended questions about the challenges facing the humanities and potentially ameliorating applications of technology. In many cases, the Bamboo workshop was the first time that scholars, librarians and IT professionals from the same institution had met one another, or discussed issues of common concern. These conversations had wide-reaching impact on local campuses, and were among the most well-regarded aspects of the Bamboo Planning Project.

Bamboo's discussion-centric approach to workshops was intended to shape the enterprise IT-oriented vision of shared services and service lifecycles that was at the core of the project's vision. Instead, it quickly became clear that many of the workshop participants—most emphatically, but not limited to, the humanities scholars—felt that the emphasis on tools and services was misplaced. Instead of focusing primarily on identifying services that could best be run centrally by a consortium, and how those services could be integrated with one another, subsequent workshops included tracks for participants interested in improving scholarly networking by connecting people, projects and tools;[1] and for participants interested in sharing faculty stories, exemplary practices, and curricula (both student-oriented, and curricula for improving scholars' fluency with digital tools and resources). The long-term, visionary "Bamboo Program" that gained consensus among participants in June 2009 included a scholarly network; a tool and content guide; a collection of scholars' narratives about the use of digital technologies across the arts, humanities, and interpretive social sciences; a repository for documenting workflows; and an "exchange" where individuals in need of assistance could connect with others who have the necessary expertise. These areas of work were included alongside the infrastructure components: a cloud-hosted service delivery appliance, the establishment of a service lifecycle, and the formalization of partnerships with tool developers and collection holders.[2] The scope was ambitious to the point of risky, but it addressed the stated needs and interests of almost all the participants who were involved throughout the planning project. Rather than simply voicing their support for the work in principle, most participants expressed an active desire for their institutions to be involved in building these aspects of Bamboo in subsequent phases.

Multiple factors contributed to the significant downscoping of Project Bamboo between June 2009 and the funding of the Bamboo Technology Project in September 2010, including the impact of the world financial crisis and a change in program leadership at the Mellon Foundation. The community-oriented aspects of the project—which fueled the interest of many planning project participants—were eliminated. This did not, however, result in a single, unified vision. One area of the project, led by librarians, focused on developing collection interoperability connectors for text repositories. Another area, led by technologists, worked on the development of a centrally-run Java-based services platform that would run core services for identity and access management, policy management, notification, and result caching, and could either run or proxy services that provided scholar-oriented functionality (e.g. morphological analysis, geoparsing, concordance). Groups of technologists also attempted to integrate these infrastructure services into "Work Space" web-based applications (Alfresco ECM and hubZero). A scholar-led group focused on the development of a "corpus tool set" application that could provide an integrated scholar-facing interface for tools that can work together to analyze textual corpora. Deviating from the planning project model that required scholars, librarians and technologists to collaborate, rather than focusing only on the areas of the project that were the most comfortable fit, led to a sense of disconnect approaching animosity between sub-groups. While progress was made in each of these areas, both tactical and strategic issues around internally- and externally-oriented communications made it difficult to convey these accomplishments intelligibly to the digital humanities community. The lack of a clearly articulated scholar or developer use case that would be facilitated

by Bamboo's technical developments exacerbated the impression that Bamboo was not developing anything of use.

As the Bamboo Planning Project illustrated, a high level of engagement from a broad community can invigorate a project, allowing it to strive for significant goals that have wide-reaching impact. However, this level of engagement comes with the risk of alienating a large swath of the community if exigencies force the project to scope down significantly, without the project fostering "spin-off" efforts to address those needs that are no longer in scope. A clearly-stated scope connected to use cases that resonate with the target community is essential. Infrastructure projects face a number of unique challenges around communication and demonstrating progress to a broad community, in large part because their technical deliverables do not provide functionality that itself directly furthers scholarship (e.g. via an innovative search algorithm or a novel text analysis interface). There are multiple approaches to ameliorating this, including working closely with digital humanities developers on pilot integrations or mockups for integration with scholar-facing tools, developing mockups, and providing documentation and a sandbox environment that developers can explore, share, and blog about. Regardless, long lapses in communication jeopardize a project comparably to significant delays in technical deliverables.

At the request of the Mellon Foundation, Project Bamboo began to shut down active development work and transition to a wrap-up phase in December 2012. By the end of March 2013, all "artifacts" from the project will be publicly available via links from projectbamboo.org .

References

Blackwell, C., and G. Crane (2009). Conclusion: Cyberinfrastructure, the Scaife Digital Library and Classics in a Digital Age. *Digital Humanities Quarterly* 3(1).

Project Bamboo. (2008). Bamboo Planning Project: An Arts and Humanities Community Planning Project to Develop Shared Technology Services for Research. January 28, 2008 public document revision. Retrieved at http://www.quinndombrowski.com/sites/default/files/blog/bamboo_planning_project_proposal.pdf

Ribes, D., and T. A. Finholt (2009). The Long Now of Technology Infrastructure: Articulating Tensions in Development. *Journal of the Association for Information Systems*. 10(5). 375–398.

Svensson, P. (2011). From Optical Fiber to Conceptual Cyberinfrastructure. *Digital Humanities Quarterly*. 5(1).

Notes

1. http://quinndombrowski.com/projects/project-bamboo/wiki/w3-implementation-proposal
2. http://quinndombrowski.com/sites/default/files/bamboo/pbw4_discussion_draft.pdf

Academic Migrants: A Digital Discussion of Transnational Teaching and Learning

Donaldson, Olivia
odonaldson@nu.edu.kz
Nazarbayev University, Kazakhstan

1. Overview

This short paper offers a progress report on the first digital humanities project being conducted at Nazarbayev University, a newly established English-language institution in Kazakhstan. During the 2012-2013 and 2013-2014 academic years, international faculty members and Kazakh students have the opportunity to contribute personal narratives of academic migration to an open-access digital database. The ensemble of narratives documents personal experiences of migration while investigating how academic migration impacts teaching and learning on trans/national scales. Putting the digital accounts into dialogue with print research on academic migration and migration more generally, the short paper probes the following questions: How does the collection of digital narratives from the Nazarbayev University campus community enhance or problematize current research on migration? What does the digital collection reveal about the formation of individual and national identities in an era of increased migration and information technology? And, how does the project's digital storytelling framework promote transnational dialogue about teaching and learning, thereby impacting trends in higher education?

2. Context and Content

"Academic Migrants" investigates experiences and impacts of academic migration through digital storytelling. The project focuses on the unique case study of academic migration to Nazarbayev University in Astana, Kazakhstan. Migration for academic purposes is integral to the mission of this English-language university. In 2011, Nazarbayev

University opened its undergraduate programs in partnership with major international institutions of higher learning including the University of Wisconsin-Madison and the University College London. The faculty consists primarily of scholars trained in North America and Europe, while the student body hails from all regions of the vast country. Over the course of the 2012-2013 and 2013-2014 academic years, students and faculty are invited to contribute personal accounts of academic migration to a project website.

This case study of academic migration contributes to developments in the field of migration studies. In the post-war period of the 1940s, the field of migration studies began to take root in the United States, the United Kingdom and other Western European countries. According to Castles and Miller, a great deal of research in the field has and continues to focus on the political and economic causes and consequences of migratory patterns implicating the U.S. and other global powers. Though much of this research has centered on the migration of unskilled laborers, more recent scholarship takes into consideration the migration of skilled workers. In the past five years, the subfield of "academic migration" has emerged to shed light on the experiences of educated individuals migrating because of academic pursuits. A growing number of social scientists are investigating the impacts of the migration of academics from "developing" countries to North American and European nations. An annual conference on the subject began in 2009 (http://www.fbmk.upm.edu.my/ICAMM3/) and an anthology appeared in 2011 (Dervin). Existing studies overwhelmingly focus on what is termed a "brain drain" phenomenon, studying for example the movement of highly educated individuals from Africa to Europe or from former Soviet republics, like Kazakhstan, to Russia. "Academic Migrants" moves away from the "brain drain" framework to present a different narrative of academic migration. In offering a case study of academic migration that hinges on the movement of North American and European scholars to the post-Soviet capital of Astana, this project addresses new trends in both migration and higher education.

3. Methods

Essential to the project is its use of digital storytelling to document and investigate the causes and implications of academic migration on personal, national and transnational levels. The use of first-hand narratives is inspired by the work of scholars like Homi Bhaba, Stuart Hall and Avtar Brah who draw from their personal experiences of migrating from the "global south" to the "global north" to deconstruct problematic North/South and West/East paradigms. These authors' cultural theories of hybrid identity and diasporic space challenge the Western-centric foundation and bi-directional focus of migration studies. Similarly, "Academic Migrants" draws from the personal narratives of students and faculty to challenge problematic West/Rest paradigms in its study of migration to and within Kazakhstan in the post-Soviet period. Participants have the option of contributing narratives in written, oral and/or visual formats, and contributions will be arranged on the project website according to particular themes or questions such as the home/land, language and translation, globalization, and so forth. The general public will have the opportunity to interact with the campus community by responding to the digital stories and contributing to online discussions about topics pertinent to academic migration.

4. Outcomes

In using interactive digital storytelling to document and generate discussion about academic migration, this project probes how life narratives can enhance our understanding of the causes and consequences of migration. Specifically, the ensemble of audio-visual and written narratives explores various reasons for which individuals become academic migrants, how personal experiences of migration overlap and/or vary, and how migration impacts teaching and learning on national and transnational scales. In addition to illuminating the varied experiences of academic migrants at Nazarbayev University and informing our understanding of migration, the digital collection invites public dialogue about migration that extends beyond national and academic frontiers.

References

Bhabha, H. (1994). *The Location of Culture*. London: Routledge.

Brah, A. (1996). *Cartographies of Diaspora: Contesting Identities*. New York: Routledge.

Castles, S., and M. Miller (2003). *The Age of Migration: International Population Movements in the Modern World*. London: Guilford.

Dervin, F. (2011). *Analysing the Consequences of Academic Mobility and Migration*. New Castle upon Thyne: Cambridge Scholars.

Hall, S. (1990). Cultural Identity and Diaspora. In Rutherford, J. (ed), *Identity: Community, Culture, Difference*. London: Lawrence. 222-237.

Bootstrapping Delta: a safety net in open-set authorship attribution

Eder, Maciej
maciejeder@gmail.com
Pedagogical University, Krakow, Poland

Introduction

In non-traditional authorship attribution, the general goal is to link a disputed/anonymous sample with the most probable 'candidate'. This is what state-of-the-art attribution methods do with ever-growing precision. However, it is similarly important to *validate* the obtained results, especially when one deals with a faultily-collected or incomplete reference corpus. This is a typical situation of an 'open' attribution problem: when the investigated anonymous text might have been written by *any* contemporary writer, and the attributor has no prior knowledge whether a sample written by a possible candidate is included in the reference corpus. Then the attributor faces the question whether, supposing that all the contemporary writers were represented in the corpus, the results in fact suggest a different person as the most likely author. A vast majority of methods used in stylometry establish a classification of samples and strive to find the *nearest neighbors* among them. Unfortunately, these techniques of classification are not resistant to a common mis-classification error: any two nearest samples are claimed to be similar, no matter how distant they are.

Given (1) a text of uncertain or anonymous authorship and (2) a comparison corpus of texts by known authors, one can perform a series of similarity tests between each sample and the disputed text. This allows us to establish a ranking list of possible authors, assuming that the sample nearest to the disputed text is stylistically similar, and thus probably written by the same author. However, the calculated distance is usually not followed by an estimation of its reliability. While testing the novel *Agnes Gray* against a corpus of English novels, one will probably have Anne Brontë as the most likely author. However, testing Emily Brontë's only novel *Wuthering Heights*, one is guaranteed to obtain wrong results (because no comparison sample is available), but the ranking of candidates will suggest a most likely author anyway, perhaps another Brontë sister. In a controlled authorship experiment, identifying such a fake candidate is easy, but how can we decide the degree of certainty in a real-life authorship attribution case? Although this problem has been discussed (Burrows, 2002, 2003; Hoover, 2004a; Koppel et al., 2009; Schaalje et al., 2011), we still have no widely-accepted solution. The method introduced below provides a new approach to this problem.

Mater semper certa: Burrows's Delta

Among a number of machine-learning methods used in stylometry, a special place is occupied by Burrows's Delta, a rare example of a made-to-measure technique designed particularly for authorship attribution (Burrows, 2002, 2003). Delta is one of the simplest and, at the same time, one of the most effective methods (as evidenced in a benchmark presented by Jockers and Witten, 2008); described in detail, tested, and improved (Hoover, 2004a, 2004b; Argamon, 2008; Smith and Aldridge, 2011; Rybicki and Eder, 2011).

However, despite obvious Delta's advantages, there are also some drawbacks, shared with a vast majority of state-of-the-art attribution methods. Particularly, Delta relies on an arbitrarily specified number of features to be analyzed. Even if most scholars agree that the best style-markers are the counted frequencies of the most frequent words (MFWs), there still remains the nasty question of the number of words that should be taken into analysis. While some practitioners claim that a small number of function words provide the best results (Mosteller and Wallace, 1964), others prefer longer vectors of MFWs: 100 words (Burrows, 2002), 300 (Smith and Aldrigde, 2011), 500 (Craig and Kinney, 2009), up to even 1,000 or more (Hoover, 2004a). However, a multi-corpus and multi-language study (Rybicki and Eder, 2011) shows that there is no universal vector of MFWs and that the results are strongly dependent on the corpus analyzed.

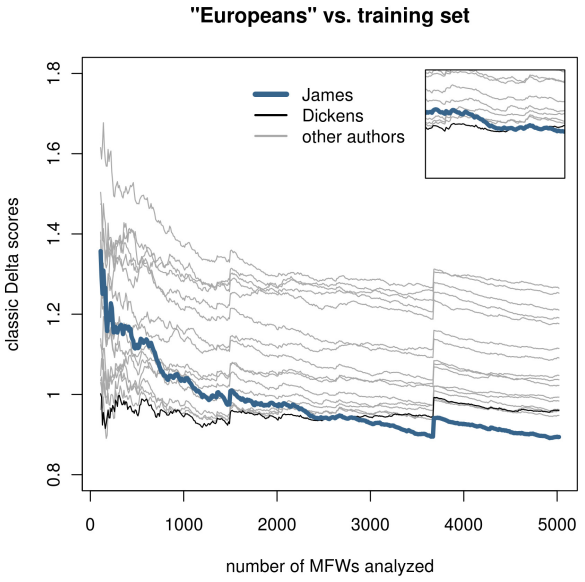


Figure 1.
Who wrote The Europeans? Rankings of 5,000 Delta tests performed on increasing MFW vectors. If the number of MFWs analyzed is lower than 2,700, then Dickens (black thin line) is ranked as the most likely candidate; to get the real author (thick line), one needs to choose a very long vector of words.

What is worse, most attribution techniques, including Delta, allow to measure one vocabulary range (i.e. one vector of MFWs) at once to obtain a ranking of candidates. As shown in Fig. 1, the distances between James's *The Europeans* and the members of the training set are unstable and strongly depend on the number of MFWs tested: e.g. one needs to analyze at least 2,700 words to see James ranked as the most likely author of *The Europeans* instead of Dickens. The danger of cherry-picking is more than obvious here. Although an attributor can perform a series of independent tests with different MFW vectors, a final comparison of thus obtained results is not straightforward at all.

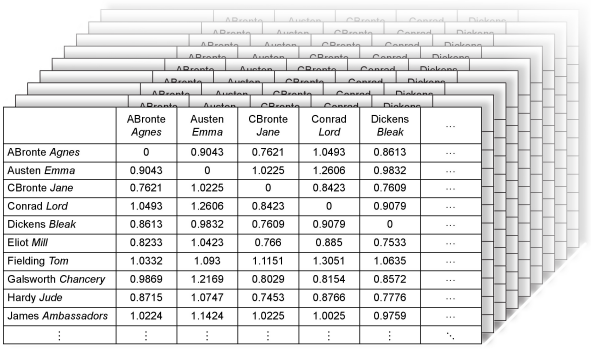
Pater familiae: Bootstrap

In Rudolf Erich Raspe's collection *The Surprising Adventures of Baron Münchhausen*, there is a scene where the main character, trapped in a swamp, pulls himself out by his own bootstraps. Although it is still disputed if Münchhausen was pulling by his boots or by his hair, the bootstrapping became a metaphor for statistical methods that make up for absence or unreliability of parameters with intensive resampling of the original population. The bootstrapping procedures are widely used in biometrics and social sciences, and their idea is quite simple: in a large number of trials, samples from the original population are chosen randomly (with replacement), and this chosen subset is analyzed in substitution of the original population (Good, 2006).

When Delta meets Bootstrap

The aim of the technique presented below is to overcome the disadvantages of the existing nearest neighbor classifications. It has been based on Delta and extended with the concept of bootstrap. The method relies on the author's empirical observation that the distance between samples similar to each other is quite stable *despite* different vectors of MFWs tested, while the distance between heterogeneous samples often displays some unsteadiness depending on the number of MFWs analyzed.

The core of the procedure is to perform a series of attribution tests in 1,000 iterations, where the number of MFWs to be analyzed is chosen randomly (e.g., 334, 638, 72, 201, 904, 145, 134, 762, …); in each iteration, the nearest neighbor classification is performed. It could be compared to taking 1,000 photos from different points of view. Thus, instead of dealing with one table of calculated distances — as in classic Delta — one obtains 1,000 distance tables. Next, the tables are arranged in a large three-dimensional table-of-tables, as visualized in Fig. 2.



	ABronte *Agnes*	Austen *Emma*	CBronte *Jane*	Conrad *Lord*	Dickens *Bleak*	…
ABronte *Agnes*	0	0.9043	0.7621	1.0493	0.8613	…
Austen *Emma*	0.9043	0	1.0225	1.2606	0.9832	…
CBronte *Jane*	0.7621	1.0225	0	0.8423	0.7609	…
Conrad *Lord*	1.0493	1.2606	0.8423	0	0.9079	…
Dickens *Bleak*	0.8613	0.9832	0.7609	0.9079	0	…
Eliot *Mill*	0.8233	1.0423	0.766	0.885	0.7533	…
Fielding *Tom*	1.0332	1.093	1.1151	1.3051	1.0635	…
Galsworth *Chancery*	0.9869	1.2169	0.8029	0.8154	0.8572	…
Hardy *Jude*	0.8715	1.0747	0.7453	0.8766	0.7776	…
James *Ambassadors*	1.0224	1.1424	1.0225	1.0025	0.9759	…
⋮	⋮	⋮	⋮	⋮	⋮	⋱

Figure 2.
Results of 1,000 bootstrap iterations (tables of distances between texts samples) arranged in a three-dimensional table.

The next stage is to estimate the mean and standard deviation of each cell across 1,000 layers of the composite table. This is a crucial point of the whole procedure. While classical nearest neighbor classifications rely on *point estimation* (i.e. the distance between two samples is always represented by a single numeric value), the new technique

introduces the concept of *confidence interval*. Namely, the distance between two samples is a *range of values* represented by the mean of 1,000 bootstrap trials plus $1.64\sigma_{ij}$ below and $1.64\sigma_{ij}$ above the arithmetic mean.

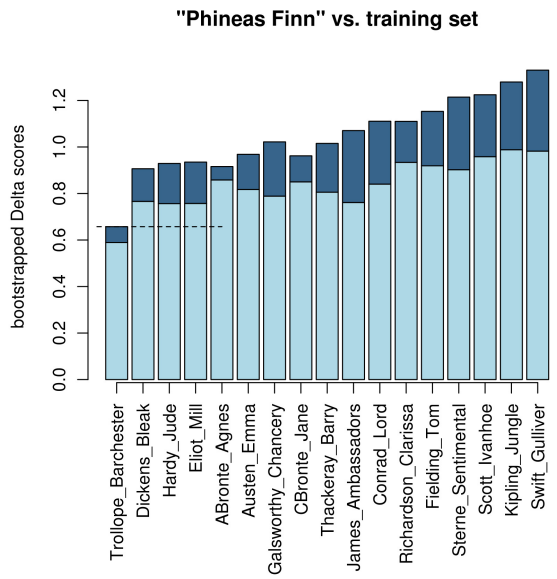


Figure 3.
Who wrote Phineas Finn? Ranking of candidates using confidence intervals.

An exemplary ranking of candidates is shown in Fig. 3. The most likely author of *Phineas Finn* is Trollope (as expected), and the calculated confidence interval does not overlap with any other range of uncertainty. This means that Trollope will be ranked first with a 100% probability.

The real strength of the method, however, is evidenced in Fig. 4 and 5, where *The Portrait of Dorian Gray* is tested against a training set which *does not contain* samples of Wilde. Classic Delta simply ranks the candidates, Hardy being the first (Fig. 4), while in the new technique, confidence intervals of the first three candidates partially overlap with each other. In consequence, the assumed probability of authorship of *Dorian Gray* is shared between Galsworthy (54.2%), Hardy (34.8%) and Charlotte Brontë (11%). The ambiguous probabilities strongly indicate fake candidates in an open-set attribution case.

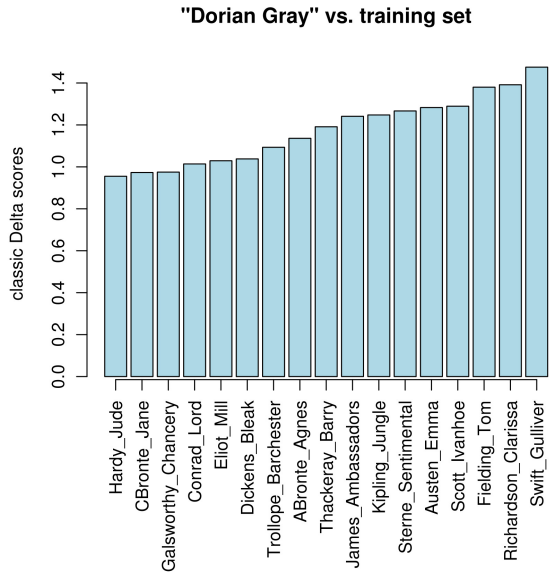


Figure 4.
Who wrote The Portrait of Dorian Gray? Ranking of candidates using classic Delta procedure (500 MFWs tested).

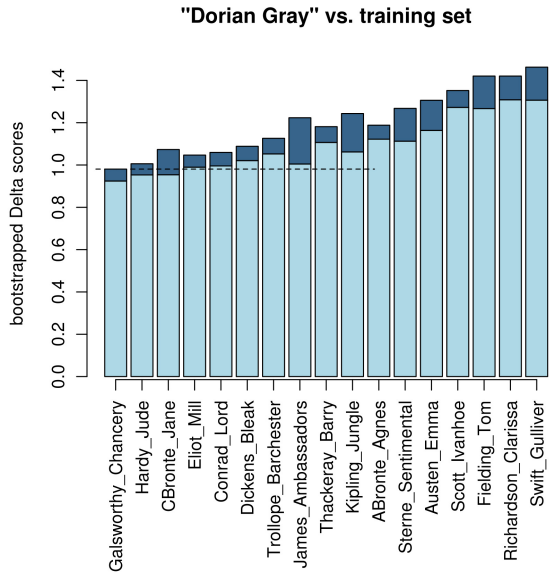


Figure 5.
Who wrote The Portrait of Dorian Gray? Ranking of candidates using confidence intervals.

First benchmark

The first exemplary results of two attribution experiments are shown in Table 1. In both approaches, Jane Austen's *Sense and Sensibility* was assumed to be an 'anonymous' sample to be attributed, and the comparison corpus consisted of 17 texts by known authors. In the first experiment, the behavior of 1,000 single bootstrap trials led to the final ranking with Jane Austen as the only probable candidate (as expected). In the second experiment, Austen's sample was excluded from the comparison corpus, so that the real author could not be guessed. However, the method refused to point out a most likely candidate with a high probability. As one can see, there is uncertainty about the first three candidates.

rank	candidate	probability
1.	Austen	100%
2.	Trollope	0%
3.	A. Brontë	0%
4.	Fielding	0%
5.	Dickens	0%
…	…	…

rank	Candidate	Probability
1.	Trollope	31.8%
2.	A. Brontë	31.2%
3.	Fielding	29.7%
4.	Dickens	7.2%
5.	Eliot	0.1%
…	…	…

Table 1.
Who wrote Sense and Sensibility? A ranking of candidates: (1) where the real author (Jane Austen) is available in the comparison corpus (left); and (2) where the comparison corpus does not contain samples by the real author (right).

The procedure presented above, displays an accuracy comparable to the state-of-the-art methods used in stylometry, but it is far more sensitive to fake candidates. While the existing methods provide two possible answers to the problem of attribution: *X is the author* or *X is not the author*, the procedure proposed introduces a third answer: *unclear results*, an important safety net against false attribution.

References

Argamon, S. (2008). Interpreting Burrows's Delta: geometric and probabilistic foundations. *Literary and Linguistic Computing*, 23(2). 131-47.

Burrows, J. F. (2002). 'Delta': a measure of stylistic difference and a guide to likely authorship. *Literary and Linguistic Computing*, 17(3). 267-87.

Burrows, J. (2003). Questions of authorship: attribution and beyond. *Computers and the Humanities*, 37 5-32.

Craig, H. and A. F. Kinney (2009). *Shakespeare, Computers, and the Mystery of Authorship*. Cambridge & New York: Cambridge University Press.

Good, P. (2006). *Resampling Methods*. Boston: Birkhäuser.

Hoover, D. L. (2004a). Testing Burrows's Delta. *Literary and Linguistic Computing*, 19(4). 453-71.

Hoover, D. L. (2004b). Delta prime? *Literary and Linguistic Computing*, 19(4). 477-95.

Jockers, M. L., and D. M. Witten (2010). A comparative study of machine learning methods for authorship attribution. *Literary and Linguistic Computing*, 25(2). 215-23.

Koppel, M., J. Schler, and S. Argamon (2009). Computational methods in authorship attribution. *Journal of the American Society for Information Science and Technology*, 60(1). 9-26.

Mosteller, F. and D. Wallace (1964). *Inference and Disputed Authorship: The Federalist Papers*. Stanford: CSLI.

Rybicki, J., and M. Eder (2011). Deeper Delta across genres and languages: do we really need the most frequent words? *Literary and Linguistic Computing*, 26(3). 315-21.

Schaalje, G. B., P. Fields, M. Roper, and G. L. Snow (2011). Extended nearest shrunken centroid classification: a new method for open-set authorship attribution of texts of varying sizes. *Literary and Linguistic Computing*, 26(91). 71-88.

Smith, P. W., and W. Aldridge (2011). Improving authorship attribution: optimizing Burrow's Delta method. *Journal of Quantitative Linguistics*, 18(1). 63-88.

Unsupervised Learning of Plot Structure: A Study in Category Romance

Elliott, Jack
jack.elliott@uon.edu.au
University of Newcastle, Australia

There are two broad approaches to machine plot analysis: annotation-based systems (Lendvai et al. 2010) and formal models of plot (Lakoff and Narayanan 2010). Annotation-based systems are inspired by markup languages such as XML, while formal models of plot are offshoots of artificial intelligence research. This paper proposes a new approach, based on gene sequencing, and derives a model of plot directly from a very large corpus of novels without training or a pre-defined model. The technique reduces novels to their narrative components, classifies these components according to type, then recombines these constituent elements to typify the plots of a group of texts.

This technique is applied to an entire genre, the category romance imprint *Harlequin Presents*.

Harlequin Presents publishes roughly eight books every month, and is probably the most commercially successful fiction genre in the world (*Harlequin Company History*). The genre can be characterized by recourse to a limited number of types of plot, although there are distinct sub-categories. Most importantly, the genre is available as an ebook, so each novel in the imprint has a definitive edition that is easily subjected to machine analysis. This study uses 1500 novels — over 15 years of *Harlequin Presents*. This is one of the first studies of popular culture to use machine analysis on an entire genre.

Although the conscription of machines to the task is relatively recent, the study of narrative is not. Traditional narratology can be traced back to Propp's work on folklore in the early twentieth century (Propp 1968). Propp collected a set of functions that described all possible actions in his collection of folk tales. The plot of any single folk tale could be described as a subset of these functions laid end-to-end. Propp's work was rejected (Lévi-Strauss 1976a), integrated (Dundes 1997, 47) and then conflated with that of the structuralists, whose work with myths extended Propp's ideas to cover much more than folklore.

Romance novels have two important parallels with Propp's folk tales and Lévi-Strauss' myths. Firstly, all three genres are, or were, contemporary. Propp's folk tales were a living art form in the early twentieth century (Haney 2009, xiii). Lévi-Strauss recorded many of the oral myths that he later integrated into his theories (Lévi-Strauss 1976b, 35-65). While stretching back 15 years, the most recent Harlequin Presents novels in our sample have been published this month. Secondly, all three genres are curated by others. Propp used a standard edition of folktales and Lévi-Strauss tapped indigenous traditions to define his myths. In our case, *Harlequin Presents* has been categorized by the publisher. Yet, unlike either folk tales or myths, romance novels have never had an oral form — which makes them ideal for machine analysis.

The technique itself is a modified version of Weighted Gene Co-Expression Network Analysis (Zhang and Hovarth 2005). This technique has been developed to allow mining of gene sequencing information, although the application to written language is a natural extension. Like words, genes are typically redundant, in that many genes signal at once to achieve a desired effect, similar to the manner in which words are collocated when expressing an idea. Natural language data is transformed to resemble gene sequencin information by seg- menting novels into bins and counting the words in each segment. A correlation matrix is then computed, giving the strength of relationship between each word to each other. Words are then clustered together into *co-expression networks* based on their frequency of co-occurrence.

Networks of genes that frequently co-occur are known as modules, and this terminology is used here to describe collocated words. The behaviour of a module throughout the genre is then typified, giving a cardinal behaviour for all words in the module. External factors, such as author and date of publication can then be related to the modules, to see how they effect the genre. It is this relationship between modules and external data that reveals the most interesting patterns within the genre. Some modules, such as those relating to the status of the hero, are correlated with the beginning of the novel. Other modules, such as those relating to pregnancy or marriage, are strongly correlated with the final segments of a novel. Other modules are related to authorship, and others can be used to classify the entire genre according to narrative strategy.

Unlike purely stylometric studies, modules are typically closely related to theme and incident - concerns directly under the control of an author. Corre- lation of modules to individual authors is not truly useful for authorship dis-crimination, but reflects preferences that an author can be expected to show as they specialise in particular narrative forms or explore certain themes. Similarly, changes in a genre over time can be seen as a direct reaction to external events rather than changes in an author's internal mental state.

One criticism of traditional narratology is the difficulty it has relating abstract categories back to the mechanics of the writing (Shen 2005, 146). Machine analysis based on annotations or artificial intelligence research both go some way to alleviating this problem. Deriving a model directly from the text eliminates this problem entirely, although it introduces another: modules of words do not always tie closely into our received notions of narrative. In particular, the abstract categories structuralists leveraged to study the similarities between cultures (Lévi-Strauss 1981, 64-66) are not found by this technique. While modules are illustrative of the texts and genres at hand, they do not really generalize beyond them, providing an insight that is deep but not broad.

Broad insights are the specialty of mark-up based and artificially intelligent narrative systems. These other systems have recourse to categories not derived from the texts at hand, and are much more able to draw links between different groups of texts. Mark-up based systems, although they cannot easily scale to working with thousands of novels as we do here, are able to leverage the (often formidable) skills and intelligence of their users. The more formalistic systems, with their pre-programmed categories are also able to generalize from a single genre. This reflects the very different design goals of these approaches: we are concerned here with mere analysis, whereas markup tools are often a

form of scholastic augmentation and artificially intelligent systems typically have plot generation as an ultimate aim (Gervás 2012).

Stylometry has typically focused on high-frequency function words to show the mechanics of language at work. Techniques derived from computational biology allow the extraction of thematic and narrative components, and allows these to be related to authorship, date of publication or other external factors. Other approaches to modeling narrative structure have their strengths, but frequently have broader objectives than the analysis of the texts at hand. Weighted Gene Co-Expression Networks sacrifice these goals but provide a flexible method of unsupervised learning of narrative structure.

References

Dundes, A. (1997). Binary Opposition in Myth: The Propp/Lévi-Strauss Debate in Retrospect. English. In: *Western Folklore* 56.1, 39–50. issn: 0043373X. url: http://www.jstor.org/stable/1500385 .

Gervás, P. (2012). From the Fleece of Fact to Narrative Yarns: a Computational Model of Composition. In: *Workshop on Computational Models of Narrative, 2012 Language Resources and Evaluation Conference (LREC'2012)*. Istanbul, Turkey.

Haney, J. (2009). *An Anthology of Russian Folk Tales*. New York: M. E. Sharpe.

Harlequin Company History. url: http://tinyurl.com/4mwtts (visited on 08/18/2011).

Lakoff, G., and S. Narayanan (2010). Toward a Computational Model of Narrative. In: *Proceedings of the AAAI Fall Symposium 2010*. Istanbul, Turkey.

Lendvai, P. et al. (2010). Propp Revisited: Integration of Linguistic Markup into Structured Content Descriptors of Tales. *Digital Humanities 2010*. Oxford University Press.

Lévi-Strauss, C. (1976a). Structure and Form: Reflection on a Work by Vladimir Propp. *Structural Anthropology* 2. Trans. by Monique Lane. London: Allen Lane.

Lévi-Strauss, C. (1976b). *The Raw and the Cooked*. Trans. by Claude Levi-Strauss. London: Harper & Row.

Lévi-Strauss, C. (1981). Structuralism and Myth. *Kenyon Review* 3.2, 64–88.

Propp, V. (1968). *Morphology of the Folk Tale*. Trans. by The American Folklore Society. Austin: University of Texas Press.

Shen, D.(2005). What Narratology and Stylistics Can Do for Each Other. *A Companion to Narrative Theory*. Phelan, J., and P. J. Rabinowitz, (eds). Blackwell Publishing, 136–150.

Zhang, B., and S. Hovarth (2005). A General Framework for Weighted Gene Co-expression Network Analysis. *Statistical Applications in Genetics and Molecular Biology* 4.

Lost in the Data, Aerial Views of an Archaeological Collection

Esteva, Maria
maria@tacc.utexas.edu
Texas Advanced Computing Center, University of Texas at Austin, United States of America

Trelogan, Jessica A.
j.trelogan@austin.utexas.edu
Institute of Classical Archaeology, University of Texas at Austin, United States

Xu, Weijia
xwj@tacc.utexas.edu
Texas Advanced Computing Center, University of Texas at Austin, United States of America

Solis, Andrew J.
asolis@tacc.utexas.edu
Texas Advanced Computing Center, University of Texas at Austin, United States of America

Lauland, Nicholas E.
lauland@austin.utexas.edu
Liberal Arts Instructional Technologies Service, University of Texas at Austin, United States

Introduction

A guiding principle in digital curation is that continuous management of data renders digital collections sustainable for continuous access and reuse (Higgings 2008). This assumes embedding data organization, documentation, and preservation practices in the research process, ideally from inception, as well as the persistence of resources to support past, current, and future technologies. In reality, these goals remain unattainable for many projects, especially those with

long histories. Large gaps must still be bridged to build solutions to accommodate the realities of contemporary humanities data (Borgman, 2009).

Humanities projects can be long and multi-faceted, leading to convoluted collections that reflect technological and methodological changes over time. Researchers tend to adopt new technologies to solve specific problems and intend to deal with preservation later. Meanwhile, technologies evolve, putting data at risk. Research can be segmented due to funding constraints, changing teams, and new domain directions, making it difficult to enforce standards and achieve consistency. This is especially true in archaeological research, where fieldwork is cyclical and produces huge and complex datasets (Kansa, el al. 2011). After decades of accumulating data, it is easy to become lost in one's own collection, wondering how to make sense of it all (Trelogan et al. 2010).

While initiatives are developing (Open Context 2012; Richards 1997) to help projects prepare data for centralized repositories, there is still a dearth of tools for on-the-fly management of evolving datasets with a long history. Here we present a visual analytics tool that provides an "aerial views" of digital collections and tools to help navigate the curation process.

We present a case study that adds new functionality to a visual analytics application designed for archival analysis with support from the National Archives and Records Administration (Esteva, et al. 2011; Xu, et al. 2011). New developments provide intuitive guidance for users with large collections in need of intervention without interrupting the flow of active research. New functions include tools for locating and sorting primary data, identifying duplicated and corrupted files, and creating timelines of production. It provides a toolkit for investigating formation processes in order to inform future plans and establish priorities for a collections' documentation, preservation, and archiving. While this concept may resonate well with archaeologists familiar with the importance of formation processes and multiple viewpoints for analysis, the tool has wide application for any kind of disparate data with a long, continuing evolution.

Project Background

This study consists of an active collection of over 1,000,000 files from investigations by the Institute of Classical Archaeology (ICA) at the University of Texas at Austin, representing over forty years of research activities. It reflects the technological changes that have affected research since the mid-seventies, as well as the methodological and theoretical changes that have influenced archaeology and associated disciplines. The data are typical of a large archaeological project, from scans of photographs, drawings and field notes, to GIS datasets, 3D visualizations, and databases. Adding further complexity, ICA's multidisciplinary approach has resulted in data amassed by generations of scholars and students, reaching across disciplinary, geographical, and political boundaries, each with its own set of methods, questions, and solutions.

As ICA's focus has recently turned from fieldwork to publication, an increasing sense of crisis has arisen as researchers attempt to retrieve, assimilate, and share digital resources for study and dissemination. Past efforts to organize and document this collection have been piecemeal and have lacked consistent conventions for file naming, metadata, or organization. Data previously distributed throughout hard drives and detached storage devices have been consolidated in a networked server, but remain in serious disarray.

Help was sought from the Texas Advanced Computing Center (TACC), which provides high performance computing services, including data management support. Starting in 2009, the teams began a collaboration to assess and organize a sample dataset from one of ICA's excavations (Esteva, et al. 2010). Work has since expanded to assess the entire collection, with goals to develop data management strategies that can be adopted without interrupting ongoing research, to document and archive the collection in TACC's storage infrastructure, and to provide web access for collaboration and dissemination. Here we describe new functionality and promising results with the initial triage of the collection using the visual analytics application discussed above.

Methods

The application uses file system and file format metadata extracted from files and directories to create a visualization of the collection, akin to aerial photographs that provide encompassing views of a landscape from above. File paths and sizes represent the collection's organizational structure as a treemap (Shneiderman 1992), and file format metadata is further classified using PRONOM's file categories (PRONOM 2012) to narrow the information to a comprehensible amount. File classes are rendered as colored distributions within the collection's structure.

In the background the extracted metadata is computationally analyzed and aggregated at different directory levels, and the results are written to a comma-separated table processed for display by the visualization. Users can interact with the visual representation of the collection by navigating directory levels, searching directory names, browsing, and selecting directories for observation. They can also select and group metadata using filtering functions. Tag clouds showing directory and file names, color and file class maps, and charts representing numbers

of files per class allow users to verify and clarify the collection views.

File Class Analysis

In this project, for purposes of understanding the contents of the collection in relation to research stages, file classes are further categorized as: 1) primary, 2) process, and 3) publication data. Raster images, for example, are more likely primary data (e.g. site photographs), whereas vector images are more likely illustrations for publication. These categories were mapped to contrasting color maps, allowing for quick visual identification of directories according to classes and categories of the files they contain (Figure 1). Figure 1 is a view of the entire collection represented as a treemap showing file classes.

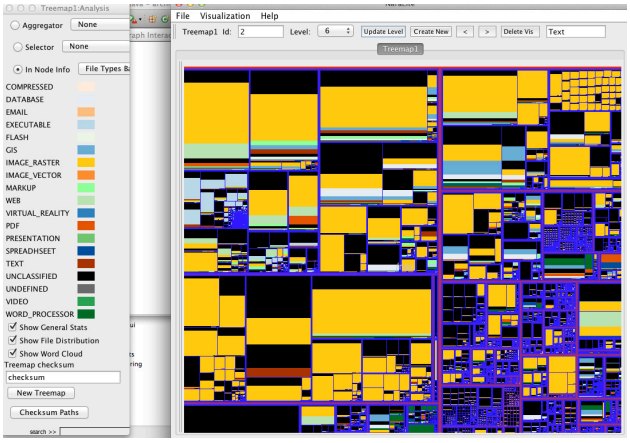


Figure 1.
5tb of archaeology data are represented as a treemap. The predominance of raster images across the collection, as well as primary, process, and publication data can be identified.

Directories can be further explored via search and browsing functions to refine the assessment. One useful discovery was a large collection of artifacts left over from obsolete operating systems and software. Using the directory labels search function we located a large collection of raw digital photographs that had been thought lost. Consequently several areas of the collection were identified as priority for archiving and others for disposal.

Checksums and Date Analyses

A checksum analysis function was developed to aid identifying corrupted and duplicated files. Identical checksums are rendered in the directories where they are located. Through this function we found that identical checksums can exist in large quantities (over ~5000 in this archive). Those are likely to belong to error messages, empty strings, cache artifacts and similar images made by databases for speeding retrieval, are generated automatically, and can span many directories. This led the team to find multiple copies of old web-based databases to mark for deletion. Instead, checksums for duplicate files likely created by copying files or directories, are generally distributed across only two or three directories. Several sets of duplicated files were located and marked for deletion, freeing up space and allowing a clearer picture of the collection.

In analyzing repeated checksums, associating them with file class information aids in deciding if they should be kept or deleted. In addition, learning the complete file path allows identifying if these are temporary files, svn-related, duplicates, backups, and their provenance. The location of these files and their distribution are indicative of the archive's formation process and can lead to clues about associated data (Esteva 2008).

We are working on incorporating dates into the analysis so that file classes and checksums can be viewed in a timeline. Last modified dates will be aggregated by year and shown in relation to file classes. Users will be able to select viewing features and amounts for a given year[1] , and many treemaps can be opened at a time to allow comparison of technologies in a timeline.

Much like in archaeological excavations, the knowledge brought by various experts regarding the collection's history, work processes, workflows, and technological change, aids in understanding the collection's context and planning for its preservation. With these tools, we are able to "dig" through the archive much more effectively and gain a clearer picture of its content and significance.

Conclusions

This application provides a framework to study collections with interactive tools for discovery, condition assessment, and a roadmap to make inferences about the processes by which they were formed. It has wide-ranging implications for collections that must be explored from multiple viewpoints — as with archaeological landscapes — in order to understand their condition and plan their future trajectories.

A work in progress, the tool requires further development from multiple ends. Presently, not all file formats can be identified and the file classes may be too broad for the study of specific datasets. Ideally, the tool should be implemented to update dynamically for purposes of continuous monitoring of active collections. We will

illustrate the presentation with images, report findings, and will discuss functionalities and the challenges ahead.

References

Borgman, C. L. (2009). The Digital Future is Now: A Call to Action for the Humanities. *Digital Humanities Quarterly,* 3(4) http://digitalhumanities.org/dhq/vol/3/4/000077/000077.html%20/000077.html (accessed 1 November 2012).

Esteva, M. (2008). Formation Process and Preservation of a Natural Electronic Archive. In *Proceedings of the American Society for Information Science and Technology 2008 Conference* held 24-29 October 2008 in Ohio. 45(1): 1-9. doi: 10.1002/meet.2008.1450450270

Esteva, M., J. Trelogan, A. Rabinowitz, D. Walling, and S. Pipkin (2010). *From the Site to Long-Term Preservation: A Reflexive System to Manage and Archive Digital Archaeological Data. Proceedings of the IS&T's Archiving 2010 Conference* June 1-4, 2010, Den Haag, The Netherlands. http://test.imaging.org/ScriptContent/store/epub.cfm?abstrid=43763 (accessed 1 November 2012).

Esteva, M, W. Xu, S. D. Jain, J. Lee, and W. K. Martin (2011). Assessing the Preservation Condition of Large and Heterogeneous Electronic Records Collections with Visualization. *International Journal of Digital Curation,* **6:1** UKLON, University of Bath. Digital Curation Center. http://www.ijdc.net/index.php/ijdc/article/view/162 (accessed 1 November 2012).

Higgins, S. (2008).The DCC Curation Lifecycle Model. *International Journal of Digital Curation.* **3:1** 134-140. doi:10.2218/ijdc.v3i1.48.

Kansa, E. C., S. W. Kansa, and E. Watrall (2011). Archaeology 2.0: New Approaches to Communication and Collaboration Location: Cotsen Institute of Archaeology. http://escholarship.org/uc/item/1r6137tb (accessed 1 November 2012).

Open Context. Alexandria Archive Institute. http://www.opencontext.org. (accessed 1 November 2012).

Richards, J. D. (1997). Preservation and Re-use of Digital Data: the Role of the Archaeology Data Service. *Antiquity* **71** 1057–1059.

Shneiderman, B. (1992). Tree Visualization with Tree-maps: 2-d Space-filling Approach. *ACM Trans. Graph.* 1992: **11**, 92-9.

The National Archives. PRONOM,The Technical Registry. http://www.nationalarchives.gov.uk/aboutapps/pronom/ (accessed 1 November 2011).

Trelogan, J., A. Rabinowitz, M. Esteva, and S. Pipkin (2010). What Do we Do with the Mess? Managing and Preserving Process History in Evolving Digital Archaeological Archives. In Contreras, F., and F. J. Melero (eds.), *Proceedings of the 38th Conference on Computer Applications and Quantitative Methods in Archaeology.* held April 6-9 in Granada, Spain.

Xu, W., M. Esteva, S. J. Dott, and V. Jain (2011). Analysis of Large Digital Collections with Interactive Visualization. VisWeek. Conference. In *Proceedings of the IEEE Conference on Visual Analytics Science and Technology.* held 23-28 October in Providence, Rhode Island. 241-250, doi: 10.1109/VAST.2011.6102462.

Notes

1. The assumption is that last modified dates are valid for analysis, which is the case in this study collection.

Mapping Homer's Catalogue of Ships

Evans, Courtney
cme2c@virginia.edu
University of Virginia, United States of America

Jasnow, Ben
bbj9t@virginia.edu
University of Virginia, United States of America

Our project, *Mapping the Catalogue of Ships*, created by Classicist Jenny Strauss Clay, illustrates the route of Homer's poetic journey around Greece and provides a digital forum in which to test and develop theories about the poet's knowledge and use of geography. This paper will detail the ways in which the construction of these digital exhibits has both deepened and changed our initial theories about Homer's familiarity with ancient routes.

The Research Question

In Book 2 of Homer's *Iliad*, the poet embarks upon a seemingly impossible feat: to enumerate the commanders of the 29 contingents of the Greek expedition, along with the number of ships and troops belonging to each, and the almost 190 towns from which they came. It must have seemed a superhuman performance to the audience of the poet's day. Homer was an oral poet, composing his verses aloud and extemporaneously, without the use of writing. Nor did Homer have the benefit of looking at a map, since the first maps of the Greek world came about only after the composition of the *Iliad*. Yet the poet

presents his 250-verse masterpiece as an organized tour of the Greek world, subdividing the commanders and their contingents according to geography. It was such a convincing performance that the late antique geographer Strabo would name Homer as the "father of geography." Yet the degree to which Homer was familiar with the details of Ancient Greek landscape remains unclear.[1]

The places named in the Catalogue may be divided into two types: large kingdoms and the cities within those kingdoms. A well-known geographical principle clearly underlies the narrative order in which Homer relates the 29 large kingdoms that make up the Greek fleet. Beginning from Boeotia, in Central Greece, the poet narrates three circuits of these kingdoms, moving from one geographical region to the next in a continuous fashion (Clay 2011; Minchin 2001). This well organized plan or mental roadmap serves the oral poet as a "spatial mnemonic" (Clay 2011), allowing Homer to traverse the nearly 190 places he mentions without getting lost in the details (Clay 2011; Minchin 2001). One function of *Mapping the Catalogue of Ships* is to illustrate this large-scale navigation of Greece, making clear for students and scholars the fundamental order that underlies Homer's tour-de-force of memory.

Although the principle according to which Homer moves from kingdom to kingdom is well understood, the poet's use of geography remains, in its other aspects, mysterious. Several scholars have suggested that Homer may have used ancient travel itineraries to organize the Catalogue (Clay 2011; Kirk 1985). However, we lack a detailed analysis of the particular routes and landscape features that Homer would have employed. The main purpose of *Mapping the Catalogue of Ships* is to fill this scholarly gap. Does the geographical ordering of the 29 kingdoms represent an actual itinerary? Moreover, does the narration of the cities within those kingdoms constitute an itinerary? Does the narrative order of the place-names in the catalogue reflect ancient roads and landscape features? These are the questions *Mapping the Catalogue of Ships* seeks to answer.

Neatline: Mapping the Catalogue of Ships

Our project uses an exciting new tool under development by the Scholars' Lab at the University of Virginia. That tool is *Neatline* (neatline.org) a "geotemporal exhibit-builder," which combines the resources of an interactive archive and narrative timeline with a custom mapping capability, making it an ideal means to engage with the poetic, geographical and cultural challenges of Homer's Catalogue.

A full understanding of the Catalogue of Ships requires simultaneous exposure to many different types of information. The text of the *Iliad* must be integrated into a map display capable of presenting detailed information and scholarship about ancient routes and landscape. Moreover, since our investigation depends upon understanding the Catalogue from a traveler's point of view, the map display must highlight the link between Homer's narrative order and the order of the points upon the map. Our *Neatline*-supported exhibition of the Catalogue of Ships will do all these things. With help from *Pleiades* (http://pleiades.stoa.org/), an online, open source gazetteer of the ancient world, we have already compiled a database containing the locations of the sites featured in the Catalogue. Drawing on this database, we have begun to create maps to illustrate the various stages of Homer's narration (each with selectable layers). When the project is finished, the user will be able to zoom out for a macroscopic view of Homer's entire journey around the Greek world, or zoom in for a more detailed look at the routes and geography belonging to each one of the 29 contingents mentioned by the poet. Each map will feature a menu-bar, listing every place that Homer names. By selecting a place-name in the menu-bar or on the map itself, visitors may browse through textual analysis, photographs as well as scholarship about archeology and ancient routes. Where archaeological evidence does not reveal the track of ancient roads, we will propose probable routes using least-cost-path GIS analysis. But the Catalogue is more than a list of places; it is a poetic narrative. Every map, therefore, will also feature an interactive text of the *Iliad*, which acts as a narrative timeline. As the user scrolls through the text of the *Iliad* (Greek and English), the map will follow along, moving to that portion of the display appropriate to the narrative. Likewise, when the user selects a particular map-exhibit, the text will jump to the appropriate portion of the Catalogue.

Mapping the Catalogue of Ships takes advantage of an increasing movement towards digital modeling and mapping in the Classics. *Digital Atlas of the Roman Empire* (francia.ahlfeldt.se/imperium.php) and the *Digital Atlas of Roman and Medieval Civilizations* (http://darmc.harvard.edu/icb/icb.do) are both useful resources and contain a certain amount of information about ancient routes. A closer parallel to *Mapping the Catalogue of Ships*, however, is *ORBIS* (http://orbis.stanford.edu/#), which uses maps not merely as a way to get a bird's-eye-view, but to understand the cultural context of ancient geography. *ORBIS* analyzes ancient travel networks, mapping least-cost-paths by land and sea while accounting for many different variables (e.g. wind patterns, travel by donkey-cart vs. by foot, civilian vs. military travel). Our project also uses least-cost-path analysis of travel networks, but seeks to understand how such networks may have shaped a literary work. In its attention to the use of space by oral poets,

Mapping the Catalogue of Ships has a clear predecessor in Professor Clay's earlier project, *Homer's Trojan Theater* (http://www.homerstrojantheater.org/), which demonstrated that Homer's disposition of the *Illiad's* battlefield serves as a mnemonic device.

Unexpected Results

Although *Mapping the Catalogue of Ships* is not yet complete, the process of plotting routes and building exhibits has already yielded interesting results. For example, we have found that the theory of Homer's catalogue as a series of itineraries appears to be correct in some respects. The poet frequently finishes enumerating the cities of one region at the geographical point closest to the area to which he is about to proceed. In some cases, moreover, Homer appears to possess detailed geographic knowledge about the disposition of cities along routes or landscape features, as in the narration of the Mycenaean contingent. Here the narrative order reflects local routes, supporting the theory, proposed by Clay and others, that the Catalogue reflects an actual itinerary. In instances where Homer seems to possess detailed geographic knowledge, his syntax mirrors that familiarity: regional subdivisions also constitute syntactic subdivisions, with a single verb governing each of the places grouped around a particular geographic area or landscape feature.

In the narration of certain regions, however, Homer seems to lack intimate knowledge, or else is unconcerned to construct his list in an order that could reflect a plausible travel-route. Such regions pose a difficulty to Clay's theory of the Catalogue as an itinerary. The Boeotian contingent is a good example of an instance in which the poet is unconcerned with listing locations in an order that could be followed by a traveler. Instead, Homer here narrates in a rough circuit around the city of Thebes, the major power of the area. This new way of understanding the narration of the Boeotian contingent finds a striking parallel in Homer's Catalogue of Trojans, which moves in a series of spokes around the city of Troy (Clay 2011).

The process of constructing this tool has yielded unexpected results and deepened our understanding of Homer's use itineraries. This paper will report on these and other unexpected findings as we continue to expand the tool.

Credits

Jenny Strauss Clay, William R. Kenan, Jr. Professor of Classics, University of Virginia
Courtney Evans, Graduate Student, Department of Classics, University of Virginia
Ben Jasnow, Graduate Student, Department of Classics, University of Virginia

Scholars' Lab Collaborators

Bethany Nowviskie, Director, Digital Research & Scholarship
Wayne Graham, Head, Research & Development
Jeremy Boggs, Design Architect
Chris Gist, GIS Specialist
Kelly Johnson, GIS Specialist

References

Clay, J. S. (2011). *Homer's Trojan Theater: Space, Vision, and Memory in the Iliad.* Cambridge: Cambridge University Press.

Kirk, G. S. (1985). *The Iliad: A Commentary,* vol. I. Cambridge: Cambridge University Press.

Lattimore, R., (1951). *The Iliad of Homer*. trans. by Lattimore. Chicago: University of Chicago Press.

Minchin, E. (2001). *Homer and the Resources of Memory.* Oxford: Oxford University Press.

Notes

1. As the composition of an oral poet, the Catalogue of Ships is the product not only of Homer, but also the entire oral tradition, consisting of highly formulaic language and content, altered on the occasion of every bard's performance, stretching back hundreds of years prior to the written form of the poem as we have it. In examining the use of space and geography in Homer's Catalogue, we are also examining the larger Homeric tradition.

Responding to the frame: classification, material boundaries, and expressiveness in personal digital bibliography

Feinberg, Melanie
feinberg@ischool.utexas.edu

The University of Texas at Austin, United States of America

A *systematic bibliography* is a thematic collection of references to intellectual works, such as the subject guides produced by academic librarians. Free of the constraints of physical libraries, these metadata representations are able to describe and relate any resource in the bibliographic universe at the will of the bibliographer, to fulfill an endless variety of goals and express any number of associated opinions. A consumer health bibliography might provocatively endorse certain elements of alternative medicine; a guide to anthropology might focus on "scientific" methods as opposed to ethnographic ones, as espoused by the department at a particular institution.

Accordingly, the ultimate collection, comments Roger Chartier, is not one of actual books but of citations, a "library without walls," as in Conrad Gesner's monumental sixteenth-century compendium (Chartier, 1994). Instead of shelves, Gesner's intricate classification scheme endows the bibliography's contents with form and structure. Although expressed in words, a classification of this sort is meant to establish relationships between abstract concepts, for which the identifying terms are merely replaceable labels. A category labeled "Geometry" indicates the *idea* of geometry, not the word. Current standards, such as the 2005 NISO standard for controlled vocabularies and the 2010 IFLA recommendations for the representing subjects in the library catalog (FRSAD) continue this mode. (Jonathan Furner [2012] contributes a useful critique of the FRSAD approach.) Even library classifications are designed as abstractions, independent of the material world, to be potentially expressed in any shelving configuration; in these classifications, as well, the concept that designates a class is conceived as a purely mental construct. Its proper identification is a notation or code (e.g., GN 790 in the Library of Congress Classification), for which any label is just a form of convenient documentation.

The interpretive nature of these concepts, of their system of relations in a classification scheme, and of their application to any set of physically or virtually collected works, has been widely accepted within information studies (as in, for example, work by Birger Hjorland, Jens-Erik Mai, and Hope Olson). A subject class like GN 790 (Anthropology — Prehistoric archaeology — Megaliths) embodies a remarkable complex of judgments: that the significance of megaliths has to do with anthropology, that a subject is both an identifiable and important characteristic of documents, that the set of documents assigned to GN 790 are equally about "megaliths," that "aboutness" is a decision easily and concretely made, as only some examples of the complex of appraisals that GN 790 expresses.

The recognition of classification as a mode of interpretation, and an accompanying sense that a citation collection or bibliography is itself a form of creative expression, is similar to the recognition of textual scholars that the application of markup languages such as TEI to digital texts is equally an interpretive act, and that critical editions expressed as digital collections are interpretive efforts. In textual studies, such realizations have led scholars such as Jerome McGann (2001), Johanna Drucker (2006), and Bonnie Mak (2011) to examine both the ways in which the material qualities of printed documents contribute to meaning and the means in which these material effects are transformed through digitization. Mak, for example, traces the semantic contributions of the page in manuscript, print, and digitized versions of a fifteenth-century Latin text. While the page signifies differently across these versions, the page as a material and conceptual construct always "matters."

In contrast, due to the longstanding emphasis on the nature of classification as abstraction, the contribution of any material component to the meaning of a systematic bibliography would initially seem quite strange. Classifications are most often studied as self-contained systems of their own, and not as applied to works in practice. Even when classifications have been studied as infrastructural components of document collections, the relation between class and text is seen as a dialectic of ideas, and not as a means of constituting a thing with a specific material presence. The translation of paper-based systems to digital environments has only confirmed this rejection of presentation as having anything at all to do with the meaning of either classification schemes or the systematic bibliographies structured by them.

This paper describes how a user study focused on the authoring of personal digital collections, or personal systematic bibliographies, demonstrated a link between the material instantiation of a bibliography and the expressiveness of its structuring. Findings from this study suggest that meaning making for both writers and readers of personal digital bibliographies inheres to some degree within the abilities of the bibliographic environment to enact framing devices that clarify the system of relations between items as the primary focus of the text. Just as Gesner's famous bibliography achieved notoriety not merely for its comprehensive selection of works but for its instantiation of relations between its contents, so does the expressiveness of personal digital collections result from the interplay between structure and contents, and not merely from the contents alone. In this paper, I contend that the ability to articulate and perceive this set of relations requires a frame with material presence, and not just mental presence.

The study described here is part of a larger project to investigate personal digital collections, or personal systematic bibliographies, as a form of creative expression. On the Web, social media services that enable users to aggregate content through the creation of personal digital

collections have proliferated. Through communities such as Pinterest and GoodReads, users select and share resources linked (or cited) from elsewhere. Personal collections are enabled for retail sites, such as Amazon, and content providers, such as YouTube and Spotify. Cultural heritage institutions, including libraries and museums, have also encouraged users to create personal collections, through services like MyMet from the Metropolitan Museum of Art in New York. Marty and Kazmer (2011) contend that such personal bibliographies facilitate the co-construction of knowledge between institutions and their user communities. As opposed to Gesner's opus, however, most Pinterest boards and YouTube playlists are trivial and bland. The project's initial investigation identified, out of the masses, a few salient examples that did constitute compelling expression and defined characteristics that contributed to these examples' expressiveness: an original purpose for creating the bibliography, a distinct authorial voice in the presentation of the bibliography, and a sense of emotional intimacy underpinning the bibliography's contents (Feinberg 2010, 2011).

The second stage of the project began to consider the design of environments to author such collections: could we encourage writers of personal bibliographies to generate more expressive examples? (One can see this as similar to an investigation focused around describing a markup language to more effectively characterize the poetic nature of texts as opposed to their informative nature, a complaint against TEI articulated by Jerome McGann.) A first experiment examined whether exposure to bibliographies that embodied all three of the identified expressive characteristics would affect the process or product of collection design (Feinberg, et al, 2012). Twelve participants created bibliographies by selecting items from specially constructed digital video libraries of source content. After creating one bibliography, participants interacted with expressive examples that were also created from the that video library. Participants then created a second bibliography using a different library of video source material. This simple intervention did not work; participants' bibliographies and design processes did not change after experiencing the examples. Participants did, however fluently read and describe the distinctive characteristics of the expressive bibliographies. Participants especially noticed the use of descriptive infrastructure (titles, annotations, and so on) in the examples to convey meaning, yet participants made little use of such descriptive elements themselves.

This paper focuses on a second experiment to encourage descriptive infrastructure as a means of facilitating expressiveness in participant bibliographies. We examined whether a structured task that asked participants to separately consider the mechanisms of resource selection, description (titles, labels, annotations), and arrangement (systems of ordering, or relations between items) would enhance expressiveness. To focus on task, and not interface, we switched from a digital environment to a simple physical one, using small libraries of print books as source materials and bulletin boards as document substrates (with preprinted paper slips for citations, along with index cards and Post-its for descriptive elements). To our surprise, while task structure did not influence designs, all the bibliographies were more expressive, compared to the examples, than the previous experiment, and 19 of 24 participants implemented classificatory structure in their work. Many of these structures were reinforced via complex visual arrangements. This paper describes how what initially appeared to be small, inconsequential changes in the material conditions of our experiment resulted in markedly different authoring processes, as well as significantly more complex products. I discuss what the framing devices illuminated via this experiment signify for the development of digital environments to support the authoring of personal expressive bibliographies. As Johanna Drucker (2005) delineates the areas on a page where graphic elements contribute to its meaning, I will describe how material elements of personal digital bibliographies may contribute to (or detract from) their textuality.

References

American National Standards Institute/National Institute of Standards Organization (ANSI/NISO). (2005). *Guidelines for the Construction, Format, and Management of Monolingual Controlled Vocabularies*. (ANSI/NISO Z39.19 — 2005). http://www.niso.org/apps/group_public/project/details.php?project_id=46 (accessed 3 March 2013).

Chartier, R. (1994) *The Order of Books*. trans by Cochrane, L. Stanford, CA: Stanford University Press.

Drucker, J. (2006). Graphical Readings and the Visual Aesthetics of Textuality, *Text*. 16. 267–276.

Feinberg, M. (2011a). Expressive Bibliography: Personal Collections in Public Space. *Knowledge Organization*. 38 (2). 123–134.

Feinberg, M. (2011b). Personal Expressive Bibliography in the Public Space of Cultural Heritage Institutions. *Library Trends*. 59 (4). 588–606.

Feinberg, M., G. Geisler, E. Whitworth, and E. Clark (2012). Understanding Personal Digital Collections: an Interdisciplinary Exploration. *Proceedings of the 2012 ACM Conference on Designing Interactive Systems (DIS)*. 200–209.

Furner, J. (2012). FRSAD and the Ontology of Subjects of Works. *Cataloging and Classification Quarterly*. 50 (5-7). 494–516.

Hjorland, B. (1992). The Concept of "Subject" in Information Science. *Journal of Documentation*. 48 (2). 172–200.

International Federation of Library Associations (IFLA). *Functional Requirements for Subject Authority Data (FRSAD)*. http://www.ifla.org/files/assets/classification-and-indexing/functional-requirements-for-subject-authority-data/frsad-final-report.pdf (accessed 3 March 2013).

Mai, J.-E. (2011). The Modernity of Classification. *Journal of Documentation*. 67 (4). 710–730.

Mak, B. (2011). *How the page matters*. Toronto: University of Toronto Press.

Marty, P. and M. Kazmer (2011). Introduction to Understanding Users. *Library Trends* 59 (4). 563–567.

McGann, J. (2001) *Radiant Textuality*. New York: Palgrave.

Olson, H. (2001). The Power to Name: Representation in Library Catalogs. *Signs* 26 (3). 639–668.

Six Degrees of Francis Bacon

Finegold, Michael Andrew
mfinegol@andrew.cmu.edu
Carnegie Mellon University, United States of America

Warren, Christopher
cnwarren@cmu.edu
Carnegie Mellon University, United States of America

Shalizi, Cosma
cshalizi@cmu.edu
Carnegie Mellon University, United States of America

Shore, Daniel
ds663@georgetown.edu
Georgetown University, United States of America

Wang, Lawrence
lawrencw@andrew.cmu.edu
Carnegie Mellon University, United States of America

Six Degrees of Francis Bacon (SDFB) is a digital reconstruction of the early modern social network (EMSN) that scholars and students from all over the world will be able to collaboratively expand, revise, curate, and critique. The primary motivation for the creation of SDFB is to make possible a new way to reconstruct, represent, and study the complex relations between authors, texts, publishers, and readers in the early modern period.

Historians and literary critics have long studied the ways that early modern writers and thinkers associated with each other and participated in various kinds of formal and informal groups. Yet their findings, published in countless books and articles, are scattered, unsynthesized, and unstructured. There is currently no way to get a unified view of the early modern social network. A scholar must start largely from scratch if she seeks to do any of the following:

1 Understand the importance or type of a particular bilateral relationship
2 Identify potentially important relationships that have yet to be explored
3 Understand the extent of communities of interaction
4 Visualize the consensus opinion regarding networks, whether small or large

To grasp the scale of the problem, consider that, at a conservative estimate, a network representing key early modern figures, their friends, families, critics, biographers, adversaries, influences, etc., could easily run to over 10,000 nodes. Each actor node could potentially be connected to any of the other nodes, leading to 5 million or more potential edges to explore.

To build a network of such a size manually is scarcely feasible. A computational approach, unifying the dispersed knowledge in the scholarly literature, is essential. Indeed, the digital tool we are building is superior to prose alone in representing the complexities of the early modern social network. Unlike published prose, it is extensible, collaborative, and interoperable: extensible in that affiliations can always be added, modified, developed, or, removed; collaborative in that it synthesizes the work of many scholars; interoperable in that new work on the network is put into immediate relation to previously mapped relationships.

Currently funded by a Faculty Research Grant from Google, SDFB is an innovative approach to comprehensively learning the early modern social network. It combines computational and statistical methods, which can explore the relationships of thousands of historical figures using hundreds of thousands of source documents, with the local expertise of a growing number of humanities scholars.

SDFB does not aim to replace the work of scholars studying individual texts, persons, or historical conjunctures. It does not even aim to provide a single totalizing synopsis of the scholarly literature. Rather, its twin goals are to make it easier for scholars to grasp what the scholarly community, as a whole, already knows about

particular connections or social formations, and to enable the exploration of larger-scale patterns, alongside detailed studies. SDFB generates the social network in the following steps:

1 Identify and continuously expand the collection of sources used as input
2 Process unstructured data sources (largely text) into structured data amenable for statistical analysis
3 Apply statistical methods to infer relationships
4 Validate a sample of proposed relationships using local expertise of humanities scholars
5 Organize and visualize the social network

We illustrate these steps by describing the start-up phase of SDFB, where we used the entries in the Oxford Dictionary of National Biography, and a statistical model in which network connections are reflected in co-mentions of actors' names, to produce an estimate of the existence and types of relationships between thousands of figures in early modern Britain.

We used the ODNB as the source of documents both because of the present authors' primary interest in the target geography and time period (Britain, c. 1550–1700), and because the dense-in-data documents are all machine-readable text in largely uniform format. Existing named entity extraction software (Alias-i, 2008; Finkel et al., 2005) produced an initial list of people mentioned in each document, which was refined with semi-automatic disambiguation and de-duplication of names. From this we created a matrix, where the rows represent documents (or document sections), the columns represent historical figures, and the entries of the matrix are the number of times each document mentioned each figure.

We developed a statistical model for the data in this matrix form, assuming that direct connections between historical figures will be reflected by their being mentioned together, but that such direct connections will "screen off" indirect connections mediated by third (or fourth, etc.) parties, rendering the latter irrelevant. That is, if Irene and Joey are connected they will tend to be mentioned together in source documents, so how often Irene is mentioned can be predicted (in part) from how often Joey is mentioned. Likewise if Joey and Karl are connected, mentions of Karl predict mentions of Joey. But if there is no direct tie between Irene and Karl, then mentions of Karl convey no information about mentions of Irene not already accounted for by mentions of Joey. Under this assumption, inferring the existence of network connections is then the same problem as inferring the conditional independence structure in our statistical model, a well-studied problem in statistics and machine learning (Spirtes, Glymour and Scheines, 1993). We solve our instance of the problem through a sequence of penalized Poisson regressions (the "Poisson graphical lasso", Allen and Liu 2012). This allows us to process thousands of historical figures in tens of thousands of documents in a matter of hours, and to obtain confidence intervals on each network connection through subsampling of documents.

Having inferred the existence of network ties through conditional dependence, we infer the types of relationship by examining the distribution of key words found in small sections of documents where a pair of connected people are both mentioned. We compare this distribution to those of pairs where the relationship is known and validated by experts, building a relationship-type classifier with standard supervised learning techniques.

We then choose a sample of people in our network and create a master list of connections inferred from our model and from competing methods. These master lists are provided in random order to a select group of scholars who then order them by importance of the relationships. We use these ordered lists to tune our model and demonstrate its effectiveness relative to competing methods.

The next phases of the project will involve extending all components of the approach. Millions of texts have been digitized and are theoretically available for use in this project, but even identifying the relevant sources is a non-trivial task. The next phase will not just increase the number of source documents, but also broaden the kinds of documents used. Some historical accounts may provide evidence of relationships not found in DNB entries, and printed scholarly compilations from primary sources (e.g., membership rolls, parish records) may be invaluable for some types of relationships. This will require new methods of processing different types of unstructured data as well as new statistical models. In the first phase all the documents were of the same type (online biographies) and a unified statistical model made sense – this may not be true as we include a variety of texts. We will also consider dynamic network models that attempt to infer the timing of relationships.

Validation will be expanded to include an ever broader collection of experts and will focus on individual relationships that are not well explained by the current model. The problem is algorithmic (an automated feedback loop to refine the model) as well as managerial (allowing corrections from multiple experts while assuring quality).

Finally, the next phase will involve roll-out of a wiki front-end that allows easy exploration of actors, their connections, and network visualizations.

References

Alias-i (2008). LingPipe 4.1.0. http://alias-i.com/lingpipe .

Allen, G. I., and Z. Liu (2012). A Log-Linear Graphical Model for Inferring Genetic Networks from High-Throughput Sequencing Data, htttp://arxiv.org/abs/1204.3941.

Davidson, C. N. (2012). Humanities 2.0: Promise, Perils, Predictions. In Gold, M. K. (ed.) *Debates in the digital humanities*. Minneapolis: University of Minnesota Press.

Evans, J. A., and J. G. Foster (2011). *Metaknowledge*. Science 331.6018: 721–725. Online. Internet. 3 Nov. 2012.

Finkel, J. R., T. Grenager, and C. Manning (2005). Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling in *Proceedings of the 43nd Annual Meeting of the Association for Computational Linguistics* (ACL 2005), 363-370 http://nlp.stanford.edu/~manning/papers/gibbscrf3.pdf .

Grafton, A. (2009). *Worlds Made by Words: Scholarship and Community in the Modern West*. Harvard University Press.

Liu, A. (2011). Friending the Past: The Sense of History and Social Computing. *New Literary History* 42(1) : 1–30. Accessed 3 Nov. 2012.

Michel, J.-B., et al. (2011). *Quantitative Analysis of Culture Using Millions of Digitized Books*. Science 331(6014): 176–182. Accessed 4 Oct. 2012.

Moretti, F. (2007). Graphs, Maps, Trees: Abstract Models for Literary History. vols. Verso.

Spirtes, P, C. Glymour, and R. Scheines (1993). Causation, Prediction, and Search. Springer Lecture Notes in Statistics, vol. 81. New York: Springer-Verlag.

The Science Fiction of Science: Collaborative Lexicons and Project Hieroglyph

Finn, Edward
edfinn@asu.edu
Arizona State University, United States of America

Overview

The Center for Science and the Imagination at Arizona State University is the institutional home for the Hieroglyph project, an innovative collaboration that pairs leading science fiction writers with scientists and engineers to craft techno-optimistic narratives set in the near future. The project asks these writers to create stories that radically extend real technologies and ideas that are at least conceptually possible. Hieroglyph draws its name from the notion that we have been inspired by certain iconic fictional ideas—Robert Heinlein's rocket ships and Isaac Asimov's robots, for example—that have come to shape our visions of the future.

In addition to organizing this project and serving as a co-editor of its culminating anthology, I am also researching Hieroglyph as a site of collaborative communication. By tracking the spread and contextualization of keywords during the project, I will explore the role of shared language in framing collaborations. Using empirical data as well as interviews and close readings, my research will consider how successful Hieroglyph is in living up to its name, asking how narrative keywords can serve as catalysts for stories that extend scientific discourse into the realm of science fiction.

Background

Hieroglyph is a collaborative project between the Center for Science and the Imagination at Arizona State University and a collective of science fiction writers led by Neal Stephenson. The project's end goal will be an anthology of near-future, techno-optimistic fiction that engages with technologies already visible on the horizon. The project is unusual in that it asks these writers to collaborate directly with scientists and engineers on their ideas, and to do so primarily through the medium of a website where they can converse and share work in progress.

The Hieroglyph site serves as both a community and a serial publication platform, regularly releasing work in progress, interviews and other contributions as a kind of "Hieroglyph Magazine" so that subscribers can experience the project online and through mobile apps. In this way the site incorporates multiple discourse communities within the same platform, creating spaces for interaction, collaboration and play between science fiction writers, their fans, research scientists and engineers.

Hieroglyph is a bold experiment because it brings together several diffuse communities into a microcosm of new cultural practices. By asking science fiction authors to work directly with scientists and engineers, the site formalizes and compresses a long-running informal relationship that has mutually informed readers and writers of science fiction for many decades. At the same time, the project creates a public performance aspect for these collaborations through its serial publication, inviting a general reading public to track work in progress and engage with both the science and the personalities behind the stories under development.

Proposal

I propose to evaluate the success of Hieroglyph and its individual collaborations by bringing a literary frame to the quantitative analysis of interdisciplinary collaboration. This approach draws together Bruno Latour's Actor-Network Theory, Pierre Bourdieu and John Guillory's concepts of social and cultural capital and my own evolving mixed-methods computational and close reading approach to these collaborations (Latour 2005; Bourdieu 1993; Guillory 1993; Finn 2012). My presentation at Digital Humanities 2013 will focus on several key research questions:

How do authors and their technical collaborators converge on shared understandings of technical and domain-specific concepts? By tracing the deployment and adoption of keywords in both collaborative exchanges and the ensuing fictions, can we identify patterns of transference? What role does shared language play in successful collaboration, and how do keywords operate both behind the scenes and within the framework of the science fiction narrative?

How do authors structure relationships between multiple knowledge domains and extend those domains into uncharted territory? When writers transgress disciplinary boundaries and, indeed, the boundaries of scientific knowledge as a whole, how will their collaborators react? In parallel, how do scientists and engineers describe their own field paradigms and establish points of consilience (or not) with other academic fields (Wilson 1998)?

What relationships will these authors, scientists and engineers develop beyond individual collaborations? Will participants develop broader collective dialogs, perhaps even an emergent paradigm for the project, or remain primarily engaged in small groups? This research question extends the concept of consilience and a collaborative lexicon to Hieroglyph itself, potentially creating opportunities for writers and researchers to create technological and narrative threads that bridge multiple stories in the final anthology.

Finally, how will readers and broader fan communities engage with the ideas embedded in these collaborations? To what extent will keyword discussions among the Hieroglyph collaborators translate to the discussions conducted among the project's subscribers and fans?

Methodology

By focusing on the definition and transmission of keywords, this project adopts a social network analysis methodology to the process of collaboration (Wasserman and Faust 1994). As new conceptual keywords are introduced in various collaborations (i.e. "symmetry" or "solar sail") they will be mapped based on the sender(s) and recipient(s) of the relevant correspondence. Collaborators will have a number of communication options available to them, ranging from email and in-person interactions to private small-group blogs, public restricted-authorship blogs and totally open public forums.[1] By mapping these networks over time, we can identify relationships between keywords and evolving groups of participants.

The project will identify keywords as they propagate across subject domains, adopting an approach inverse to the typical applications of statistical modeling (where algorithms are used to identify key terms within specific knowledge domains or discourses). For this reason, the project will combine statistical modeling Termine or another term extraction tool with direct interventions, including user-initiated tagging, surveys and editorial analysis (National Centre for Text Mining). For example, collaborators will be asked to identify and define active keywords in their conversations at regular intervals, and editors will offer their own keyword analyses of the writers' work in progress. By combining these different quantitative and qualitative approaches, we can address the research questions listed above at the level of individual actors, small collaborative teams and broader discourse groups.

These multilayered groupings of terms and participants will inform a "middle ground" reading of the collaborations in Hieroglyph, looking at the collective crafting of fact-driven narratives as a form of cultural production. By including the interactions of the site's reading public as well as its official collaborators, the project also allows us to explore the discourse boundaries between scientists, engineers, science fiction writers, science fiction fans and other groups.

Conclusion

Hieroglyph defines a new direction in the study of scientific collaboration by embedding the importance of narrative and broad public engagement into the core of the collaborative process. Asking scientists and engineers to work directly with science fiction writers (and vice versa) forces each group to question assumptions and move beyond the terministic screens that define their individual artistic practices and knowledge domains (Burke 1968). Hieroglyph seeks to open up a creative space for experimentation by all participants, allowing them to step outside of traditional discourses and explore new ideas through new language.

This research project will trace the success or failure of the Hieroglyph experiment using a mixed-methods approach including keyword extraction, tagging, user surveys and traditional literary close reading of the collaborative fiction.

Its middle ground approach will allow us to see how writers, scientific experts and the general public interact in the making of science fiction, that most accessible literature of ideas.

References

Bourdieu, P. (1993). *The Field of Cultural Production*. Ed. Randal Johnson. New York: Columbia University Press.

Burke, K. (1968). *Language as Symbolic Action: Essays on Life, Literature, and Method*. University of California Press.

Finn, E. (2012). New Literary Cultures: Mapping the Digital Networks of Toni Morrison. In Lang, A. (ed.) *From Codex to Hypertext: Reading at the Turn of the Twenty-First Century*, University of Massachusetts Press.

Guillory, J. (1993). *Cultural Capital: The Problem of Literary Canon Formation*. University of Chicago Press.

Latour, B. (2005). *Reassembling the Social: An Introduction to Actor-Network-Theory*. Oxford University Press.

National Centre for Text Mining. *TerMine*. http://www.nactem.ac.uk/software/termine/ .

Wasserman, S., and K. Faust. (1994). *Social Network Analysis: Methods and Applications*. 1st ed. New York: Cambridge University Press.

Wilson, E. O. (1998). *Consilience: The Unity of Knowledge*. New York: Knopf.

Notes

1. While it will probably prove impossible to capture every interaction, the Hieroglyph digital and procedural structure is designed to archive most exchanges.

A catalogue of digital editions

Franzini, Greta
greta.franzini@gmail.com
Centre for Digital Humanities, UCL, GB

The focus of my doctoral studies at the UCL Centre for Digital Humanities is the creation of a digital edition of the oldest surviving manuscript of S. Augustine's *De Civitate Dei.* The manuscript dates back to the early fifth century and most of the existing, scarce research we have predates the 1950s. Its much debated provenance and authorship, due to it being contemporary to Augustine himself, are as intriguing as its rare palaeographical features and marginalia. My research seeks to, firstly, examine best practice in the field of digital editions by collating relevant evidence in a detailed catalogue of extant digital editions. The catalogue records features, scope, philological as well as technological aspects of each edition and aims at becoming a collaborative scholarly endeavour for the benefit of the Digital Humanities community. Secondly (and consequently), lessons learnt from the catalogue will inform the production of an electronic edition of *De Civitate Dei,* which will include transcriptions of the text and the scholia, high-definition images, a short critical apparatus, as well as background information and links to relevant resources.

What makes a good digital edition? What features do digital editions share? What is the state of the art in the field of digital editions? Why are there so few electronic editions of ancient texts?

To address these questions, I have collated relevant evidence in a detailed catalogue of digital editions. Amongst other things, the catalogue makes a distinction between scholarly and non-scholarly editions; provides a list of tools used; open source and open access projects which will help flag up potential data links; funding bodies which will serve as a popularity record; digital humanities standards compliant projects (TEI, Creative Commons License, linked open data, etc.); and texts under examination as well as their repositories as a means of assessing which countries, areas or cultural institutions are more actively digitising.

Why does this project bring to the Digital Humanities community? A catalogue of digital editions is greatly beneficial as it provides:

- an accessible, unique record of which texts have had digital editions created and the historical period they belong to;
- a data bank of features, tools, licences, funding bodies and locations;
- an insight into past, present and future projects;
- the possibility of viewing trends or patterns (e.g. what time periods are most covered or which institutions produce the largest number of digital editions);
- a platform where collaborators can engage in live discussions and update information as it becomes available;
- a means of identifying which areas need to be improved.

The editions I include in the catalogue come from numerous sources and their selection follows basic criteria: the electronic texts can be ongoing or complete projects[1] , born-digital editions or electronic reproductions of print volumes. These were gathered from existing catalogues[2]

, lists, such as *Projects using the TEI* [3] , RSS feeds,[4] publications (articles, reviews and books), Google Scholar alerts, tweets, word of mouth, web browsing and chaining.

Data is carefully collected and assessed both quantitatively and qualitatively. Content analysis is being carried out along two parallel tracks: a passive approach, whereby I contact each team with a short questionnaire aimed at gaining a deeper understanding of both the production and user ends of the project; and a more active, observational examination of the electronic editions through website and related publications analysis.

The catalogue has been produced using existing tools, Google Spreadsheets and Google Fusion Tables, whose progressive development offers new opportunities in the field of data collection and visualisation. This choice was dictated by a number of factors, namely cost, ease, collaboration, functionality and output.[5] Google Spreadsheets and Tables also provide inbuilt sharing and communication tools, opening up possibilities for live, collaborative and synergetic work.

Once all the data has been analysed, it will be possible to establish the state of the art in the field of electronic editing, draw up a best practice profile and make reliable inferences from which further research can stem and develop.

To date, the catalogue showcases some three-hundred digital editions (this is the estimate figure for July 2013), collected and examined over a period of ten months. Of course, there are many more editions left to include and, indeed, many more to come.

However, interesting facts are already beginning to emerge: several projects, for instance, have not set up analytics as a means of studying usage; projects urging the digital reunification of manuscript fragments are often internally fragmented themselves, having split the project between institutions rather than centralising the material for easy retrieval and management; and TEI guidelines are not as widely adopted in the field of digital editions as we might think.

While initially collated for personal research purposes, I am developing the catalogue into a larger resource, available at: https://sites.google.com/site/digitaleds. The website enables people to report bugs and errors, comment and make suggestions for improvement. Although initially curated by myself, a wider group of administrators is envisaged for a more reliable and smoother experience: regular and prompt updates, continuous support and wider outreach. Scholars can join the project as administrators or editors by contacting the author.

The ultimate aim of the catalogue is not only to be used as a project reference tool but also to bring together scholars in the field, thus systematically and collaboratively creating a unique bank of data which would figure alongside other prominent Digital Humanities resources such as *centerNet,* the *Digital Classicist Wiki* and the various associations (*ADHO, ACH, ALLC,* etc.).

Notes

1. Still active on the web
2. Dr. Patrick Sahle's *Catalog of Digital Scholarly Editions* (available at: http://www.uni-koeln.de/~ahz26/vlet/vlet-about.html); Dr. Paolo Monella (available at: http://goo.gl/smy6p , section 2.2); Dr. Cinzia Pusceddu (available at: http://www.digitalvariants.org/e-philology); Dr. Aurélien Berra (available at: http://philologia.hypotheses.org/corpus); the *Monastic Manuscript Project* (available at: http://earlymedievalmonasticism.org/listoflinks.html#Digital); Hunter College (available at: http://goo.gl/cBjci); the *Digital Classicist* (available at: http://goo.gl/r8eUt and http://goo.gl/GSpLl) and the *Associazione per l'Informatica Umanistica e la Cultura Digitale* wikis (available at: http://www.digitalclassicist.org/wip/index.html). Another notable catalogue is UCLA's *Catalogue of Digitized Medieval Manuscripts* which, however, records some 3126 fully digitised manuscripts as opposed to digital editions (available at: http://manuscripts.cmrs.ucla.edu/index.php).
3. Available at: http://www.tei-c.org/Activities/Projects/index.xml (Accessed: 2 March 2013).
4. The *Ancient World Online* (available at: http://ancientworldonline.blogspot.co.uk/); arts-humanities.net (available at: http://www.arts-humanities.net/); *Digital Classicist* seminars (available at: http://www.digitalclassicist.org/wip/index.html).
5. The data can be exported in different formats: Microsoft Excel (.xlsx); Open Document Format (.ods); PDF Document (.pdf); Comma Separated Values (.csv); Plain Text (.txt); Web Page (.html).

SIMSSA: Towards full-music search over a large collection of musical scores

Fujinaga, Ichiro
ich@music.mcgill.ca
Centre for Interdisciplinary Research in Music Media and Technology, McGill University, Canada

Hankinson, Andrew

andrew.hankinson@mail.mcgill.ca
Centre for Interdisciplinary Research in Music Media and Technology, McGill University, Canada

Musical scores are the central resource for music research, and research involving the capture, transmission, and analysis of these resources is a unique and largely untapped area in the Digital Humanities. For hundreds of years before the invention of audio recording music scores were the only format capable of capturing and transmitting sounds from one musician to the next. Our project, the Single Interface for Music Score Searching and Analysis (SIMSSA) targets digitized (scanned) music scores, and seeks to provide tools for searching and retrieving these resources. We seek to replicate the successes of similar initiatives for textual materials, like the HathiTrust or Google Books, in bringing large collections of musical materials to anyone with an internet connection. We have made the first steps towards this goal by developing a number of prototype systems, and have been actively seeking partnerships with music researchers and libraries.

An unprecedented number of musical scores are being made available as libraries digitize their collections. Nevertheless, there are two major challenges to using them. One is that the digitization efforts are distributed: all across the world, many different libraries, archives, and museums are digitizing their collections of music scores, both printed and manuscript, but no standards exist currently to unify these collections so that these digital scores can be easily found. The other challenge is that it is virtually impossible to perform content-based search or analysis of online scores — a sharp contrast with the situation for digitized texts. There is simply no reliable optical music recognition (OMR) software that can achieve results comparable to the optical character recognition (OCR) software that institutions use to make their text collections searchable. Until digital page images of musical scores can be converted into computer-readable format using OMR, the full potential of search, analysis, and retrieval of digital music collections is cannot be realized.

We currently have two research teams in place, developing tools to support our efforts. One team is concentrating on finding scores available as digital images on the Internet. This task requires crawling the Internet, automatically discovering digitized books, articles, and facsimiles of music sources. Each digital image is analyzed to determine whether it contains printed music. This type of large-scale music document analysis has never been attempted before, thus new and efficient algorithms need to be developed. Our preliminary study involving 659 page images, resulted in 98.7% recall; missing only 3 pages containing musical scores (false negatives). We are aware of large sites that contain music scores among the millions of books already digitized, such as Google Books, Internet Archive, HathiTrust, and the Bibliothèque nationale de France. When our system determines that there are music scores in a book, we will index the information so that in the future, each digital object will be easily locatable. In other words, we will automatically create a catalogue of digitized scores so that researchers can use a central resource to search hundreds of independent websites containing scores — much like web crawlers do today.

Our second team is working on developing content-based analysis tools for performing large-scale OMR. Current OMR tools are highly limited in scope — they can only work with a subset of music notation types, and are restricted to operating as a desktop application. We are currently developing new, web-based OMR tools that will allow us to operate large, flexible OMR systems through a web browser. We are also developing new methods of "crowdsourcing," allowing us to distribute the steps of the OMR process to a wide, global audience. This work represents a significant advance in the state-of-the-art for OMR systems.

The outcomes of the SIMSSA project will prompt further exploration into large-scale digitization, transcription, retrieval, and analysis of music documents. The larger agenda behind SIMSSA is to make all musical documents available in electronic format to the wider public — an ambitious goal, but one that has some precedent in the library and musicological domain. To achieve this goal, we recognize that there are both technological and intellectual issues that need to be addressed.

The technological outcome of the SIMSSA project will be the development of powerful software tools that are accessible and usable by our constituent communities. Through the creation of web crawling and music image indexing systems, we hope to unlock the contents of existing digital music page images and convert them into searchable and analyzable data. The creation of an advanced, online toolkit for optical music recognition will also assist researchers in performing their own content analysis, both for improving the results of our indexing system, as well as operating on their own document collections.

Intellectually, the SIMSSA project will open up the possibility of performing search and analysis on the world's musical collections, which will create new avenues of exploration in music theory and history. Currently researchers are limited to small personal data sets or to music that they have transcribed and analysed by hand. Researchers need tools that provide the ability to search across thousands of documents. This will promote discoveries about the nature of music that would have taken years or even lifetimes to do manually.

We also hope to tackle some of the issues surrounding the creation of the digital scholarly edition for music. While

the digital edition represents several important advances over the traditional print edition by virtue of being in a dynamic, interactive environment, it also presents some difficulties in how these musical editions are represented at the computational level, and the amount of complexity involved in making these editions a reality. The large-scale nature of the SIMSSA project presents a unique opportunity for practitioners in musicology, library science, and computer science to develop standards, tools, and best practices for creating the digital scholarly edition.

The SIMSSA project is well into its second year of operation and we have already presented a number of prototype projects and software packages useful for both our project, and the digital humanities in general. We have developed the Diva.js image viewer, a web-based software package that allows users to quickly and efficiently view extremely high-resolution (multi-gigabyte) document images on the web. We have demonstrated both the Liber Usualis prototype for performing musical search and retrieval, and the Salzinnes Antiphonal prototype that combines high-quality scholarly information from the CANTUS website with a unique exploration interface that allows users to see this information in situ with the original page images.

We are continuing to develop a number of important relationships with the large, national libraries (British Library, Library of Congress, Bibliotheque Nationale de France, Bayerische Staatsbibliothek) who will provide access to their digital collections. We are also working closely with the Music Encoding Initiative, an international group of scholars, technologists, and librarians based at the University of Virginia who are developing standards and best practices for the digital music edition.

Musical scores is a unique and, as-yet, untapped area for digital humanities research. We have a very limited understanding of how people use and interact with vast amounts of musical information at their fingertips, since there are no large-scale initiatives that offer this. To address this we are actively creating a community of technologists, musicologists, librarians, and other interested parties to begin to uncover the many questions that must be answered, and to explore the new areas of research that will emerge from this work.

Notes

1. http://ddmal.music.mcgill.ca/diva/
2. http://ddmal.music.mcgill.ca/liber/
3. http://ddmal.music.mcgill.ca/salzinnes/
4. http://www.music-encoding.org/

Counting Words with Henry James: Towards a Quantitative Hermeneutics

Fyfe, Paul
pfyfe@fsu.edu
Florida State University, United States of America

> Her function was to sit there with two young men—the other telegraphist and the counter-clerk; to mind the 'sounder,' which was always going, to dole out stamps and postal-orders, weigh letters, answer stupid questions, give difficult change and, more than anything else, count words as numberless as the sands of the sea […][1]

The anonymous telegrapher in Henry James 1898 London-based novella *In the Cage* has drawn significant attention in recent decades from literary critics and media historians. These scholars have read her story—of projecting herself into the domestic drama of the aristocrats who use her office—for its lessons about the anxieties of surveillance, the instability of public and private domains, the regulation of gendered information workers, and the impact of telegraphic mediation on discourse including James's own style.[2] But while *In the Cage* is rich with possible readings for the work of media and cultural history, our own critical reflex to do 'readings'—to suspect the text for what aesthetic, cultural, and historical lessons it encodes or conceals—overlooks the story's own emphasis on a very different mode of textual encounter: counting words. By the light of recent interest in quantitative literary analysis, we can see the textual transactions of *In the Cage* from other angles: as a historical signal for when 'distant reading' (as we now call it) may have become necessary, and as a provocation to leave our conventional hermeneutics for the reflective reading that counting words might actually facilitate.

'Counting Words with Henry James' undertakes to demonstrate how the digital humanities participates in the recent turn in literary theory from the 'hermeneutics of suspicion' to what Rita Felski calls post-critical or 'reflective reading.'[3] I argue that critiques of the digital humanities as being anti- or even non-theoretical fundamentally misrecognize its alliance with such recent theoretical initiatives. Instead, the self-critical methodologies of digital humanities also manifest the 'the intricate play of perception, interpretation, and affective orientation' that characterizes critical reading after

suspicion.[4] Thus, I ultimately hope to move beyond the unfortunate associations of quantitative literary analysis with 'not reading' or data crunching by offering a different theoretical vocabulary for its continuing work in partnership with literary theory. At the same time, I propose that the emerging theoretical program of 'reflective reading' can address methodological insufficiencies in how quantitative analysts move from data to interpretation or from signal to concept.[5]

My paper will share the results of ongoing text analysis and topic modeling trials of the works of Henry James at multiple levels of address. Using off-the-shelf tools like the text analysis suite Voyeur and the topic modeling toolkit MALLET, I undertake a series of experiments at an increasing scale, beginning with the text of *In the Cage* and scaling up through James's entire collected works and prefaces. Quantitative approaches to James are not new, nor are these tools necessarily cutting edge, but my argument is instead about the appropriateness and timeliness of their application for ongoing critical discussions: about *In the Cage* as well as possible affiliations of surface and distant reading. [6] *In the Cage* uniquely warrants such treatment, in two senses: first, because of its own conspicuous thematics of textual abundance, word counting, and interpretation; and second, because of James's own editorial efforts to control interpretations of *In the Cage* as the story soon joins another massive textual corpus: his collected works for the New York edition.

Throughout his novella, James describes late-nineteenth-century telegraphy with an arithmetical lexicon; the telegrapher is constantly counting, figuring, adding, calculating, working out meaning in the margins, and building a hypothesis about the relations of Captain Everard and Lady Bradeen from her process. Because, in the story, the telegrapher ultimately misreads these relations, James seems to deprecate her quantitative methods compared with his own elaborate narrative procedures. In his later preface to *In the Cage*, James describes its interest in the 'wonderment' of the telegraph office's cacophonous information field, its investigation into 'the question of what it might 'mean'.' In other words, the story is about challenges to reading practices and, more abstractly, about the methodologies of information processing that might lead to 'meaning.' And as the telegraphist tries to deal with the problem of texts at scale, she occupies a similar position of alienated curiosity with respect to digital humanists and large data sets.[7]

How would the telegrapher-as-quantitative-analyst read her own story? Or all of James's works? The telegraphist is faced with proliferating telegraphic fragments and 'words as numberless as the sands of the sea.' These messages are more than merely fragments to reconstruct: they are also units of scalable information. Their textual compression and transcoding make possible the 'massive addressability' that characterizes large collections of digitized text.[8] Text analysis and topic modeling lets us further identify different levels by which to approach meaning in the text—e.g. the word, the genre, the topic cluster, the historical trend—as well as in larger corpora in which the text signifies. Those methods have reinvigorated questions familiar to the telegraphist of how to read, count, and interpret.

In addition to thematizing problems of literary interpretation across different textual scales, *In the Cage* invites them through its own revision history. Shortly after completing the novella, James began collecting his texts into a massive New York edition, revising his works and writing new prefaces which aim to direct our readings of their individual and collective significance. While James writes in his prefaces about what scale of textual address we should use to find meaning, this novella—from its oral composition and transcription, to its narrative about fragments and questions of meaning, to its tenuous status relative to James's oeuvre—also invites us to reconsider the interpretive possibilities of words and clusters at scale. In a sense, *In the Cage* embodies the problem of textual addressability.

My reading of the novella aims to link the telegraphist's wonderment, the 'hypothesis-testing mode' characteristic of recent work in quantitative literary analysis, and emergent forms of post-critical reading in literary theory.[9] In effect, post-critical reading offers a recuperative vocabulary for counting words, rescuing the telegrapher from Jamesian suspicion and perhaps bolstering the claims of quantitative literary analysis. Its recursive processes, I suggest, likewise draws upon what Felski calls 'the intricate play of perception, interpretation, and affective orientation that constitutes aesthetic response.'[10] Ultimately, this paper argues that the digital humanities are not post-theoretical, but they may be productively post-critical in generating a reflexive, quantitative hermeneutics.

References

Anon. (2013). *Surface Reading/ Machine Reading: New Approaches to Texts and Data* Available from: http://raley.english.ucsb.edu/wp-content/surface-reading_flyer.jpg (Accessed 5 March 2013).

Clayton, J. (1997). The Voice in the Machine: Hazlitt, Hardy, James, in *Language Machines: Technologies of Literary and Cultural Production*. New York: Routledge. 209–232.

Felski, R. (2009). After Suspicion. *Profession*. 28–35.

Flanders, J. (2009). The Productive Unease of 21st-century Digital Scholarship. *Digital Humanities Quarterly*.

3(3). Available from: http://digitalhumanities.org/dhq/vol/3/3/000055.html (Accessed 10 December 2009).

Heuser, R. & L. Le-Khac (2011) Learning to Read Data: Bringing out the Humanistic in the Digital Humanities. *Victorian Studies*. 54(1). 79–86.

Hoover, D. (2007) Corpus Stylistics, Stylometry, and the Styles of Henry James. *Style*. 41 (2). 174–203.

James, H. (2002) 'In the Cage', in Wegelin, C., & H. B. Wonham (eds.) *Tales of Henry James*. 2nd edn. New York: W. W. Norton & Company.

Keep, C. (2011) Touching at a Distance: Telegraphy, Gender, and Henry James's *In the Cage*, in *Media, Technology, and Literature in the Nineteenth Century: Image, Sound, Touch*. Surrey, England: Ashgate. 239–255.

Liu, A. (2009) Digital Humanities and Academic Change. *English Language Notes*. 47(1). 17–35.

Marvin, C. (1988). *When Old Technologies Were New: Thinking About Electric Communication in the Late Nineteenth Century*. New York: Oxford University Press.

Menke, R. (2000). Telegraphic Realism: Henry James's *In the Cage. PMLA: Publications of the Modern Language Association of America*. 115(5). 975–990.

Stauffer, A. (2011) Introduction: Searching Engines, Reading Machines. *Victorian Studies*. 54.1. 63–68.

Witmore, M. (2010) Text: A Massively Addressable Object. Wine Dark Sea [online]. Available from: http://winedarksea.org/?p=926 (Accessed 8 August 2011).

Notes

1. James, 2002, p.229.
2. Notable examples include Clayton, 1997; Keep, 2011; Menke, 2000; Marvin, 1988.
3. Felski, 2009.
4. 2009, p.31.
5. For an accessible review of this problem, see Stauffer, 2011.
6. For an example of a quantitative approach, see Hoover, 2007. A recent conference at New York University was devoted to such issues: 'Surface Reading / Machine Reading: New Approaches to Texts and Data', 2013.
7. For reports of this phenomenon, see Liu, 2009; Flanders, 2009; Heuser and Le-Khac, 2011.
8. Witmore, 2010.
9. Heuser and Le-Khac, 2011, p.85.
10. 2009, p.31.

Automatic Detection of Reuses and Citations in Literary Texts

Ganascia, Jean-Gabriel
jean-gabriel.ganascia@lip6.fr
LIP6, University Pierre and Marie Curie, Paris, France; Labex OBVIL, PRES Sorbonne Universités, Paris, France

Glaudes, Pierre
pierre.glaudes@wanadoo.fr
Littérature Française, XIXe-XXIe, University Paris-Sorbonne, France; Labex OBVIL, PRES Sorbonne Universités, Paris, France

DelLungo, Andrea
adellungo@free.fr
ALITHILA, University Charles de Gaulle, Lille, France

1 Introduction

For more than forty years now, modern theories of literature insist on the role of paraphrases, rewritings, citations, reciprocal borrowings and mutual contributions of any kinds. The notion of *intertextuality* was introduced in the sixties to approach these phenomena. *PHOEBUS* is collaborative project that makes computer scientists from the University Pierre and Marie Curie (LIP6-UPMC) collaborate with the literary teams of Paris-Sorbonne University with the aim to develop efficient tools for literary studies that take advantage of modern computer science techniques.

In this context, we have developed a piece of software that automatically detects and explores networks of textual reuses in classical literature. Written in PROLOG this program has been extensively tested on Isidore Ducasse texts (Lautréamont, 2009) that are known to contain many reuses and on "La comédie humaine" (Balzac, 1976-1981) from Honoré de Balzac, which, according to (Duclos, 2012), reuses some texts of his friend Théophile Gautier (Gautier, 2002). We claim that our approach is more efficient than comparable ones, e.g. (Roe, 2012) (Büchler, Crane, Mueller, Burns, & Heyer, 2011). This abstract describes the principles on which is based this program, the significant results that have already been obtained and the perspectives for the near future.

2 Distinctions between Plagiarism, Pastiches, Citations and Textual Reuses

Before going into the detail of the description of the techniques that are used, let us note that the notions of literary textual reuse and citation, which we aim to automatically detect, have to be distinguished from two similar notions: the *plagiarism* and the *pastiche*.

The plagiarism consists in robbing the work of another, i.e. in fraudulently appropriating his/her texts, without mentioning explicitly their origin. As such, the plagiarism is considered as an unethical practice that has to be tracked and prosecuted. Many techniques have been developed to detect plagiarism that is considered as some plague, because intellectual work is stolen. By contrast, the pastiche is an artistic practice that imitates an artist, a style or a period. There is nothing wrong with it, except that it mocks well-known authors. Many great writers, for instance Marcel Proust, began by pastiches for the fun and to improve their style. Their detection is close to the identification of literary style (Dinu, Niculae, & Sulea, 2012), which requires capturing the essence of an artist's style or of a period. Halfway from detection of plagiarisms and identification of pastiches, the recognition of textual reuses and citations helps to track the literary influences and the spirit of the epoch. Some of the textual reuses and citations are conscious, other not. They may correspond to explicit — or implicit — and more or less distorted quotations. Usually, textual reuses proceed by transforming a piece of text, while citations are verbatim, but it's not always the case. When a sufficient part of the original text is kept, the fragments can be recognized. This is exactly what we attempt to do automatically here. Reuses and approximate citations are far more difficult to detect than plagiarisms, because the original fragments of text may be distorted, but far less than pastiches. Anyway, their detection could be of great interest for scholars interested in intertextuality.

3 Criteria

As previously said, text reuse and citation discovery is inspired from plagiarism detection, but it has to take into account all the alterations that may have transformed the initial text. To precise the type of distortions that affect a text, we started from a hand made study realized by Tania Duclos who shows in (Duclos, 2013) [cf. *figure 1*] how some parts of the *Human Comedy* (Balzac, 1976-1981) reuse fragments of texts from Théophile Gautier.

For instance, some of the passages highlighted in *figure 1* are identical, while others are somehow different. For instance "*en brocatelle à plis soutenus et puissants, s'entouraient de fraises godronnées*" becomes "*de brocatelle aux plis soutenus et puissants, les hautes fraises godronnées*" and "*des manches à crevés et à sabots de dentelles d'où la main sortait comme une fleur de sa capsule*" becomes "*les manches à crevés et à sabots de dentelles, dont la main sort comme le pistil du calice d'une fleur*". Lastly, some fragments look far more difficult to identify, because they are composed of isolated words or even different words (e.g. "*diamants*" and "*pierreries*" or "*tableau*" and "*gravures*") of which meanings are closed. Here, we try to detect string homologies where some words may be missing, especially stop words, i.e. articles, pronouns or prepositions. It may also happen that the number and the genre of nouns, adjectives or verbs change as when "*d'une main mignonne frappée de fossettes*" is transformed in "*des mains mignonnes frappées de fossettes*".

Béatrix (H. Balzac)	Jenny Colon – Portraits (Th. Gautier)
« Si elle pouvait par un artifice quelconque porter le costume dues ~~anci~~ temps où les femmes avaient des corsets pointus à échelles de rubans s'élançant minces et frêles de l'ampleur étoffée des jupes en brocatelle à plis soutenus et puissants, s'entouraient de fraises godronnées cachaient leurs bras dans des manches à crevés et à sabots de dentelles d'où la main sortait comme une fleur de sa capsule, et qui rejetaient ~~leurs~~ les mille boucles de leur chevelure ~~sur leurs épaules~~ au delà d'un chignon ficelé de pierreries, elle lutterait avec avantage ~~eont~~ avec les beautés les plus célèbres que vous voyez vêtues ainsi dit-elle en montrant un tableau à Calyste, ### debout, ~~devant un~~ tenant ~~une m~~ un papier et chantant avec un seigneur brabançon, pendant qu'un nègre verse dans un verre à patte du vieux vin d'Espagne et qu'~~une~~ la vieille femme de charge arrange des biscuits. »	« Les costumes romanesques de Piquillo conviennent beaucoup au type de beauté de Mlle Colon; les grandes robes de lampas ou de brocatelle aux plis soutenus et puissants, les hautes fraises godronnées et frappées à l'emporte-pièce, comme on en voit dans les dessins de Romain de Hooge; les manches à crevés et à sabots de dentelles, dont la main sort comme le pistil du calice d'une fleur, les feutres à ganse de perles, à plumes crespelées, les chaînes et les rivières de diamants écaillant d'étincelles papillotantes la blancheur mate de la poitrine, les corsets pointus à échelles de rubans s'élançant minces et frêles de l'ampleur étoffée des jupes: - toute la toilette abondante et fantasque du seizième siècle s'adapte merveilleusement à la physionomie de Mlle Colon, que l'on prendrait, dans un de ses costumes capricieux, pour un de ces belles dames des gravures d'Abraham Bosse, qui marchent gravement une tulipe à la main, suivies du petit page nègre qui porte leur queue, leur chien et leur manchon, dans les allées bordées de buis d'un parterre du temps de Louis XIII. »

Figure 1:
example of hand coded comparison (Duclos, 2013) between a fragment of Béatrix (Balzac, 1976-1981) on the left and a fragment of Théophile Gautier (Gautier, Portraits contemporains, 1874) on the right.

4 Detection of Fragments with Holes

Among the more efficient existing plagiarism detection techniques, many are based on fingerprints built with the hash coding of character strings (Potthast, Eiselt, Barron-Cedeño, Stein, & Rosso, 2011), (Potthast, Stein, Barron-Cedeño, & Rosso, 2010), (Burrows, Tahaghoghi, & Zobel, 2006). Other techniques evaluate the statistical distribution of vocabulary with a vector space model of the texts and a cosine similarity measure that evaluate their closeness, however they don't seem to be appropriate to our purpose, even if we use the distribution of words.

We have implemented and adapted the fingerprint method and we have evaluated it on the reuses isolated by Tania Duclos. It helped us to optimize the values of the different parameters. To do this, we have first eliminated the "stop words", i.e. article, preposition, pronoun, auxiliary verbs, etc. We have also used the Snowball (Porter, 2001) (Tomlinson, 2004) stemmer to reduce the words to their root, which allows being independent from the inflected forms used in the text. For instance, the words "fishing", "fished", "fish", "fishes" and "fisher" are reduced to the same root word "fish".

The second part consists in extracting sequences of words characterized by their minimal size, i.e. by the minimal number of consecutive non-"stop words" they contain, which we call the window size. In addition, because we want to allow missing words, we also introduce possible holes. This means that a window of size 4 does not necessarily correspond to 4 consecutive words.

Once the similar fragments are discovered, they are adjoined end to end, which build blocs. Lastly, we have manually defined what we call "weak words", which are not very significant, and we filter the blocs of similar words of which number of non-"weak words" is bigger than a minimal threshold, for instance 4. This allows eliminating noise, without loosing a lot of information.

5 Obtained Results

The program has been implemented in SWI-Prolog (cf. (SWI-Prolog's home)) using an external table to store hash-coded texts. It is quite efficient, for instance it took less than 10 minutes to index all the Balzac's *Human Comedy* (Balzac, 1976-1981) that contains more than 25 millions of characters on a 2GHz MacPro. Then, it takes a couple of minutes to discover text reuses on entire novels.

Using this program, we were able to retrieve all the handed coded reuses of (Duclos, 2013), except the "yellow" one (see *figure 1*). We have also detected many interesting citations and reuses, for instance a reuse of the Gautier's Novel entitled "Mademoiselle de Maupin" (Gautier, Romans, contes et nouvelles, 2002) in the Balzac's Novel "Modeste Mignon" (Balzac, 1976-1981), which has not been mentioned before, or a citation of Lyttleton both in "Delphine" (de Staël, 1869) and in "Ursule Mirouët" (Balzac, 1976-1981). We also tested the system between Lautréamont's work (Lautréamont, 2009) and Buffon one the one hand and the French moralists like Pascal, La Rochefoucauld or La Bruyère on the other. We have retrieved many text reuses among which some interesting distortions like, for instance the Pascal aphorism "*Nous naissons injustes; car chacun tend à soi: cela est contre tout ordre*." that has been rewritten in "*Nous naissons justes. Chacun tend à soi. C'est envers l'ordre*."

6 Perspectives

For the near future, we plan to extensively use our system in many fields of literature, especially on the 19th century French literature, with Balzac's work, which is the aim of the PHOEBUS project funded by the CNRS. More precisely, PHOEBUS is intended to investigate the textual reuses between the Balzac's youth novels and the *Human Comedy*, and between Balzac's work and his contemporaries' work like Théophile Gautier, Benjamin Constant, George Sand etc. We also plan to digitalize the journals where many authors published either under their own names, or anonymously and to compare them with the *Human Comedy*. Lastly, we will conduct a thorough comparison with similar approaches.

References

Büchler, M., G. Crane, M. Mueller, P. Burns, and G. Heyer (2011). One Step Closer To Paraphrase Detection On Historical Texts: About The Quality of Text Re-use Techniques and the Ability to Learn Paradigmatic Relations. in Thiruvathukal, G. K. & Jones, S. E., (Éds). *Journal of the Chicago Colloquium on Digital Humanities and Computer Science*.

Balzac, H. (1976-1981). *La comédie humaine* (Vol. I-XII). (C. L. Pléiade, Éd.) Paris: Gallimard.

Burrows, S., S. Tahaghoghi, and J. Zobel(2006). Efficient Plagiarism Detection for Large Code Repositories. *Software — Practice and Experience*. 37. 151-175.

de Staël, G. (1869). *Delphine*. Paris: Garnier frères.

Dinu, L. P., V. Niculae, and O. M. Sulea (2012). Pastiche detection based on stopword rankings. Exposing impersonators of a Romanian writer. EACL 2012 — *Workshop on Computational Approaches to Deception Detection*. Avignon: Association for Computational Linguistics. 72-77.

Duclos, T. (2013). *L'intertextualité dans une Fille d'Eve et Béatrix d'Honoré de Balzac*. Paris: Sorbonne University.

Gautier, T. (1874). *Portraits contemporains*. Paris: Charpentier et Cie.

Gautier, T. (2002). *Romans, contes et nouvelles*. Pléiade, C. D. (ed.) Paris: Gallimard.

Lautréamont. (2009). *Œuvres complètes*. Pléiade, C. D. (ed.) Paris: Gallimard.

Porter, M. (2001). *Snowball: A language for stemming algorithms*. http://snowball.tartarus.org/texts/introduction.html (consulté le October 22, 2012).

Potthast, M., A. Eiselt, A. Barron-Cedeño, B. Stein, and P. Rosso (2011). Overview of the 3rd International Competition on Plagiarism Detection. Petras D. V. , P. Forner, and P. D. Clough (eds.), *Notebook Papers of CLEF 11 Labs and Workshops*.

Potthast, M., B. Stein, A. Barron-Cedeño, and P. Rosso (2010). An Evaluation Framework for Plagiarism Detection. Dans Huang, C.-R. and D. Jurafsky (eds.), *23rd International Conference on Computational Linguistics (COLING 10)*. Stroudsburg, Pennsylvania. 997-1005.

Roe, G. R. (2012). Intertextuality and Influence in the Age of Enlightenment: Sequence Alignment Applications for Humanities Research. *Digital Humanities*. Hamburg.

SWI-Prolog's home. SWI-Prolog: http://www.swi-prolog.org/ (consulté le Octobre 22, 2012).

Tomlinson, S. (2004). Lexical and Algorithmic Stemming Compared for 9 European Languages with Hummingbird SearchServer TM at CLEF 2003. Dans Peters, C. (ed.), *Working Notes for the CLEF 2003 Workshop*. Springer.

Agent-Based Modeling and Historical Simulation

Gavin, Michael
michael.a.gavin@gmail.com
University of South Carolina, United States of America

Overview

This paper will discuss Agent-Based Modeling (ABM) and its application in the humanities, with special focus on questions of concern to literary history. I begin with an introduction to ABM. Unlike text mining, topic modeling, and social-network analysis, which apply quantitative analysis to already existing text corpora, ABM creates a simulated environment and measures the interactions of individual agents within that environment. Like video games, agent-based models simulate rule-bound behaviors and generate outcomes based on those rules. However, unlike most games, where the "procedural rhetoric" of the game "persuades" users (Bogost), ABM does not depend on human interaction, but can be run many times with changing variables. Researchers can alter the parameters of agent behavior and compare how different models generate different outcomes. In the fields of ecology, economics, and political science, ABM has been used to show how the behaviors of individual entities—microbes, consumers, and voters— collectively alter large emergent phenomena. ABM offers a promising new way to approach long-standing humanistic questions, such as how literary genres change over time, how publics form and transform, how consumer markets influence authors, how ideologies move across national boundaries, or how family structures affect reading practices.

The bulk of my presentation will concern ABM in general and the theoretical issues posed by its application to historical data, but I will also discuss our ongoing project: a computer simulation of English print culture of the seventeenth century. Using NetLogo ABM software, we attempt to simulate some of the ways literary and political arguments moved through various textual media (print, manuscript, gossip, and the stage) of the seventeenth century. As the agents in our simulation trade texts, their networks of affiliation create "opinion formations" (groups of like- minded agents), helping us to see how literary and political values were transmitted in tandem. Book historians have shown how the history of ideas is inextricable from the networks of textual circulation through which ideas move. ABM will allow researchers to see these connections in a new way and, most importantly, to test textual relationships in a controlled, manipulable environment.

This paper aims to introduce ABM to digital humanists, not (just) because ABM is "the next big thing," but also to enrich ABM by importing a "more comprehensive understanding of human decision-making ... needed to move the technique forward" (Williams 2012). The rich traditions and methods of humanistic inquiry have much to contribute to this new and important computational method.

Background: What is Agent-Based Modeling?

Computer simulation has emerged in the past two decades as an important alternate method of quantitative analysis. The social and biological sciences, in particular, have benefited from the adoption of NetLogo, a simple-to-use open-source program which has opened ABM to a wide range of new inquiries (Railsback). In archeology, researchers used ABM to test possible causes for the disappearance of the Kayenti Anasazi, a prehistoric precursor of the Pueblo cultures (Dean, et al.), but ABM has made few inroads into the humanities. In the humanities, quantitative approaches have tended to involve the manipulation of text. Agent-based modeling differs significantly from topic modeling, which identifies and visualizes "lexical relationality among literary works" (Piper and Algee-Hewitt; McCarty). Such techniques identify word-clusters that represent core ideas and demonstrate how these clusters relate to each other within an existing

field of text-data. Similarly, text-mining and text-mapping, sometimes called "culturomics," find large-scale patterns, which sometimes confirm and sometimes unsettle prior expectations (Michel et al.; Wilkens; Underwood; Lieberman). ABM is also concerned with identifying relationships across large bodies of data. However, ABM differs from these methods significantly by moving away from a text-centric understanding of history. ABM instead simulates the causal forces that motivate change in complex systems like historical societies.

ABM is most similar to Game Simulation (GS), and my discussion will briefly outline points of overlap. Historical gaming is an important genre, and research communities like PlayThePast.org are beginning to identify the pedagogical value of historical games. Like GS, ABM can be said to create a "problem space" where the "affordances" and "constraints" of the space dictate player behavior (McCall). Much of what Ian Bogost has said about games applies to ABM as well: "Games represent how real and imagined systems work, and they invite players to interact with those systems and form judgments about them" (Bogost). ABM shares with GS this attention to systems of interaction. Accordingly, there has been a new push to incorporate ABM into GS in order to provide better simulations for gameplay (Arai et al.; Bonnett).

Project Description: Simulating the History of Print Controversy

How do texts move through social networks? How does the movement of texts affect the movement of ideas? Although detailed financial records from the early print era are necessarily sparse, we know a great deal about the history of publishing *as a system* (Darnton; St. Clair; Raven). Book historians have developed complex qualitative models for understanding how books were produced, sold, and distributed. The most famous such model is Robert Darnton's "communications circuit," which provides a schema for describing the production of books as they flow through various contact points in the book trade. Our models point to three specific points of interaction in the circuit:

I: How do readers decide to purchase books?

Working from William St. Clair's "reader-led model" for understanding the communications circuit, my first model will examine how readers' preferences influence the decisions of printers and booksellers (St. Clair). Historians know a great deal about what books were published when and about the mix of old and new titles that circulated through bookshops. This simulation will describe when and how books run through multiple editions, when they are kept for sale, and when they are reverted to pulp.

II: How does censorship affect controversial political writing?

The print marketplace was tightly controlled by several governing forces. Some of these were official institutions, like the state Licenser and the Stationers' Company (the professional association of printers and booksellers), but punishments were also meted out by influential individuals offended by books. In this model, I will adapt Joshua Epstein's simulation of 20th- century political oppression to the seventeenth-century context in order to examine how the circulation of controversial books was affected by censors (Epstein).

III: How do books change readers' opinions over time?

ABM has been used to describe voting patterns and "opinion formations" (Afshar and Asadpour; Lorenz). Following this research, my third model will examine how readers' opinions change over time in response to political rhetoric. The reader-agents will hold an array of opinions along a continuously sliding scale from [-1] to [+1]. Agents respond to the opinions of others based on the confidence they placed in the author and in their proximity within continuously shifting social networks.

References

Afshar, M., and M. Asadpour (2010). "Opinion Formation by Informed Agents." *Journal of Artificial Societies and Social Simulation* **13 (4).**: 5.

Arai, K., H. Deguchi, and H. Matsui (2006). *Agent-Based Modeling Meets Gaming Simulation*. Springer.

Bogost, I. (2007). *Persuasive Games: The Expressive Power of Videogames*. Cambridge, MA: MIT Press.

Bonnett, J. (2007). Charting a New Aesthetics for History: 3D, Scenarios, and the Future of the Historian's Craft. *L'histoire Sociale/Social History* 40. 169–208.

Darnton, R. (2006). What is the History of Books? in *Book History Reader*. 2nd edn. Routlege.

Dean, J., J. Epstein, A. Swedlund, M. Parker, and S. McCarroll Understanding Anasazi Culture Through Agent-

Based Modeling. Sante Fe Institute. http://www.santafe.edu/media/ workingpapers/98-10-094.pdf .

Lieberman, M. (2012). The 'Dance of the P's and B's': Truth or Noise? *Language Log*. http://languagelog.ldc.upenn.edu/nll/?p=3730 .

Lorenz, J. (2007). Continuous Opinion Dynamics Under Bounded Confidence: a Survey. *International Journal of Modern Physics* **18** 1819–38.

McCall, J. (2012). Historical Simulations as Problem Spaces: Criticism and Classroom Use. *Journal of Digital Humanities*.

McCarty, W. (2004). Modeling: a Study in Words and Meanings. in *Companion to the Digital Humanities*. Blackwell.

Michel, J.-B., et al. (2011). Quantitative Analysis of Culture Using Millions of Digitized Books. *Science* **331** 176–182.

Piper, A., and M. Algee-Hewitt (2012). About in *Literary Topologies*. http://literarytopologies.org/?page_id=15 .

Railsback, S. (2012). *Agent-based and Individual-based Modeling a Practical Introduction*. Princeton University Press.

Raven, J. (2007). *The Business of Books*. Yale University Press.

St. Clair, W. (2004). *Reading Nation in the Romantic Period*. Cambridge.

Underwood, T. (2012). Where to Start with Text Mining. in *The Stone and the Shell.* http://tedunderwood.wordpress.com/2012/08/14/where-to-start-with-text-mining/ .

Wilkens, M. (2012). "Canons, Close Reading, and the Evolution of Method." In *Debates in the Digital Humanities,* ed. Gold, M. K. 249–58. Minneapolis: University of Minnesota Press.

Williams, S. (2012). *Agent-Based Modeling: History and Applications. Conference Presentation.* Quantifying Social Fields. held at Berkeley, CA. http://www.irle.berkeley.edu/culture/ conf2012/williams-agents12.pdf

The Digitized Divide: Mapping Access to Subscription-Based Digitized Resources

Gooding, Paul Matthew
paul.gooding.10@ucl.ac.uk
University College London, United Kingdom

Introduction

This paper will present the findings from a PhD case study into the use and users of the British Library Nineteenth Century Newspapers Collection (BNCN).[1] Using data gathered from web analytics and user surveys, it will show that although digitization provides clear benefits to users who operate in an information-rich environment, these benefits are distributed unequally. I will therefore present an alternative geographical visualization based upon the location of subscribing institutions rather than individual users. This, combined with university rankings data and relative poverty measures, backs up the main argument of this paper: that the subscription-based model of digitization severely undermines the rhetorical embracement of universal access, and instead reinforces existing divides between information-rich and information-poor communities.

Addressing the Digital Divide

Web analytics services such as Google Analytics[2] provide map overlays that automatically visualize user locations. This important data provides web analysts with important insight into their user base. These automated mapping tools lack, though, a consideration of how user location is influenced by variations in access to digitized resource. The internet, and by extension, digitized collections, are generally viewed as an opportunity to widen participation and improve education (Norris 2001, 7; Bell 2005), but this paper will demonstrate that they can provide these benefits unevenly across society. Norris identifies that the digital divide is multifaceted; it is a global and democratic divide, but also a social divide between different groups in society (Norris 2001, 4).

This is important when considered alongside a common problem in the literature: increased quantities of digitized content have brought with them inflated expectation levels. Everett, for instance, wrongly conflates the digitization of large collections with the concept of universal access: "the problem for the twenty-first century scholar will be to limit inquiry to a manageable subset of data; because all scholars will have immediate access to all archives in the world" (Everett 2005). Commercial reality, though, makes this utopian outlook seem naïve. The social and professional environment in which users operate remains vital in deciding access to digitized content. This mirrors the wider context of the digital divide, which is increasingly manifested as an indicator of the differentiated uptake of important digital resources (Hargittai & Walejko 2008; Hassani 2006; Norris 2001; Castells 2002). Existing

research into this "second-level digital divide" (Hassani 2006) suggests that those from higher socioeconomic backgrounds generally benefit most from technological developments. As a result, mapping the location of individual users may show nothing more than high levels of connectivity in a particular demographic (Hassani 2006, 251). Similarly, in academia it appears that a 'digitized divide' could emerge between those with access to digitized content in large quantities and those without, one strongly related to social and geographical status.

Methods

While Geographical Information Systems (GIS) have the potential to answer innovative research questions (Bodenhamer et al. 2010), the automated nature of default web analytics visualizations does not allow us to interrogate this problem.

In order to study the impact of BNCN, web data was analysed for a period of one calendar year using reports generated by Google Analytics, including the referral analysis (Madsen 2010)[3] which influenced this paper; the results demonstrated that many users located outside Europe and the USA were in fact linked to institutions within these two regions. A survey was also mounted, which showed that many BNCN users still found access difficult; indeed some were forced to travel in order to access the collection digitally. The GIS was therefore created to test concerns about potential inequalities in access.

The visualization is based upon a list of institutions with current subscriptions to BNCN, including educational institutions, UK public libraries, and national libraries. This list was collated using online subscriber lists,[4] manual searching and referrer lists derived from web analytics. These were then mapped to a web-based GIS using Google Maps Fusion Tables,[5] and combined with demographic information and university ranking data. A separate layer was created for English public libraries which combined access information with UK Government measures of relative deprivation, population, and public spend on libraries.

Findings

The findings demonstrate that the divide in access correlates strongly with the status of a university: more highly ranked institutions were more likely to have current subscriptions. Additionally, English public library authorities were far more likely to have access if they were in less deprived regions, or served a relatively large population. This backs up qualitative data from the survey; some respondents were worried about the impact of working at institutions without appropriate subscriptions, and the prospect of losing access when fixed-term academic posts expire. This mirrors Hargittai's assertion that "the societal position that users inhabit influences aspects of their digital media use such as the technical equipment to which they have access" (Hargittai 2008, 940). As digitized content proliferates, the expectation that scholars will use it also increases. The idea of democratized access to digital content (Bell 2005) can therefore become a damaging myth for those left behind. Unequal access to resources has been a longstanding problem for researchers, but the rhetorical shift towards universal access has ignored it. Technological inequality never entirely disappears, as Castells points out: "as one source of technological inequality seems to be diminishing, another one emerges" (Castells 2002, 256).

Second, rather than facilitating the disintermediation of information, the current glut of commercially digitized content increases the importance of library services in relation to access. In the wake of the Google Books project, Roush questioned "the 'value proposition' they [libraries] offer in a digital future" (Roush 2005)., These findings suggest that this value proposition will centre on the library's ability to supply relevant subscriptions to its users in a timely manner, for as long as access is too costly to maintain at an individual level.

Conclusion

The study of web impact would benefit from a more realistic appraisal of the digital divide in relation to digitized content. Commercial digitization, while an effective way to fund projects, has implications for scholars working outside information-rich institutions, or indeed outside institutional frameworks entirely. Similar case studies should therefore be done with other digitized resources to discover whether these patterns are replicated elsewhere. Additionally, we must consider how project developers and library services can work to address the inequalities discovered by this project. In keeping with the conference's theme, this paper provides an analysis of how far digitization truly provides researchers with freedom to explore the content being produced.

References

Bell, D. (2005). The Bookless Future: What the Internet is Doing to Scholarship. *The New Republic*. http://www.tnr.com/article/books-and-arts/the-bookless-future (accessed March 27, 2012).

Bodenhamer, D. J., J. Corrigan, and T. M. Harris (eds.) (2010). *The Spatial Humanities: GIS and the Future of Humanities Scholarship*, Bloomington: Indiana University Press.

Castells, M. (2002). *The Internet Galaxy: Reflections on the Internet, Business, and Society*, Oxford: Oxford University Press.

Everett, G. (2005). Electronic Resources for Victorian Researchers — 2005 and Beyond. *Victorian Literature and Culture*. 33. 601–614.

Hargittai, E. (2008). The Digital Reproduction of Inequality. In Grusky, D. (ed). *Social Stratification*. Boulder, CO: Westview Press.

Hargittai, E. & G. Walejko. (2008). The Participation Divide: Content Creation and Sharing in the Digital Age. *Information, Communication & Society*. 11(2). 239–256.

Hassani, S. N. (2006). Locating Digital Divides at Home, Work, and Everywhere Else. *Poetics*. 34(4-5).

Madsen, C. (2010). What is Referrer Analysis? *Toolkit for the Impact of Digitised Scholarly Resources (TIDSR)*. http://microsites.oii.ox.ac.uk/tidsr/kb/45/what-referrer-analysis (accessed June 11, 2012).

Norris, P. (2001). *Digital Divide: Civic Engagement, Information Poverty and the Internet Worldwide*, Cambridge: Cambridge University Press.

Roush, W. (2005). The Infinite Library. *Technology Review*. 108(5) 54–59.

Notes

1. newspapers.bl.uk/blcs/
2. www.google.com/analytics/
3. See the Toolkit for the Impact of Digitized Scholarly Resources: http://microsites.oii.ox.ac.uk/tidsr/.
4. http://gdc.gale.com/products/19th-century-british-library-newspapers-part-i-and-part-ii/evaluate/customer-list/ and http://www.bl.uk/reshelp/findhelprestype/news/database.pdf
5. https://developers.google.com/maps/

Schooling the Scholar, Poaching the Fan: Fannish Intellectual Production and Digital Humanities Methods

Goodwin, Hannah
hmlgoodwin@gmail.com
University of California, Santa Barbara, United States of America

D'Silva, Alston
alston.dsilva@gmail.com
University of California, Santa Barbara, United States of America

Scholars have often maintained a critical distance between their own forms of knowledge making and those of fans, a distance that we find worth interrogating in the setting of digital humanities. After all, the kind of mapping and charting of vast amounts of cultural data that digital humanists are beginning to do seems closely tied to fan practices of collective textual analysis (and production). We argue against a general academic hesitation to seriously incorporate or interrogate fannish intellectual production as compatible with academic work, and we see digital humanities as a promising site for cooperation between fans and academics. Indeed, digital humanities appear indebted to the intellectual productions of fan communities in an age of media convergence.

Specifically, this paper engages with fan material in a way that acknowledges how it may inform and work alongside academic work in television studies. Fan communities' collective work analyzing and producing digital work around television shows, including mapping characters and their affective relationships, has heavily influenced our own television analysis method, which involves graphing and analyzing social networks of television characters. This method also draws on the academic work of Franco Moretti and other digital humanities scholars who have demonstrated the usefulness of social network analysis of various texts, but who have not discussed the relationships between their analysis and that of fans. We use digital graphing tools to investigate affective relationships, as defined primarily by fans, between television characters across lines of racial or sexual "difference" in ensemble character-driven dramas. Our data sets have in some cases been gleaned from fan forums, from which we have constructed affective social network graphs using tools like Gephi, ManyEyes, and Mathematica. Analysis of these graphs, we found, illuminates trends and "underlying structures" that might otherwise be difficult to notice (Moretti).

To explore the usefulness of graphing social networks, we demonstrated the absence of queer characters and relationships in the television series *Lost*, a show with a sizable multi-ethnic cast that draws heavily on the sexual, affective, and familial tension of its characters for dramatic effect. First establishing the overwhelming preponderance of heterosexual pairings, we reorganized the nodes so that the most frequently appearing characters gravitate

towards the center of the field, and the less frequently appearing characters radiate outwards. As we might expect, characters who appear most frequently have multiple lines of (heterosexual) relationships, while less frequently appearing characters are likelier to be unattached or to have only one relationship over the course of the series. The suggestion of an undergirding heterosexual matrix, visible here, is borne out in the conclusion of the show. In a "flash-sideways," the storyline in a parallel timeline where Oceanic Airlines Flight 815 never crashes, characters recall the moments on the island when they make contact with their principal love interests. This forces a heteronormative pairing as an organizing force (possibly queering the pair of John Locke and Ben Linus, who recall their past lives but without the benefit of a significant other). The graph also makes legible the outliers, exceptions, and outsiders that relate to the frequently appearing characters in ways that support the structuring heterosexual matrix (See Figure 1). One interesting cluster that the graph brings to our attention, for instance, is the set of characters who are positioned as paternal figures to specific nodes.

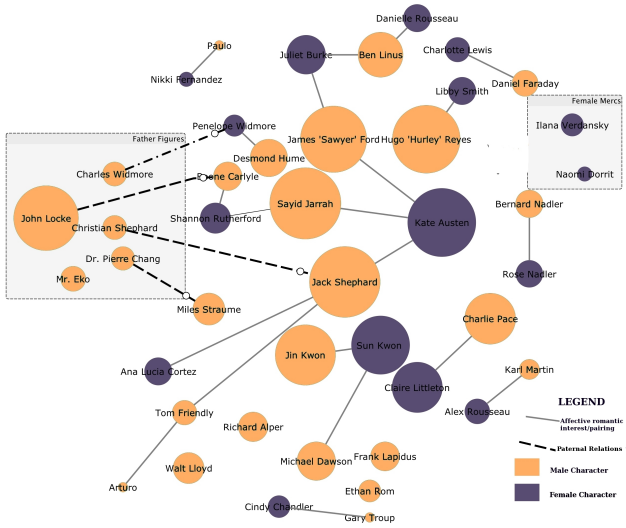


Fig. 1:
Affective and Parental Relations in Lost (Nodal size indicates relative frequency of character's appearance). Constructed using Gephi.

The lines of paternity as a special category that our graphing highlights, and the general heteronormative drift of the show, also invites us to consider the separateness of maternity as a category with a unique valence. Apart from the running theme of pregnancy dramatized by repeated attempts to capture or rescue the visibly pregnant character, maternity we uncover is an organizing type distinct from paternity. While the characters Kate and Jack both co-parent a child, only Kate's status as a mother symbolically revokes her candidacy as protector of the island in the finale.

Along similar lines, social network graphs of *Friday Night Lights* reveal the extent of the characters' racial segregation in the early seasons.[1] We chose this show to analyze because race is a central theme as characters negotiate racial tensions in a small Texas town. In the case of *Friday Night Lights*, segregation is apparent through conventional modes of analysis, but a graph makes it more starkly visible. This graph of a *Friday Night Lights* episode, for example, shows a notable degree of segregation between black and white characters. The mix of such characters on the television screen hides the fact that there are very few interracial conversations, which this graph depicting all interactions in the episode maps clearly. Here we see that the majority of white characters are only connected to other white characters, and, importantly, that the white characters tend to be more strongly socially connected than black characters—they are given more social power (See Figure 2). This graph then provides in one image a sense of how racial interaction plays out across a whole episode.

Fig. 2:
Interactions in Friday Night Lights Season 2 Episode 8. Constructed using ManyEyes.

One way we have used social network graphs to heighten visibility of character segregation (or commonality) is through "deformance," a playful textual reimagining that in some ways resembles remix culture. Deformance is described by Lisa Samuels and Jerome McGann in reference to poetry analysis, as a kind of

"reading backwards," a reconsideration of a text by undoing it, upsetting its order, revealing its gaps (30). In the context of this project, deforming graphs has entailed removing certain networked nodes—and thus removing certain characters from the plot—to see what new connections come to the fore, as well as to reveal the role of those removed nodes in the networked system. If a network falls apart when one node is removed, that often speaks more to the significance of that character than does looking at the graph before he or she is removed from the network. In our analysis of *Friday Night Lights*, for instance, deforming social networks created from fan-generated relationship information (or from our own fannish data generation) revealed that interactions across race hinged on just one or two key characters. Deforming the episode graph above, for example, by removing the central black character Smash and all those characters who interact exclusively with him, revealed the extent to which he functioned as a "bridge" node connecting two otherwise largely distinct communities of characters (See Figure 3).

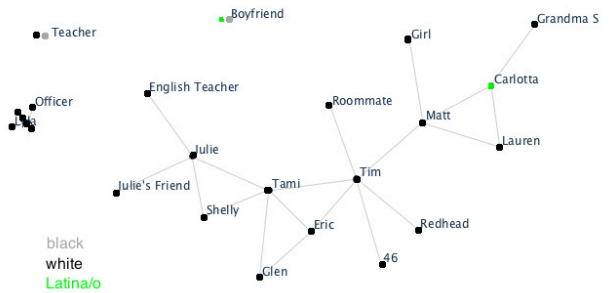


Fig. 3:
Interactions in Friday Night Lights Season 2 Episode 8, with Smash and characters who speak only to him removed. Constructed using ManyEyes.

As this work demonstrates, we are not interested in fan labor solely as sources for data, but also as rich and challenging sites for methodological exchanges. The intellectual production of fans has been acknowledged and celebrated; indeed Henry Jenkins's now-classic text on fan cultures, *Textual Poachers*, highlights the intelligence of audiences and the seriousness of their responses, which engage in familiar practices of literary criticism. Yet to Jenkins, while fans do work that is critical and interpretive, their criticism "is playful, speculative, subjective" and directed to the fan community (284). We see our methodology as a gesture toward the possibility of digital humanities to engage this playful approach seriously, and to consider the importance of the relationships between fans and other interpretive communities.

With academics and fans less strictly monitoring their boundaries in this era of convergence, and with scholars using digital techniques that may formerly have been dismissed as too "playful" or not seriously analytical, a reconsideration of the fan/scholar relationship and possibilities of exchange (or poaching) is called for. Visualization and mapping are widely circulated by fans broaching similar themes to those we explored. In toggling between an episodic mode of analysis and a mode that exceeds the episode, we rely on the community and labor of fans, appropriating their knowledge as the basis of our inquiry. What this suggests is that the digital humanities can anticipate not just a new kind of appreciation of the fan but an acknowledgement of the rich, intellectually productive, and rigorous strategies of knowledge making that new scholarship exploits in its interpenetrative incursions into the terrain of fandom.

References

Jenkins, H. (1992). *Textual Poachers: Television Fans & Participatory Culture*. New York: Routledge.

Jenson, J. (1992). Fandom as Pathology: The Consequences of Characterization. In Lewis, L. A. (ed), *The Adoring Audience: Fan Culture And Popular Media*. London: Routledge, 9-30.

Moretti, F. (2011). Network Theory, Plot Analysis. *Stanford Literary Lab Pamphlet #2*. http://litlab.stanford.edu/LiteraryLabPamphlet2.pdf .

Moody, J. (2001). Race, School Integration, and Friendship Segregation in America. *American Journal of Sociology*, 107(3): 679–716.

Samuels, L., and J. McGann (1999). Deformance and Interpretation, *New Literary History*, 30(1): 25-56.

Notes

1. Our interest in using social network graphs specifically to look at racial interaction was partly inspired by James Moody's 2001 study of integration in high schools, in which he used social network graphs to make visible certain trends of friendship formation between high school students.

Beyond the Scanned Image: A Needs Assessment of Faculty Users of Digital Collections

Green, Harriett Elizabeth
green19@illinois.edu
University of Illinois at Urbana-Champaign, United States of America

Saylor, Nicole
nicole-saylor@uiowa.edu
University of Iowa, United States of America

Courtney, Angela
ancourtn@indiana.edu
Indiana University at Bloomington, United States of America

Introduction

The plethora of digital collections now available to humanities scholars — such as the William Blake Archive or the Civil Rights Digital Library— prompt us to ask critical questions concerning the next stage of humanities scholarship: How well are digital collections meeting the research needs of scholars? And how should digital collections evolve to sustain and strengthen their value to digital humanities research? This paper presents the results of a study of humanities faculty at twelve research institutions that surveyed scholars on their use of digital collections and types of additional functionalities they thought digital collections needed for scholarly research. This study specifically focuses on the structural functionalities in digital collections for which libraries have expertise, such as metadata, information retrieval, and other issues surrounding access and communications.

Background

Digital collections are defined in this study as dynamic and coherent aggregations of thematic digital content that provide a "dense unit for exploration or study" (Palmer 2004; Palmer, et al. 2010). Humanities scholars' use of digital resources has been explored a number of studies during the past several decades. Early studies by Gilmore and Case (1993), Duff and Cherry (2000), and Bates (1993, 1996) among others, analyzed how humanities scholars incorporated early electronic resources into their research. User studies such as those by Brockman et al. (2001), Spiro and Segal (2005), and Warwick et al. (2008) have examined how humanities researchers incorporate digital materials into their workflows and Sukovic (2008, 2011) has analyzed the usage patterns of electronic texts by humanities scholars.

The effectiveness of digital collections for scholarly use has been the focus of several recent studies, including Proffitt and Schaffner (2008), Bulger et al. (2011), and Meyers' study (2011)that led to the development of Toolkit for the Impact of Digitised Scholarly Resources (TIDSR). Thus there is an emerging and rapidly expanding body of research that examines the use and structure of digital collections. Little, however, has been written on functionalities of digital collections for research needs.

Needs Assessment Study

The study presented here consisted of a survey distributed to English and History faculty at twelve research universities, and interviews conducted with fine arts and performing arts faculty at the same research universities. The faculty were identified and recruited for the study with the assistance of librarians and academic technologists at the institutions. The survey was distributed to a randomly selected one-third of the faculty members in the English and History departments at each institution, and was conducted from October 2011 through February 2012. Interviews were conducted from January 2012 through August 2012 via email and telephone, and a random one-third of faculty members from fine arts departments were recruited for interviews. Both the survey and interviews asked respondents to describe their research work with digital collections, the benefits and disadvantages of digital materials, and functionalities that would improve digital collections for scholarly research. Survey respondents were provided with this precise definition of digital collections as curated collections and asked if they used this type of digital resource. Respondents who answered "Yes" continued the survey, while those who answered "No" were taken to the end of the survey.

The quantitative survey responses were analyzed in Excel for statistical percentages. The open-ended survey responses and qualitative interview data were hand coded for themes, and then automated coding was applied with the ATLAS.ti software. In the analyses of these gathered responses from humanities scholars, this paper asks: How do scholars incorporate digital collections into their

research? What do digital collections do well and what functionalities are needed in them?

Two primary needs emerged in the scholars' responses: sustained access and discovery of digital collections, and the ability to mix and reuse digital materials. These needs map to the issues of digital curation and interoperability, which will be explored in-depth in this paper.

Curation

Among survey respondents, the most frequently used materials were texts at 100 percent and images at 94 percent, followed by maps at 58 percent, video at 42 percent, and audio at 39 percent. For all of these materials, curation was paramount.

The responses on the requirements for preparing these materials for scholarly use corresponded to varying processes within the Data Curation Lifecycle (JISC, 2010). Respondents were asked to identify the most needed functionalities for collections of types of digital objects: texts, images, and multi-format media materials. For text collections, detailed metadata and provenance information were the most desirable features. Respondents also strongly expressed the importance of annotating texts, and access to the text files for analyses. For collections of images, the most frequently identified functionality was the ability to download images, followed by the need for consistent, high quality images. Similarly strong responses were expressed for the availability of annotation and editing tools. One survey respondent noted, "The easier objects are to repurpose, remix, and reuse the better."

The interviewed respondents had similar needs for curation, citing content of collections as the most critical need. This included the temporal coverage of content, transcriptions, the inclusion of non-textual sources in collections, and access to broader content. Such steps for curation enable the discovery of content, and reveal to scholars the ways in which they can synthesize the digital materials together for scholarship. Yet synthesis also requires interoperability.

Interoperability

Users also need digital collections that contain interoperable content that functionally facilitates the synthesis conducted by humanities scholars in their comparative examinations of digital materials. As noted in a study by Brockman et al. (2001), the research practices of humanities scholars prominently includes the gathering of sources from multiple collections, in order to create a customized corpus that enables them to explore particular research questions. As such digital collections need effective interoperability between collections' content and metadata to support scholarly research, and the respondents in this study clearly expressed this need.

In the survey and interviews, robust search tools across multiple digital collections were another strongly expressed need among interview and survey respondents. Search functionalities that were particularly valuable included keyword searching, faceted searching, previewing of files, and general browsing of all types of materials. Responses also identified the need for comprehensive metadata in digital collections to enable comparative analysis of collections' content, particularly the identification of specific scholarly editions. The cross-collection use of digital materials results in remixing and reuse of materials for teaching and research, as one respondent explained that ideal digital collections allowed them to be "exporting files and creating my own text and visual files either for teaching or research purpose."

Analysis

The respondents' expressed needs for digital curation and interoperability in digital collections highlight the imperative for libraries to re-evaluate their approach to building and enhancing digital collections. A number of recent studies — notably including the 2010 CLIR report *The Idea of Order: Transforming Research Collections for 21st Century Scholarship* — argue for a dramatic shift in research libraries' conceptualization of collections. This shift is marked by an active, user-centered perspective toward collections — both digital and physical. In particular, the principle of contextual mass (Palmer 2001, 2010), which prioritizes users' research practices in determining collection content, is more imperative than ever in the development of digital collections. Thus as scholarly users demand greater functionality and reliability in digital collections, it is critical that libraries anchor the scope and functionality of their collections in the needs of users, and in doing so, begin establishing deeper research partnerships with users.

Conclusion

This study presents an initial exploration of the needs of researchers when using digital collections. While there are vast differences among the scholarly needs of individual researchers, this study begins to reveal that libraries must work to build collections that exist in a sustainable, networked and iterative environment, and that the content is responsive to the evolving needs for digital humanities research. The study not only reinforces the need for content

providers such as libraries and museums to collaborate with humanities scholars when making curatorial decisions about digital collections, it also emphasizes the essential nature of this partnership in shaping digital collection development.

References

American Council of Learned Societies. (2006). *Our Cultural Commonwealth: The report of the ACLS Commission on Cyberinfrastructure for the Humanities and Social Sciences.* New York: American Council of Learned Societies. http://www.acls.org/uploadedFiles/Publications/Programs/Our_Cultural_Commonwealth.pdf

Brockman, W. S., L. Neumann, C. L. Palmer, and T. J. Tidline (2001). *Scholarly Work in the Humanities and the Evolving Information Environment.* CLIR Publication 104. Washington, D.C.: Digital Library Federation and Council on Library and Information Resources. http://www.clir.org/pubs/reports/pub104/contents.html

Bulger, M., E. T. Meyer, G. de la Flor, M. Terras, S. Wyatt, M. Jirotka, K. Eccles, and C. Madsen (2011). *Reinventing Research?: Information Practices in the Humanities. A Research Information Network Report.* London: Research Information Network.

Council of Library and Information Resources. (2010). *The Idea of Order: Transforming Research Collections for 21st Century Scholarship.* CLIR Publication No. 147. Washington, D.C.: Council of Library and Information Resources. http://www.clir.org/pubs/reports/pub147/reports/pub147/pub147.pdf

Currall, J., M. Moss, and S. Stuart (2004). "What is a Collection?" *Archivaria* 58: 131-146. http://journals.sfu.ca/archivar/index.php/archivaria/article/view/12480/13594

Duff, W. M., and J. M. Cherry (2000). "Use of Historical Documents in a Digital World: Comparisons with Original Materials and Microfiche." *Information Research* 6.1: http://informationr.net/ir/6-1/paper86.html

Gilmore, M. B., and D. O. Case (1992). "Historians, books, computers, and the library." *Library Trends* 40.4: 667-686. https://www.ideals.illinois.edu/bitstream/handle/2142/7802/librarytrendsv40i4g_opt.pdf?sequence=1

Meyer, E. T., Joint Information Systems Committee Report. (2011). *Splashes and Ripples: Synthesizing the Evidence on the Impacts of Digital Resources.* London: JISC. http://ssrn.com/abstract=1846535

Meyer, E. T., K. Eccles, and C. Madsen (2009). Digitisation as e-Research Infrastructure: Access to Materials and Research Capabilities in the Humanities. *5th International Conference on e-Social Science*. held 14 June 2009 in Cologne, Germany. http://www.ncess.ac.uk/resources/content/papers/Meyer(2).pdf (accessed June 14, 2012).

Palmer, C. (2004). Thematic Research Collections. In *A Companion to Digital Humanities.* ed. by Schreibman, S., Siemens, R., and Unsworth, J. Oxford: Blackwell. ch. 24.

Palmer, C., O. Zavalina, and K. Fenlon (2010). 'Beyond Size and Search: Building Contextual Mass in Digital Aggregations for Scholarly Use'. *Proceedings of the American Society for Information Science and Technology* 47.1: 1-10.

Proffitt, M., and J. Schaffner (2008). *The Impact of Digitizing Special Collections on Teaching and Scholarship: Reflections on a Symposium about Digitization and the Humanities.* Dublin, OH: OCLC Programs and Research. http://www.oclc.org/programs/reports/2008-04.pdf

Spiro, L., and J. Segal The Impact of Digital Resources on Humanities Research. http://library.rice.edu/services/dmc/about/projects/the-impact-of-digital-resources-on-humanities-research

Sukovic, S. (2008). Convergent Flows: Humanities Scholars and Their Interactions with Electronic Texts. *Library Quarterly* 78.3: 263-284.

Sukovic, S.. (2008). E-Texts in Research Projects in the Humanities. In Woodsworth, A. and Penniman, W. D., (eds), *Advances in Librarianship*. Bingley, UK: Emerald Group Publishing, 2011.

Warwick, C., M. Terras, P. Huntington, and N. Pappa (2008). If You Build It Will They Come? The LAIRAH Study: Quantifying the Use of Online Resources in the Arts and Humanities through Statistical Analysis of User Log Data. *Literary and Linguistic Computing* 23.1: 85-102.

Linked Data for Music Collections: A User-Centred Approach

Grimes, Jonathan

jonathangrimes@gmail.com
Trinity College Dublin, Ireland

Lawless, Séamus

seamus.lawless@scss.tcd.ie
Trinity College Dublin, Ireland

The emergence of the Linked Data movement offers music collections with the potential to change how they structure and share their data[1] . The ability to use Linked Data to combine music metadata with other information sources offers rich semantic possibilities for both access to and promotion of music, and provides considerable scope

for music collections to remodel their data in a way which makes a significant contribution to the growing web of data[2] .

Previously, collections have focused on building and maintaining relational database models and have used, where possible, common metadata schemas to help build interoperability into their collections. While this has achieved a degree of interchange of structured data between collections, much of this approach has been limited by the type of data structure, the standards used, and above all the development of customised solutions such as APIs for sharing such data[3] . The technologies associated with Linked Data hold the promise of enabling music collections to achieve a high level of semantic interchange with other information resources, and examples of such interchange might include the combining of music metadata with geospatial, historical or biographical data[4] . The advantage of such an approach would allow music archives to focus on their core data collection (i.e. music metadata) and achieve a streamlining of their workflow. The publishing of algorithmic data related to musical form and content as linked data also offers the possibility of developing new applications for music information retrieval[5] , and allows exciting possibilities for music searching and recommendation.

This paper will investigate a number of approaches to the publication of music collections as Linked Open Data. It will explore the motivations for, and benefits of, taking this approach to encoding and exposing musical data. A critical appraisal of the state of the art technology will be outlined and its suitability with relation to music collections discussed. The paper will also explore some of the key changes in maintaining a digital collection and will attempt to address how the notion of a centralised collection has changed as a result of this emerging technology. This impending paradigm shift is one which challenges the notion of data ownership and places organisations and collections in a different position when it comes to managing data.

The paper uses the development of a database system for the Contemporary Music Centre (CMC), a national archive and resource centre for Irish composer's music, to illustrate the design and theoretical processes involved in enabling some of the Linked Data technologies. This database tracks the metadata on Irish composers' compositions and associated materials and is currently being redesigned as part of a new web site which is under development. In addition to replacing the current structure and underlying technologies which power the database, the project also wishes to futureproof the system for the medium-term and take advantage of the emerging trends in digital collection management.

In deciding upon the particular approach to follow to deploy Linked Data in this system, a series of interviews examining the different user groups' attitudes and views on the linking of CMC's content to external data was carried out. The feedback received from these interviews, when combined with the in-depth review of the state of the art, was a key input into developing an approach for the CMC's content. The results of this user-based research also helped to inform the design of a road map of how an approach to Linked Data for a music data source might be taken. The research also illustrates some of the potential problems with the technology and suggests some ways in which these problems might be overcome.

The user research reveals a number of notable findings in relation to attitudes towards the linking of external data to music information. While many identify with the idea of a distributed music collection, users are cautious about exposing and linking music-related data and content to external non-musicrelated content. Key among users is the need for such a Linked Data-driven system to ensure accuracy of data and maintain the trust of its users. There was also a strong view that such external links needed to be curated by both CMC and the users, and that the Linked Data provided should balance with and enhance the core collection data rather than overshadow it.

The solution proposed as a result of this user research involves the restructuring and mapping of CMC's database to Music Ontology classes and concepts and presents a number of approaches involving the presentation of the remodelled data.

Notes

1. Raimond, Yves, Christopher Sutton, and Mark Sandler. (2009). "Interlinking Music-related Data on the Web." IEEE Computer Society.
2. Raimond, Yves, and Mark Sandler. "A Web of Musical Information." Conference Proceedings on Music Information Retrieval.
3. Heath, Tom, and Christian Bizer. Linked Data: Evolving the Web into a Global Data Space. Ed. Frank Van Harmelen & James Hendler. Vol. 1. Morgan & Claypool, 201.
4. Kobilarov, Georgi, Tom Scott, and Yves Raimond.(2009). "Media Meets Semantic Web–how the BBC Uses DBpedia and Linked Data to Make Connections." The Semantic Web.
5. Dixon, Simon et al. (2010) "Towards a Distributed Research Environment for Music Informatics and Computational Musicology." Journal of New Music Research 39.4.

Computing Place: The Case of City Nature

Grossner, Karl
karlg@stanford.edu
Stanford University, United States of America

A large proportion of digital humanities projects entail mapping and therefore computation about places. Much of that computation concerns spatial attributes and relations — location, distribution, interaction flows, and so on — the list of spatial analytic operations of interest for humanistic studies is long and growing. Much of this decidedly quantitative spatial (often spatial-temporal) analysis is performed to better understand *place*, a decidedly humanistic concept defined here as 'experiential space' or 'space as experienced by humans.'[1] This distinction can be confusing because the term space is often appropriated by humanities scholars and critical theorists, in speaking for example of human-constructed spaces. In my biased view as a practitioner of geographic information science (GIScience) with a humanist bent, we need both terms and we need for them not to be conflated.

To date, computational geography (born in the late 1950s but distinguished as GIScience for a couple decades now) has concerned itself with quantitative analyses of spatial and spatial-temporal aspects of natural phenomena at geographic scales. However social scientists, and more obviously, humanist scholars, have questions concerning human experience of space and space-time. In recent years, GIScience has increasingly added methods more directly supporting qualitative analysis (Bodenhamer et al 2010; Dear et al 2011). Examples of this include (i) the semantic analysis of texts joined with spatial analysis of locations associated with their production (Cooper and Gregory 2011), and (ii) the space-time prisms of Time Geography applied to urban residents daily movements by critical human geographers (e.g. Kwan and Schwanen 2009).

In my own recent work on the City Nature project[2] , I have undertaken to characterize "Naturehoods" as areas within cities that are distinctive for their level of "nature-ness;" to capture the human sense of being close to nature in a classification which integrates several physical measures including satellite imagery and social variables including wealth and ethnicity. The overarching goals of City Nature have been first, accounting for the enormous variation in quantity and quality of natural areas in large U.S. cities, and then helping planners arrive at best practices for ensuring reasonable quality of life as cities world-wide grow at an unsustainable pace.

The Naturehoods profile combines satellite measures of mean distance to park-level greenness and non-impervious surface with areal percentages of park and open space, as well as walking distance to parks, and demographic variables like age, household income, diversity, affluence and race.

Somewhat surprisingly, initial results show no appreciable correlation between environmental facts on the ground and social factors. Explanation must now be sought in historical contingency — both in cities' planning processes and external events. Textual analysis of cities' comprehensive plans has been undertaken, and is helping us develop hypotheses to investigate further. Some historical investigation has begun as well, with a study of Los Angeles park planning history. Both of these are outside this paper's scope, but will be discussed in the delivered paper.

Finally, we are also designing a human-subject experiment that will attempt to validate our measures of urban nature on the basis of human reactions to traversing through disparate "Naturehoods." That is, to learn whether this carefully crafted statistical profile of earth surface characteristics joined with demographic statistics in fact corresponds with peoples' "sense of place." Living in an area dominated by strip malls and new tracts is not like living in an area with leafy boulevards, nor a downtown with a few sprinkled mini-parks, nor a wealthy enclave with large yards dense with foliage. Can digital methods predict residents' affect in this case?

Preliminary results of these attempts at "Computing Place" appear in an interactive scholarly work[3] — a mapping and visualization application that allows researchers and the general public to explore will ultimately tell a story of the variation in people's experience of nature in US cities.

References

Bodenhamer, D. J., J. Corrigan, and T. M. Harris (eds.) (2010). *The spatial humanities: GIS and the future of humanities scholarship.* Bloomington, IN: Indiana University Press.

Cooper, D., and I. N. Gregory (2011). Mapping the English Lake District: a literary GIS. *Transactions of the Institute of British Geographers,* 36(1).

Dear, M., J. Ketchum, S. Luria, and D. Richardson (2011) (eds),(2011). *Geohumanities: Art, history, text at the edge of place.* London: Routledge.

Kwan, M. and T. Schwanen (2009) (eds), Critical Quantitative Geographies. *Environment and Planning A,* 41(2).

Massey, D. (2005). *For space.* London: Sage

Tuan, Y. F. (1977). *Space and place.* Minneapolis: University of Minnesota Press: *The perspective of experience*

Notes

1. There are of course many understandings of the term Place; my own are influenced heavily by Massey (2005) and Tuan (1977).
2. undertaken with colleagues Jon Christensen, Elijah Meeks and Maria Santos at Stanford
3. http://citynature.stanford.edu

A Digital Humanities Approach to the Design of Gesture-Driven Interactive Narratives

Harrell, D. Fox
fox.harrell@mit.edu
Massachusetss Institute of Technology, United States of America

Chow, Kenny K. N.
knchow@polyu.edu.hk
The Hong Kong Polytechnic University

Loyer, Erik
erik@song.nu
Song New Creative

Abstract

Technologies allowing for gestural input and output have become more prevalent, e.g. the iPhone/iPad, Nintendo Wii and 3DS, and laptops with multi-touch-screen input and accelerometers to measure motion. However, there is a need to develop theory and technology for incorporating gestural technologies into expressive digital humanities systems. Toward addressing this need, we have developed interdisciplinary theory and technology for expressive gestural interfaces. In particular, we produced a platform for building interactive narrative systems that change emotional tone, theme, perspective, and other content elements in response to embodied user input on multi-touch devices.

We have also produced scholarship that examines the implications and impacts of these emerging technologies[1] .

Overview

The digital humanities include the development of computing technologies to create and better understand subjective expression and narrative, topics usually studied in fields such as literary and cultural studies. Our open-source platform, the GeNIE system, implemented in Objective C, allows authors to create and better understand culturally salient, effective, gesture-driven interactive stories. Three primary outcomes of this project are described below.

Theoretical Framework

Our new models conceptually build upon:

- Theory and Technology for Interactive Narrative (Harrell, 2007)
- Studies of human gesture (Ekman & Friesen, 1972; McNeil, 1992)
- Studies of human-computer gestural input (Wexelblat, 1994)
- Semiotics (Peirce, 1965)
- Study of narrative (Goguen, 2001; Labov, 1972)
- Study of conceptual metaphor (Lakoff & Johnson, 1980; Lakoff & Turner, 1989)

On this basis, we developed our own synthesizing process and framework discussed below.

Results

Our research resulted in:

- *New Theory*: We created a taxonomy of relationships between gestures performed by users as input to devices and narrative meanings in digital stories.
- *New Technology*: We built an engine for implementing gesture-driven interactive narratives for mobile devices featuring touchscreens and a sample interactive narrative to instantiate and assess our outcomes.

New Theory

Our expansive definition of “gesture” encompasses a range of non-verbal communication types including hand gestures, posture, facial expression, and other forms of

embodied meaning expression. When it comes to digital storytelling, there are two meanings of gesture that are likely to be used. These are:

- Input Gesture: Gesture as a user input mechanism on specific device (such as a user clicking and dragging using a touchscreen)
- Storyworld Gesture: Gesture as narrative act/expression *within* a particular media experience (such as a character in a game pointing at another character)

In C.S. Peirce's classic work in semiotics, he describes multiple types of relationships that representations can have to meanings (Peirce, 1965). One of these types of relationships is termed "indexical." It describes a function between the representation (representamen) and a meaning (object). The indexical relationship describes a function between these, i.e. how they (often indirectly) relate to one another. For example, "smoke" can be an index for "fire" when the presence of smoke indicates the presence of fire. This is relevant here, because when gestures performed by a user, such as moving a finger up and down (Input Gesture), causes gestures to be performed by a character, such as nodding her head (Storyworld Gesture), then we can say that the two gestures have an indexical relationship to one another.

So, the task of a gesture-driven interactive narrative system is to *implement a set of indexical relationships suited for effective interactive narrative*.

There are many such indexical relationships. Based on the references in our theoretical framework, some of the most useful in developing interactive narrative works were:

- **Pantomimic**: user action is echoed as an avatar action
  Example: swinging a device to swings a storyworld tennis racket
- **Iconic**: user action depicts the form of an avatar action
  Example: a "<>" motion with fingers makes a character place its hands on hips
- **Metonymic or Metaphoric**: user action is associated with the same meaning as an avatar action
  Examples: *Metonymic* – shaking the device makes a character angry; *Metaphoric* —downward swiping makes a character appear emotionally down (SAD is DOWN)
- **Manipulative**: user action tightly manipulates an object
  Example: Dragging flips a light switch on/off
- **Semaphoric (non-diegetic)**: user action controls something outside of the storyworld
  Example: double-clicking pauses

New Technology

GeNIE consists of two components, one structures narrative events and the other renders them using animated graphical images and text.

Narrative system component:
The narrative event-structuring component allows authors to represent stories based on a model of sociolinguist William Labov (see **Figure 1**). We chose this venerable model as an initial test case since it is empirically-based, easily extensible, and, most importantly, since oral narratives of personal experience are the bases for many more complex forms of storytelling.

Narrative category	Narrative question	Narrative function	Linguistic form
ABSTRACT	What was this about?	Signals that the story is about to begin and draws attention from the listener.	A short summarising statement, provided before the narrative commences.
ORIENTATION	Who or what are involved in the story, and when and where did it take place?	Helps the listener to identify the time, place, persons, activity and situation of the story.	Characterised by past continuous verbs; and Adjuncts (see A3) of time, manner and place.
COMPLICATING ACTION	Then what happened?	The core narrative category providing the "what happened" element of the story.	Temporally ordered narrative clauses with a verb in the simple past or present
RESOLUTION	What finally happened?	Recapitulates the final key event of a story.	Expressed as the last of the narrative clauses that began the Complicating Action.
EVALUATION	So what?	Functions to make the point of the story clear.	Includes: intensifiers; modal verbs; negatives; repetition; evaluative commentary; embedded speech; comparisons with unrealised events.
CODA	How does it all end?	Signals that a story has ended and brings listener back to the point at which s/he entered the narrative.	Often a generalised statement which is 'timeless' in feel.

Figure 1:
Labov's model of narratives of personal experience (Labov, 1972)

Story specifications use the well-known XML format to make it easily usable by non-expert programmers.

Graphical System Component:

The graphical rendering system uses appropriate animated illustrations to express underlying narrative content (**Figure 2** shows a screenshot of a prototype narrative).

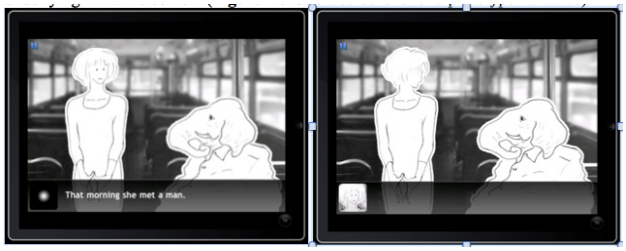


Figure 2:
Screenshots from a test narrative implemented on a mobile "smartphone." Gestural input causes events to occur in a storyworld and drives the narrative forward.

In a sample narrative, we used these to affect important aspects of storytelling such as emotional tone. For example, the metaphoric gesture of pinching in or pinching out can be used to express introverted or outgoing feelings (see **Figure 3**).

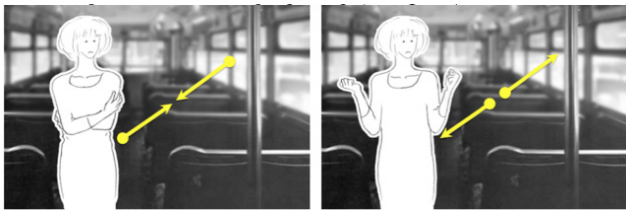


Figure 3:
Pinching in or pinching out affects the character's emotional state.

In our prototype, we implemented emotional states resonating with James A. Russell's idea of core affect (Russell, 2003) by representing both the sense of *arousal* and the degree of *pleasure* characters felt as shown in **Figure 4** (illustration by Chow):

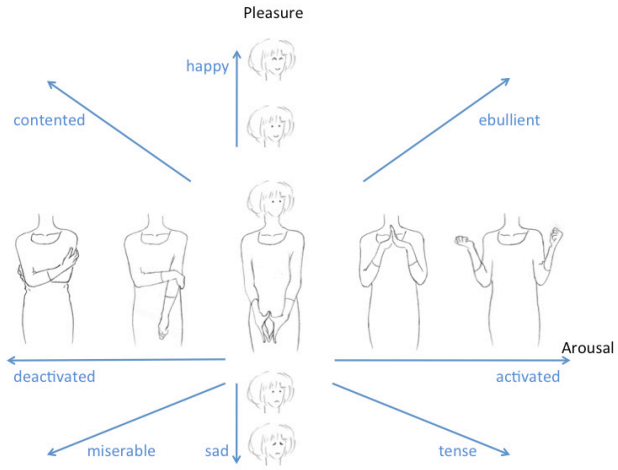


Figure 4:
A character's emotional state is displayed via storyworld gestures. The vertical axis shows how the character's facial expression changes to represent the degree of pleasure. The horizontal axis shows how the character's body language changes to represent the sense of arousal. Combinations of two dimensions result in emotions such as "tense" or "ebullient."

Furthermore, the system utilized conventions from cinema in order to shift between gestures of different types that affect storytelling differently (see **Figures 5** and **6**). For example, a close-up shot allows the user to alter the character's facial expressions and reactions in greater detail, while still keeping an eye on the actions and reactions of the other characters. Dialogue balloons appear to indicate character speech.

Figure 5:
A semaphoric gesture: touching a hotspot for the first time causes an explanation of the related interaction to appear. An iconic gesture: dragging in a ""U shape causes the character to smile (an inverted ""U causes the character to frown).

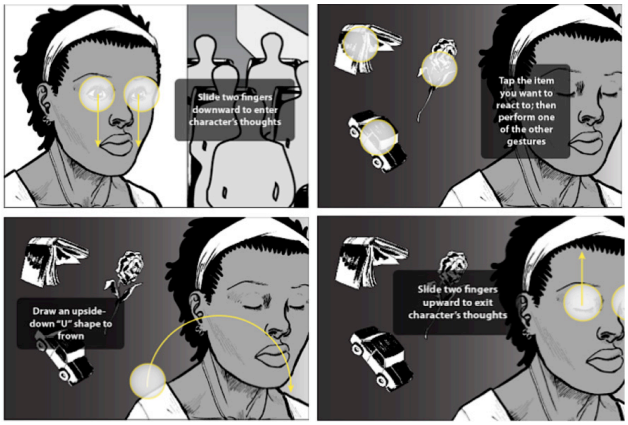


Figure 6:
Manipulative gestures: closing the character's eyes reveals the character's thoughts (other gestures can then change the character's emotions toward objects the character is thinking about); eye-opening exits the thoughts.

Similarly, the medium shot allows the user to get a sense of the environment, and allows for manipulation of the player character's overall posture (**see Figure 7**).

Figure 7:
Manipulative gestures: swiping up and down causes the character to agree with a statement by another character, while swiping left and right causes the character to disagree.

Culturally Specific Expressive Interfaces

We were also interested in implementing culturally diverse gestural models and how culturally-specific gestures can be conveyed via digital media. For example, in some speakerly texts (Gates Jr., 1988), actions such as eye-rolling or placing one's hands on her/his hips have served as markers for a self-possessed "attitude." Gestural walk cycles can convey culturally meaningful differences such as a "cool strut" or "uptight stride." In our sample narrative demo, we implemented a specifically Japanese anime-based gestural model explicitly to exploit the notion of cultural discomfort felt between two characters.

Intuitive Interfaces

At the same time, gestures can also implement relatively universal forms of communication such as the intuitively aggressive act of shaking a device. As another example, tilting a device from side to side can be used as in intuitive way to switch between two characters.

Harrell has developed evaluation methods for games and interactive narratives based on grounded theory analysis augmented by metaphor-based analyses from cognitive science. After assessing early prototypes, we began user-testing the first complete interactive narrative produce using the system called *Mimesis*, which is used to educate users about social discrimination. To summarize early observations, users found the test demo to be effective in conveying our core aims: the system has been found to be intuitive and users conceive of themselves as "puppeteers" for characters. Another outcome is that the system design holds implications that would be a fruitful interaction mechanism for videogames at a diverse user-set. Computer scientists found the gestural input taxonomy to be informative for developing interfaces.

Conclusion

Better understanding and designing digital storytelling systems, as they fit within the broader purview of storytelling in media at large, is a digital humanities endeavor. Our theory helps with better understand digital storytelling systems and enables development of systems that are culturally and technically grounded in a greater diversity of cultural models than most current systems.

References

Ekman, P., and W. V. Friesen (1972). Hand Movements. *Journal of Communication*, 22:353-374.

Gates Jr., H. L. (1988). *The Signifying Monkey: A Theory of African-American Literary Criticism*. New York: Oxford.

Goguen, J. (2001). Notes on Narrative, from http://www.cse.ucsd.edu/~goguen/papers/narr.html (accessed 8/31/2011)

Harrell, D. F. (2007). *Theory and Technology for Computational Narrative: An Approach to Generative and Interactive Narrative with Bases in Algebraic Semiotics and Cognitive Linguistics*. Ph.D. Dissertation, University of California, San Diego, La Jolla.

Labov, W. (1972). The Transformation of experience in narrative syntax. *Language in the Inner Cit*y 354-396. Philadelphia, PA: University of Pennsylvania Press.

Lakoff, G., and M. Johnson (1980). *Metaphors We Live By*. Chicago, IL: University of Chicago Press.

Lakoff, G., and M. Turner (1989). *More Than Cool Reason — A Field Guide to Poetic Metaphor*. Chicago, IL: University of Chicago Press.

McNeil, D. (1992). *Hand and Mind: What Gestures Reveal about Thought*. Chicago, IL: University of Chicago Press.

Peirce, C. S. (1965). *Collected Papers of Charles Saunders Peirce*. Cambridge, MA: The Belknap Press of Harvard University Press

Russell, J. A. (2003). Core Affect and the Psychological Construction of Emotion. *Psychological Review*, 110:1, 145-172.

Wexelblat, A. (1994). *A feature-based approach to continuous-gesture analysis*. Massachusetts Institute of Technology.

Notes

This material is based upon work supported by a National Endowment for the Humanities Digital Humanities Start-Up Grant.

The Advanced Identity Representation (AIR) Project: A Digital Humanities Approach to Social Identity Pedagogy

Harrell, D. Fox
fox.harrell@mit.edu
Massachusetss Institute of Technology, United States of America

Abstract

The Advanced Identity Representation (AIR) Project[1] develops theory and technology for expressive and empowering self-representations (profiles, player characters, avatars) in social networks, games, and related technologies. Here, I report on the development processes and pedagogy of the AIR Project. In particular, I describe two systems implemented as a part of the AIR Project: (1) *Mimesis*, a critical pedagogy game that increases awareness of covert forms of social discrimination, and (2) *Chimeria*, a system that computationally models social group membership phenomena and narrates them via a social networking interface. These systems emerged from integrated research and pedagogical aims based upon a unique interdisciplinary theoretical framework and have been published on in peer-reviewed conference proceedings with students as co-authors.

Introduction

Over the past several years, I have run a course at MIT on Advanced Identity Representation in which students use software platforms developed in my lab[2] to implement empowering computational systems allowing users to better understand social identity phenomena. In particular, we have explored topics such as microaggression (covert discrimination studied in critical race theory and clinical psychology) and modeling social group membership and naturalization.

Most computational identity systems incorrectly *reify* identity categories by implementing them as simple data fields (e.g., selecting gender from a brief drop down menu) or collections of attributes (e.g., races represented as modifiers to numerical statistics and constrained graphical characteristics in computer games). In contrast, the AIR Project results in computational models of *subjective* identity phenomena related to categorization such as specific forms of marginalization that are often overlooked in engineering. Simply put, we intervene in engineering practice by replacing conventional narrow categorization models with systems informed by more critically-aware humanities/social science research.

The AIR Project approach allows us to construct systems, for example, simulating phenomena such as systematic patterns of discrimination or experiences of moving between social groups. These systems cannot express the nuances of real world identities, yet they provide *advances*[3] *over current systems* in their foci on phenomena shown by humanities scholars and social scientists to be important for understanding issues such as oppression and supporting user empowerment. The resultant systems are often necessarily reductive (from real life experience to data structures and algorithms) in order to be implementable, yet this reduction is done knowingly with the benefit of expanding the expressive capacity of computational systems to address social identity phenomena.

Theoretical Framework

Our systems are based on a theoretical framework including the following areas:

Digital Humanities/Game Studies

Current user representations in digital media are inadequate for capturing complex phenomena involving subjective experience of social identity. Current character creation tools allow for user representations to be customized on the basis of attributes associated with models of race, class, profession, and similar classifications, along with physical choice and construction of character models, skin tone, gender characteristics, and the like. Many popular current games duplicate and amplify many disempowering existing social structures. Such games hardcode stereotypes into their infrastructures. They reduce social constructs such as race to sets of numerical variables, abstract data structures and cosmetic changes to avatar appearance (Harrell, 2010).

Cognitive Science of Categorization

Results from cognitive science have revealed that there are many basic, entrenched metaphors that inform everyday cognitive categorization, including social stereotyping. (Lakoff, 1987) These concepts are often structured by image schemas, "skeletal patterns" that recur in our motor-sensory experiences such as "Center-Periphery" or "More is Up" as expressed respectively by everyday cognitive metaphors such as "marginalized peoples" or "upper-class." (Lakoff, 1987)

Sociology of Classification

A great deal of personal suffering has been identified in cases where individuals exist at the interstices and boundaries between social classifications, for example individuals of ambiguous racial classification in Apartheid South Africa (Bowker & Star, 1999), or the classification of "mixed-race" individuals in the United States Census recordings. Such situations, in which there are conflicts between individual biographies, identity self-perception, and social metrification of identity, result in the experience of what Bowker and Star term *torque*. AIR project systems take up the challenge to computer scientists posed by Bowker and Star by implementing models of social identity that are dynamic, imaginative, and explicitly socially-constructed.

Pedagogy and Development

Building upon the theoretical framework above, my repeatable lab/seminar course CMS.628/828 Advanced Identity Representation is populated by MIT Ph.D. students in Electrical Engineering and Computer Science, S.M. students in Comparative Media Studies, a small number of undergraduates from a range of departments, and occasional cross-registered students such as from the Graduate School of Education at Harvard University. My teaching approach draws upon critical pedagogy research and theories of computational literacy (diSessa, 2000; Freire, 1973; Street, 1993). A brief account of two projects developed in the course follows.

Mimesis

Mimesis is an interactive narrative system exploring a particular social discrimination phenomenon via metaphor. As illustrated in **Figure 1**, in *Mimesis* the player character encounters others (sea creatures) who perform discriminatory microaggressions. We use metaphor to convey a generalized notion of microaggression from the definition of racial microaggression as "brief and commonplace daily verbal, behavioral and environmental indignities, whether intentional or unintentional, that communicate hostile, derogatory, or negative racials slights and insults to the target person or group." (Sue, et al., 2007) Racial microaggressions have been clinically found to have strong cumulative effects on health and happiness, and restrict understandings between groups. (ibid.) *Mimesis* aims to explore the efficacy of computational identity representation systems as tools for bringing awareness of microaggression at large. In Mimesis, each encounter between the player and a non-player character progresses according to a conversational narrative schema in which moods such *oblivious*, *confused*, *suspicious*, or *aggressive* are mapped to strategies of conversationally responding to microaggressions. We aim for *Mimesis* to be an effective tool for increasing awareness of this subtle form of social discrimination.

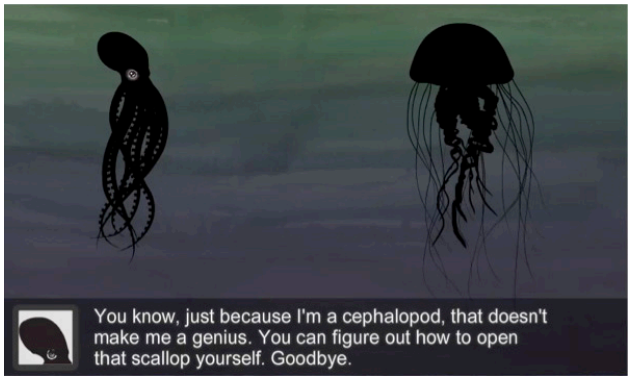


Figure 1:
This screen shot shows the player's character (left) in a microaggressive encounter in which a non-player character (right) metaphorically represents the theme of stereotypical ascription of skill (e.g., assuming someone is good at mathematics because of her ethnicity).

Chimeria

Chimeria consists of: (1) the *Chimeria Engine*: a dynamic algorithmic model of users' degrees of membership in multiple social groups, and (2) the *Chimeria Social Narrative Interface (ChimeriaSN)*: a narrative social networking interface for expressing experiences of membership and marginalization in social groups (see **Figure 2**). The *Chimeria Engine's* model of users' dynamic gradient category memberships in relation to central members enables more nuance than binary statuses of

member/nonmember. (Bowker & Star, 1999; Harrell, 2010; Lakoff, 1987)

ChimeriaSN is a streamlined, aestheticized social networking interface. The screen is dominated by a photowall: a dynamic collage of photos representing the user's musical taste preferences. A feed of recent updates, posts, and invitations appear in an adjacent vertical timeline. One initial application of *Chimeria* attends to the phenomenon of *passing—presenting* oneself as a member of another group to gain social acceptance (e.g., a multiracial person passing for white or a rock fan passing as a jazz fan). Music is our initial test domain since people often identify with groups based on musical preferences. We generate categories of social groups using music data (e.g., genres, artists, moods) from the Rovi Cloud Services API.

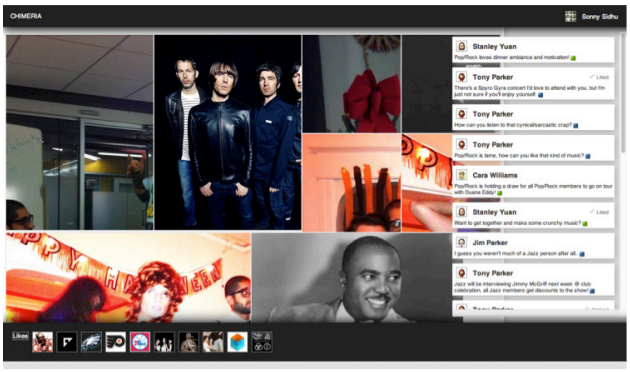


Figure 2:
This screenshot shows the Chimeria social-networking/ interactive narrative interface.

ChimeriaSN begins by initializing a set of musical preferences from the user's Facebook music likes or by manual entry. A *Chimeria* profile is then generated, initiating a hybrid real/fictitious simulation presented via narrative structured posts by the user's friends. The user may click "like," "dislike," or simply ignore these posts, resulting in group membership changes illustrated by alterations to the photowall and subsequent posts. The resulting narrative may tell of passing as a member of a new group, reinforcing a prior group affiliation, or even being marginalized in every group. Furthermore, some groups are deemed oppositional, privileged, or marginalized relative to others.

Novel aspects of *Chimeria* include:

- a compelling interface that is familiar and accessible to most users who have used a social networking application,
- the combining of real world data with authored data, comprising a new type of *alternate reality narrative* (blending fact and fiction to convey social themes), and
- generalizability to a wide range of applications.

Conclusion

The AIR project looks at the underlying data structures and algorithms and how they implement cultural identity effects, and posits a technical framework for more deeply engaging identity semantics of classification and categorization. Additionally, a primary application of our project is educational software such as *Mimesis* and *Chimeria* that utilize digital humanities technology and theory to allow students to represent themselves as learners and doers within subject domains. Digital humanities approaches to identity such as this, we feel, are especially attuned to the pedagogical needs of today's students.

References

Bowker, G. C., and S. L. Star (1999). *Sorting Things Out: Classification and Its Consequences*. Cambridge, MA: MIT Press.

diSessa, A. A. (2000). *Changing minds: Computers, learning, and literacy*. Cambridge, MA: The MIT Press.

Freire, P. (1973). *Pedagogy of the oppressed*. New York: Seabury Press (Originally 1968).

Harrell, D. F. (2010). Toward a Theory of Critical Computing: The Case of Social Identity Representation in Digital Media Applications. *CTheory*.

Lakoff, G. (1987). *Women, Fire, and Dangerous Things: What Categories Reveal about the Mind*. Chicago, IL: University of Chicago Press.

Street, B. V. (1993). Introduction: The new literacy studies. In B. V. Street (Ed.), *Cambridge studies in oral and literate culture: Cross-cultural approaches to literacy*. Cambridge, UK: Press Syndicate of the University of Cambridge.

Sue, D. W., C. M. Capodilupo, G. C. Torino, J. M. Bucceri, A. M. B. Holder, K. L. Nadal, et al. (2007). Racial Microaggressions in Everyday Life: Implications for Clinical Practice. *American Psychologist*, 62(4), 271-286.

Notes

1. This material is based upon work supported by the National Science Foundation under Grant No. 1064495.
2. My research group is called the Imagination, Computation, and Expression Laboratory (ICE Lab).
3. The meaning of "advanced" in the AIR acronym refers to advances over the limitations of many current identity

categorization systems that are unengaged with theories of identity from the humanities and social sciences.

TXM Platform for analysis of TEI encoded textual sources

Heiden, Serge
slh@ens-lyon.fr
ICAR Laboratory, ENS de Lyon - CNRS, France

Lavrentiev, Alexei
alexei.lavrentev@ens-lyon.fr
ICAR Laboratory, ENS de Lyon - CNRS, France

Textometry is a methodology of text corpora analysis combining qualitative and quantitative techniques (kwic concordances, word frequency lists, collocations, factorial analysis, etc.) and producing valuable results for various fields of the humanities (linguistics, literary studies, history, geography, etc.).

The first generation of textometric software operated mainly on "raw text" with limited metadata and structural markup. In the recent years, a great number of digital resources with complex markup have been created. These can include multiple languages, various readings and other forms of critical apparatus, annotations like word lemmas or part of speech, syntactic structures, etc. As a general markup environment, the TEI guidelines provide a common framework for encoding all kinds of textual resources, although this framework allows a great flexibility and sometimes the same information can be encoded in many different ways. It is a challenge for the software and for the researcher to interpret these data correctly out of the context of their original project but it is also an opportunity to make the textometric analysis deeper and more precise.

A new generation of textometric open-source software called TXM was initiated by the Textométrie research project[1] funded by the French ANR agency (2007-2010) bringing together previous textometric techniques and state-of-the-art text encoding and corpus-building technologies: Unicode, XML, TEI, NLP (Heiden, 2010; Heiden et al., 2010). The TXM platform can be downloaded for free at http://sf.net/projects/txm with its sources. This article presents the design and the current state of the import environment being developed since to allow the platform to analyze various kind of TEI encoded sources.

The TXM platform addresses the challenge of importing TEI encoded corpora by "translating" the source document structure into the terms (or objects) relevant for textometric analysis. The main objects are: "text units" (define text limits in a corpus), "text metadata" (associate texts with their properties), "lexical units" (the way the word forms are separated), "word properties" (how to get their lemma or morpho-syntactic description if available), "text divisions" (book parts, sections, paragraphs...), primary and secondary "text surface" (what is the main language of the text to run NLP tools on the right tokens and possible secondary languages: foreign quotations or section titles provided by the editor of a historical source text), "out-of-text": parts not to be considered as part of the source text (critical apparatus, encoding comments, etc.), "pagination" (to build an edition of the texts), etc.

For each type of source corpus, one has to precisely define how the textometric objects are encoded in the TEI sources and how to extract them to express the corresponding objects in a specially designed XML-TXM pivotal format before being instantiated inside the platform. The XML-TXM format is specialized in analytic data categories, in a way similar to the "TEI Analytics" format of the MONK project (Brian L. Pytlik Zillig, 2009), but is richer in data categories and is a formal TEI extension.

The extraction process is implemented by a combination of specific XSLT stylesheets, XPATH expressions and Groovy script parameters[2] .

We will describe how that approach has been validated on a comprehensive set of completely unrelated TEI encoded corpora: "Bibliothèques Virtuelles Humanistes" corpus (BVH collection of 16th century books: http://www.bvh.univ-tours.fr), Flaubert's "Bouvard et Pécuchet" 19th century novel corpus: http://dossiers-flaubert.ish-lyon.cnrs.fr, corpus of 5 years issues of the "DISCOURS" linguistic journal: http://discours.revues.org/?lang=en) and the TEI version of the Brown 1 million words corpus from the NLTK project: http://nltk.org .

TXM TEI import environment and its XML-TXM pivotal format have proven to be flexible enough to process various data sources efficiently. In further developments, we will define a complete ODD description of the XML-TXM format to document it better for the TEI community and to contribute to the discussion on the ability of software tools to analyze TEI encoded data.

References

Heiden, S. (2010). The TXM Platform: Building Open-Source Textual Analysis Software Compatible with the TEI Encoding Scheme. *24th Pacific Asia Conference on Language, Information and Computation.* Éd. Kiyoshi Ishikawa Ryo Otoguro. Institute for Digital Enhancement of

Cognitive Development, Waseda University, 4-7 November 2010. 389-398.

Heiden, S., J.-P. Magué, and B. Pincemin (2010). TXM : Une plateforme logicielle open-source pour la textométrie – conception et développement, in Bolasco, S., et al. (eds.), *Statistical Analysis of Textual Data - Proceedings of 10th International Conference* JADT.

Pytlik Zillig, B. L. (2009) TEI Analytics: converting documents into a TEI format for cross-collection text analysis. *Literary and Linguistic Computing* 24(2):187-192; doi:10.1093/llc/fqp005.

Notes

1. http://textometrie.ens-lyon.fr/?lang=en
2. The TXM import environment is implemented by several dynamic scripts written in the Groovy progamming language. All that software environment is directly accessible to the user to be modified and adapted: platform sources, import scripts, XSLT stylesheets, etc.

Digitizing Serialized Fiction

Hess, Kirk
kirkhess@illinois.edu
University of Illinois at Urbana-Champaign, United States of America

Introduction

One barrier to locating serialized fiction in a digital newspaper archive is the fact that the serialized fiction themselves are not indexed, and individual articles do not have subject terms or tags associated with them that would identify them as fiction. As a result, articles are difficult to find unless the reader browses a large volume of issues or simply hits upon a salutary keyword search. Keyword searching of the collection is more effective for articles on topics in farming or farm life than for works of fiction. Unless the reader is looking for stories by a specific author or for a known story title, keyword searching of fiction is highly ineffective.

While the software used by most archives does a good job of connecting articles in a single issue, the reader does not know where to find the next installment in a serialized work of fiction, so he or she has to find it manually by browsing the collection or doing a keyword search. Finally, while the OCR scanning was done to the highest standard, this is an imperfect process, and much of the content cannot be adequately OCR'd due to background noise and broken letters, features of the original newspapers that impede scanning.

This paper summarized the process of our project that was completed over 15 weeks in the summer of 2012. Our goal was to complete the manual indexing process that had already been started previously, display serialized fiction articles in a new repository, evaluate multiple software packages to see which ones were the most promising for use in the future, and evaluate any automated ways of finding serialized fiction.

Method and Results

Manual Indexing

We experimented with three digital library systems, a Drupal/Fedora based repository (Islandora) (http://islandora.ca/) , converting the fiction into TEI P5 and displaying it in the California Digital Library's eXtensible Text Framework (XTF) (http://www.cdlib.org/services/publishing/tools/xtf/) and Omeka (http://omeka.org) a PHP-based publishing platform for digital library objects. We were unable to get Islandora's OCR correction module installed so we stopped using it in favor of Omeka. We used XSLT to transform the PrXML into very simple TEI5 files, which we were able to upload to XTF, but the lack of an editor and the intensely manual process of text encoding was also rejected in favor of Omeka. TEI Example: http://uller.grainger.uiuc.edu:8080/xtf/search

Ultimately, we decided on using Omeka with the Scripto Plugin for correction. Serialized fiction articles in one title, the Farmer's Wife, was manually indexed in a spreadsheet, and graduate assistants converted those stories from PrXML into an exhibit, added Dublin Core metadata and links to the newspaper archive from the new serialized fiction collection. The end result was index of serialized fiction that would increase the accessibility of these articles. Omeka Exhibit: http://uller.grainger.illinois.edu/omeka/

Crowdsourcing OCR correction

The University of Illinois Digital Newspaper collections are in Olive Software ActivePaper Archive, which has a method for administrators to correct text but not users. Omeka provides a plugin called 'Scripto' for text correction that we were able to successfully use to correct the text in selected articles. We also evaluated Veridian (http://www.dlconsulting.com/veridian/) , which is a commercial digital newspaper library solution used by Trove Digitised

Newspapers (National Library of Australia), From the Page (http://beta.fromthepage.com/) and Islandora (http://www.islandlives.ca/) . From the Page and Islandora were both very difficult to install and administer, and while not free we felt Veridian was a much better approach and we are evaluating it as part of our future newspaper digitization efforts.

Text Analysis

How can we identify serialized fiction without having to have a human find it, index it in a spreadsheet and manually extract it from the archive? Certain n-grams are common within serialized fiction such as 'chapter', 'the end', 'to be continued' and could be used to simply search for keywords within documents; we could also calculate which words occur most frequently in fiction vs. other types of articles and use those terms to automatically tag articles.

We also evaluated using topic analysis to find fiction. We evaluated the 580 articles we had already identified as serialized fiction using Mallet to find 25 topics with 25 words each. Figure 1. shows the top 25 topics modeled as a network using Gephi, while figure 2 shows the topic words ordered by frequency.

Figure 1

Figure 2

Nodes were ranked by betweenness centrality and topic 14 had the highest at 51,321.01 and its component n-grams along with the other top topics could be used to find serialized fiction in other titles.

One final text analysis technique that could be useful is identifying proper names is Named Entity Extraction. While we made an effort to manually remove names from the topic analysis, as you can see they kept reappearing in the results. By using named entity extraction we could eliminate proper names from the topic analysis to make them more accurate, and to link fiction together by the character's names. All three of these techniques (keyword frequency, topic analysis, named entity extraction) I plan on evaluating in a future study.

Conclusion

Serialized fiction is an important component of historical newspapers and by making it more accessible to patrons and researchers we can expand the use and usefulness of our digital newspaper collections. The manual indexing approach was relatively inexpensive to accomplish but was time consuming and difficult to do over a large corpus of pages. Two promising approaches to find and digitize serialized fiction in our newspaper archive are adding a crowdsourcing feature to enable users to identify article types and correct mistakes, and utilizing text analysis techniques to identify fiction programmatically. We hope to report on our efforts at the latter at the DH 2013 conference.

References

Bastian M., S. Heymann, and M. Jacomy (2009). *Gephi: an open source software for exploring and manipulating networks.* International AAAI Conference on Weblogs and Social Media.

Brandes, U. (2001). A Faster Algorithm for Betweenness Centrality. *Journal of Mathematical Sociology* 25 163-177.

Cohen, D. (2008). *Introducing Omeka.* Dan Cohen's Digital Humanities Blog. http://www.dancohen.org/2008/02/20/introducing-omeka/ .

Islandora. http://islandora.ca/ .

Island Lives. http://www.islandlives.ca/

McCallum, A. K. (2002). MALLET: A Machine Learning for Language Toolkit. http://www.cs.umass.edu/~mccallum/mallet .

Veridian. http://www.dlconsulting.com/veridian/ .

Encoding historical dates correctly: is it practical, and is it worth it?

Holmes, Martin
mholmes@uvic.ca
Humanities Computing & Media Centre, University of Victoria, Canada

Jenstad, Janelle
jenstad@uvic.ca
Department of English, University of Victoria, Canada

Butt, Cameron
cambutt@uvic.ca
Department of English, University of Victoria, Canada

From the middle ages to the early 20th century, a bewildering variety of calendars and dating methods was in use across Europe. This presentation will address issues involved in encoding historical dates during the early modern period, and look at strategies for enhancing computability and interoperability in date-encoding.

Although other calendars are in use across the world, nearly all societies have adopted the Gregorian calendar to some degree; wherever you go in the modern world, the date and time are generally uncontroversial. In past centuries, this was far from true. Jardine (2009) cites the case of William of Orange, whose invasion fleet left the Netherlands on November 11, 1688, but landed in England on November 5, having negotiated not only the English Channel but also the ten-day discrepancy between the Gregorian calendar in use in the Netherlands and the Julian still used in England. The complicated history of the adoption of the new Gregorian calendar across Europe, between its introduction by Pope Gregory in 1582 and its final consolidation in the early part of the 20th century, is familiar to scholars (see for instance Cheney 2000 and Duncan 1998).

Our project, the *Map of Early Modern London* (*MoEML*), falls squarely within the period of maximum calendar confusion for England. The two principle sources of difficulty are the discrepancy between the Gregorian and Julian calendars (ten days during most of the period) and the change in New Year's Day, which from the 12th century onward was March 25 rather than January 1. The resulting ambiguity in actual dating is compounded by the ways in which writers, both in the period and after it, have chosen to deal with the difficulty. The terms "Old Style" and "New Style" (O.S. and N.S.) have been variously used to indicate the start-of-year convention in use, the leap-day adjustment between Julian and Gregorian calendars, or both; so if we see, for instance, "February 11 1650 New Style", it is by no means clear whether the date referred to would have been viewed by contemporaries as February 1 1650 (calendar adjustment), February 11 1649 (start-of-year adjustment), or February 1 1649 (both).

These issues of dating have been a major challenge for *MoEML*'s encoders. Our collection of TEI born-digital documents and historical texts includes dates from many calendars. When we encode dates in historical texts, we first must determine which dating method was used by the clerk or author. Dates are given in terms of many possible systems: regnal years, papal years, mayoral years, legal terms, years calculated from a particular feast day, and even Anno Mundi figures (years since the creation of the world); we follow Cheney (2000) and Fryde (1986) in parsing these references and converting them to Julian dates. We give a few examples.

> Julian date from primary source: "Alfred king of the weſt Saxons, in the yere 886..." (Stow 1598 8)
>
> regnal date: “a par[l]iament being holden at Carlile in the [...] 35. of Edwarde the firſt" (Stow 1598 11)
>
> Anno Mundi dates, combined with proleptic Julian: "[...] Eneas, the ſonne of Venus, daughter of Iupiter, aboute the yeare of the world 2855. the yeare before Chriſtes natiuitie, 1108. builded a Citie [...]" (Stow 1598 1)

We also use a wide variety of modern sources such as the *Oxford Dictionary of National Biography*. The ODNB has precise methods of expressing uncertainties in dates and date ranges, but does not specify how uncertainties arise, which means we can determine precision from the *ODNB* but not accuracy.

We have attempted to discover whether other projects are concerned with calendar issues, and if so, whether they have adopted similar encoding methods. A brief survey of the projects listed on the TEI Projects page (http://www.tei-c.org/Activities/Projects/) shows, at the time of writing, 152 projects in total. Of those, 68 projects could be expected to contain materials in the historical range that concerns us. We were able to retrieve XML from 19, 16 of which contained encoded dates that would be subject to calendar issues. Only three of those projects appear to have taken account of calendars in their encoding. The prevalent view is well expressed by Godfried Croenen (personal communication, 2012-10-05):

> "All medievalists use Julian dates to refer to any date before the introduction of the Gregorian calendar, and so all the dates before the 16th century I have ever encoded into XML TEI documents are in Julian dates. I never felt it would be useful to convert these dates to Gregorian dates, as nobody would know what I was referring to."

Croenen also expresses doubt as to whether it is practical or useful to attempt date conversion between calendars. Others have pointed out that, where date encoding involves only the year, there is no reason to worry about the calendar, and one might as well encode using @when (whose datatype is explicitly Gregorian) with a Julian date. However, in the case of England between the 12th century and 1752, assuming Julian years amounts to an allowance that nearly one in four dates is likely to be wrong, because of the New Year issue. Another objection to regularizing all date-encoding to use Gregorian is that it is unconventional to use the Gregorian calendar proleptically. However, it is a long-standing practice to use the Julian calendar proleptically, referring to dates in antiquity — in fact, Stow does this in one of the examples above, in which he glosses the Anno Mundi date 2855 as 1108 BC.

In our encoding of dates for the *MoEML* project, we have two major concerns: that dates be as accurate as possible so that we know when an event occurred (or at least that the source and scope of inaccuracy be clearly expressed), and that they be computable. We are constructing an eventography, and we want to be able to plot event sequences on timelines. We would also like to be able to integrate our data with that of other early modern projects, many of which will have data from countries whose calendar usage varies substantially from English practice. As a result of these concerns, we are early adopters of some recently-added features of TEI that are intended to formalize accurate encoding of dates from differing calendars.

The original P5 attributes for encoding dates included two distinct classes: `att.datable.w3c`, and `att.datable.iso`. These two classes allow slightly different forms of date encoding (derived from XML Schema datatypes, and ISO 8601 respectively), but both are explicitly based on the Gregorian calendar. In other words, it is clearly wrong to encode a Julian date using one of these attributes:

```
*<birth when="1566"
calendar="#julianEngland">1566</birth>
```

The recently-added `att.datable.custom class` remedies this deficiency by providing a full suite of dating attributes designed for non-Gregorian calendars, along with the @datingMethod attribute through which the calendar used can be specified (@calendar refers to the calendar used in the text content of a dating element, not its attributes). We can now encode a date with these attributes and Julian dates :

```
<birth when-custom="1566"
datingMethod="#julianEngland"
calendar="#julianEngland">1566</birth>
```

These tags show that both the attribute value and the text date use the Julian calendar. Given that our purpose is computability, though, the question arises: why not simply convert all our dates to (proleptic) Gregorian before encoding them? If we take the example above, this conversion would be the result:

```
<birth notBefore="1566-04-04"
notAfter="1567-04-03"
calendar="#julianEngland">1566</birth>
```

The conversion, in accounting for the New Year issue and the leap day discrepancy, becomes a rather unwieldy range. Moreover, this conversion is itself computable, so it is unnecessary to impose this burden on our encoders. Instead, we encode Julian dates using `@when-custom`. On the website, we generate tooltips for all such dates showing the equivalent date or date-range in Gregorian. For the purposes of interoperability, the same conversion could be used to insert Gregorian dating attributes.

Until now, the encoding of historical dates in TEI projects appears to have been haphazard, for a variety of reasons, including the lack of adequate encoding mechanisms, academic convention, and historical practice. However, we now have a set of attributes that enable us to be more precise, and we can easily create conversion functions between (for instance) Julian and Gregorian dating systems. Moreover, as we begin to integrate data from different projects, and create timelines and event sequences that require accurate dating, there is more reason than ever for developing and propagating good practice in date encoding; we do not want to end up creating inter-project timelines in which (for example) the invasion force of William of Orange arrives in England several days before it sets off from the Netherlands. In presenting our date-encoding practices and the issues we have encountered, we hope to stimulate a discussion on accurate date encoding that will encourage those working on projects involving non-Gregorian calendars to be aware of the issues, and to collaborate in creating methods for encoding and interchange that will obviate these problems.

References

Cheney, C. R. (2000). *A Handbook of Dates for Students of British history*. Revised by Michael Jones. Cambridge: Cambridge University Press.

Duncan, D. E. (1998). *Calendar: Humanity's Epic Struggle to Determine a True and Accurate Year*. New York: Avon Books.

Duncan, D. E. (1999). Calendar. *Smithsonian* **29 (11)**. 48-58.

Fryde, E. B., D. E. Greenway, S. Porter, and I. Roy (1986). *Handbook of British Chronology*. 3rd edn. London: Offices of the Royal Historical Society.

Jardine, L. (2009). *Another point of view*. London: Preface.

Stow, J. (1598). *A SVRVAY OF LONDON.* London: John Windet for John Wolfe.

Practical Interoperability: The Map of Early Modern London and the Internet Shakespeare Editions

Holmes, Martin

mholmes@uvic.ca
Humanities Computing & Media Centre, University of Victoria, Canada

Jenstad, Janelle

jenstad@uvic.ca
Department of English, University of Victoria, Canada

While the promise of interoperability has been one of the major driving forces in the adoption of standards such as TEI, it has long been recognized that interoperability has only limited practicality (McDonough 2008; Sperberg-McQueen 2008). As large-scale digitization projects have matured, it has become apparent that the most effective approach to interoperability between them is based on loose coupling through APIs and metadata exchange services such as OAI-PMH, rather than wholesale convertibility or aggregation (see, for example, Bol, Hsiang and Fong 2012; Matei 2012).

The *Map of Early Modern London (MoEML)* and the *Internet Shakespeare Editions (ISE)* are mature projects, and both are under active development. While *MoEML*'s text database is steadily growing, the literary texts in the ISE collection, on the same network and sharing some of the same research team, have become a tempting target for integration. MoEML's goal is to give users a sense of the lived space of London, particularly as that space was invoked by the implied geography of early modern plays. Shakespeare's ten history plays are rich in references to London places. *1 Henry IV* moves between Eastcheap and Westminster; the title character of *Richard III* bustles through London; and the Tower looms ominously over the action of nearly every play. Ingesting and mapping these references in the *MoEML* environment would stimulate research questions about Shakespeare and London alike. How typical is Shakespeare's invocation of London? How do his characters move through the urban environment? What is the relationship between London and the court in Shakespeare's historical vision? How does this vision compare to that of other playwrights, such as Thomas Heywood, and to that of historians like Holinshed and Stow?

MoEML maps the streets, sites, and significant boundaries of London from 1560 to 1640, basing its interface on the Agas Map of London, which dates originally from the 1560s. The project incorporates a detailed gazetteer, topical essays, and digital texts from the period, and will soon include three editions of John Stow's *A Survey of London*. At the heart of the project is an XML placeography incorporating over 720 streets, churches, wards, neighbourhoods, and sites of interest. Places are both geo-referenced and linked to the Agas Map.

One goal is to use the Agas Map as a platform on which to visualize the locations in texts of the period. To that end, *MoEML* includes a library of early modern texts with all the toponyms identified and tagged. With dramatic texts, we have until now included only the "Dramatic Extracts" that contain London toponyms. It would be preferable, though, to extract toponyms dynamically from existing digital editions and plot them on the Agas map. The simple data visualization in Figure 1 shows a prototype for *Richard III*, with each location sized according to the number of references to it, demonstrating the manner in which the Tower dominates the action.

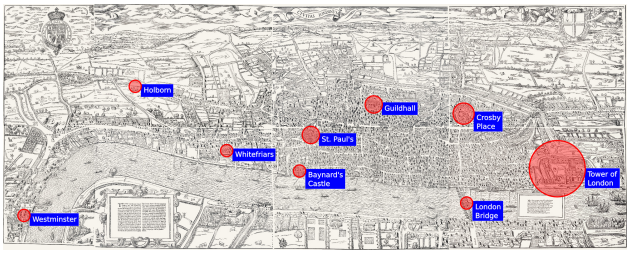


Figure 1:
The London locations in Richard III on the Agas Map, sized according to the number of references to them.

The *Internet Shakespeare Editions* is primarily an open-source digital anthology of Shakespeare's plays. The *ISE*'s programming platform also runs the *Queen's Men's Editions*

(QME) and the *Digital Renaissance Editions (DRE)*. Between these three projects, all the plays of Shakespeare and his contemporaries from 1500 to 1640 will be available in standardized XML base texts. At the heart of the projects are carefully edited texts of each play, in both their early printed forms and in modern editions with spelling and punctuation regularized. *ISE* editors already tag the base texts with simple tags that are converted to XML. We could ask the *ISE*, *DRE*, and *QME* editors to add in the London references; however, they are likely to turn to *MoEML* for help with identifying specific locations, so it is preferable to process their XML files and identify the toponyms ourselves. We propose to begin our prototyping with the ten history plays because five of them are complete or nearly complete.

TEI versions of the *ISE* plays are currently indexed in an eXist XML database (like the *MoEML* texts) in order to provide search capabilities. The ISE textbase also includes modern-spelling versions of the texts, and all its versions of each core text are linked using "through line numbers" (TLNs) based on the *First Folio*. These features provide the basis for a comprehensive system to identify placename references throughout the texts.

We will use a multi-phase approach to identifying relevant placename instances. First, we will deploy a Named Entity Recognition tool such as the *Stanford Named Entity Recognizer*, trained on a subset of texts selected to provide sufficient variety of genre and known to include a useful number of London place references. We will combine the results with the entries in a dictionary of spelling variants of London placenames extracted from our *MoEML* collection. We will generate results in a form that includes:

- Candidate placename
- Surrounding context (paragraph, line selection, etc.)
- Link to online version of the text using TLN
- ID and name of candidate match location in *MoEML* database (if there is one)
- Link to *MoEML* location data

This manner of reporting the results will allow research assistants to rapidly accept, reject, or correct the placename instance. Confirmed references will be stored in a TEI document in the form of `<linkGrp>` elements: `<linkGrp target="mol:CHAR1" n="Charing Cross">`
`<link target="ise:1H4/M/scene/2.1#tln-659|76-88"/>`
`</linkGrp>`

This tagging encodes a link between the *MoEML* placeography (Charing Cross, which has the @xml:id "CHAR1") and the *ISE*'s modern-spelling version of *Henry IV Part 1*. Any other links to the same location will be encoded using `<link>` elements inside the same `<linkGrp>`. These links use Private URI Schemes for the sake of convenience. The pointers prefixed with "*mol:*" are dereferenced in the context of the *MoEML* database through XPath (`//TEI[@xml:id="CHAR1"]` in this case). The `"ise:"` prefix can be similarly dereferenced to construct a full URI to the target location in the document: http://internetshakespeare.uvic.ca/Library/Texts/1H4/M/scene/2.1#tln-659 . The last component of the pointer contains the character offset range for the placename. A formal method for documenting and mechanically dereferencing private URI schemes and similar abbreviated pointers has been proposed (Holmes 2012) and is being considered for adoption by the TEI Council.

Once candidate placenames have been encoded for the modern-spelling editions of the plays, the TLN referencing system in use by the *ISE* can be used to identify corresponding references in the other editions. We will use this automated process for identifications:

1. Retrieve the text following the corresponding TLN from an original-spelling edition of the text.
2. Search for the placename as it appears in the modern-spelling edition. If found, record its offsets and generate a `<link>`.
3. If not found, try a search for each variant spelling of the placename known to the *MoEML* database.
4. If a match is still not found, tokenize the target text, create bigrams and trigrams, and run similarity metrics between the original-spelling placename and each n-gram. If a similarity threshold is reached, assume a match and create a `<link>`.

Various similarity metrics might be appropriate here, including the Universal Similarity Metric (USM; see Holmes 2010). Where a similarity metric is invoked, the results will be flagged for manual checking. Pursuing this particular example, the *First Folio* has Charing Cross with the spelling "Charing-crosse". A Java implementation of the USM gives these a similarity score of 0.206, which represents high similarity (scores are between 0 and 1, with 0 representing identity). Processing the *First Folio* edition would generate a second `<link>`: `<link target="ise:1H4/F1/scene/2.1#tln-660|37-50"/>`

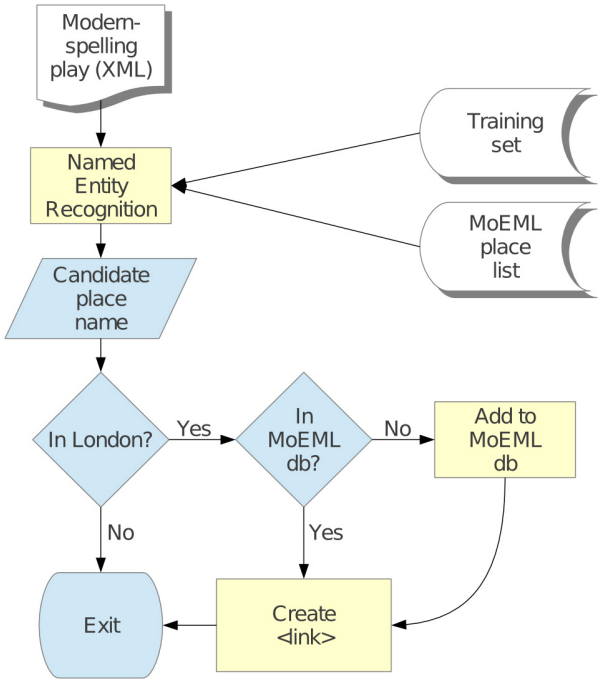


Figure 2:
A flowchart representing the placename identification process.

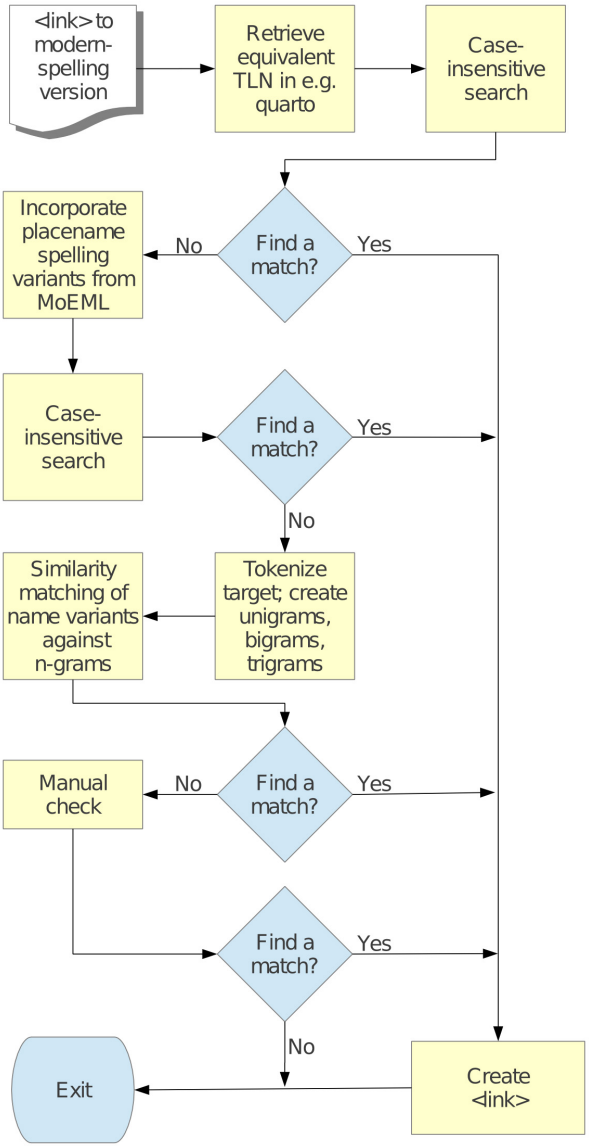


Figure 3:
Mapping placename identification in the modern-spelling texts onto original-spelling versions.

The complete process is represented in the flow charts in Figures 2 and 3. This approach will enable us to generate a large number of matches and resulting links without excessive human labour. The link groups will be stored in the *MoEML* database. No modification of *MoEML* or *ISE* texts is required; this is the "loose coupling" mentioned above. Links to instances of London placenames in the *ISE* texts can be provided as part of *MoEML*'s online placeography. Meanwhile, the *ISE* team has expressed interest in linking out to *MoEML* location data, which could easily be achieved either by processing the *MoEML* link groups to add annotations directly into the *ISE* texts, or

(pursuing the loose coupling methodology) by making calls to an *API* provided by *MoEML* when rendering sections of *ISE* texts to incorporate relevant links.

References

Bol, P. K., J. Hsiang, and G. Fong (2012). "Prosopographical Databases, Text-Mining, GIS and System Interoperability for Chinese History and Literature." *Digital Humanities 2012 Conference Abstracts*, 43-51.

Hirsch, B. (ed.) *Digital Renaissance Editions*. http://digitalrenaissance.uvic.ca/ .

Holmes, M. (2010). Using the Universal Similarity Metric to Map Correspondences between Witnesses. http://dh2010.cch.kcl.ac.uk/academic-programme/abstracts/papers/html/ab-693.html .

Holmes, M. (2012). "Prefix Definition Proposal." http://wiki.tei-c.org/index.php/Prefix_Definition_Proposal .

The Map of Early Modern London. Dir. Janelle Jenstad. http://mapoflondon.uvic.ca .

Matei, S. A. (2012). "ARTL@S and BasArt: A loose coupling strategy for digital humanities." *Artl@s Bulletin*, 1: 27-30.

McDonough, J. (2008). “Structural Metadata and the Social Limitation of Interoperability: A Sociotechnical View of XML and Digital Library Standards Development.” Presented at Balisage: The Markup Conference 2008, Montréal, Canada, August 12 - 15. In *Proceedings of Balisage: The Markup Conference 2008*. Balisage Series on Markup Technologies, 1. doi:10.4242/BalisageVol1.McDonough01. http://www.balisage.net/Proceedings/vol1/html/McDonough01/BalisageVol1-McDonough01.html .

Ostovich H., A. Griffin, P. Cockett, and J. Roberts-Smith (eds.) Queen's Men Editions. http://qme.internetshakespeare.uvic.ca/ .

Rasmussen, E., M. Best, et al. (eds.) *Internet Shakespeare Editions* http://internetshakespeare.uvic.ca .

Sperberg-McQueen, C. M. (2008). “But wait, there's more!” Presented at Balisage: The Markup Conference 2008, Montréal, Canada, August 12 - 15. In *Proceedings of Balisage: The Markup Conference 2008*. Balisage Series on Markup Technologies, 1. doi:10.4242/BalisageVol1.Sperberg-McQueen02. http://www.balisage.net/Proceedings/vol1/html/Sperberg-McQueen02/BalisageVol1-Sperberg-McQueen02.html .

Databases in Context: Transnational Compilations, and Networks of Women Writers from the Middle Ages to the Present

Hoogenboom, Hilde M.
hilde.hoogenboom@asu.edu
Arizona State University, United States of America

Overview

In the 1990s, databases began to expand the potential of quantitative approaches to create new connections between women, their writings, critics, readers and nations. However, in my research on an overlooked reference genre in women’s literary history, compilations — of biographies, bibliographies, and selected works of women, initially as (in)famous women in history, then as learned women and writers — I discovered that most databases enhance rather than transform the national narratives they have inherited, with a few important exceptions. Those national narratives became evident as I traced the development over the past 600 years of over a 100 compilations, a highly coherent, dynamic genre that began with Boccaccio’s *Famous Women* (1375), and gradually spread in waves from Italy to France, England, Denmark, Germany, Spain, Russia, the United States, and many smaller nations, initially as manuscripts, later as books, in the form of anthologies, biographies, bibliographies, treatises, and literary histories, and now as databases. Despite their quite varied generic properties, which are often hybrid, they are similar in how they function. Compilers list compilations, and rely on, compete and disagree with, and often simply borrow from their predecessors’ work not only nationally, which we would expect, but transnationally — features that make the genre cohere over centuries across national and linguistic boundaries. A comparative, historical survey of compilations reveals that women’s literary history is fundamentally relational, between nations (Hoogenboom, 2013). My project thus situates digital and quantitative scholarship on women in a long historical continuum that invites deeper reflection on the quantitative methods and assumptions underlying digital scholarship on women,

and what Ann Blair terms our historically shared "info-lust" (Blair, 2010).

The national limitations of existing databases on women writers, and opportunities to question priorities and structures that reflect the nationalist, canonical narratives of literary histories, are apparent from some recent quantitative papers. Women's writings remain underrepresented in textual and linguistic corpora, and thus in data mining, despite evidence that in some countries, at certain periods, women were writing more than men (for example, England from 1800 to 1830 (Garside et al., 2000), and Australia since 1990 (Bode, 2012)). In one project on data mining gender differences in French literature, the researchers note that, "The female corpus was assembled first, due to the more limited digital collection of women's writing at our disposal" (Argamon et al., 2009). Among researchers, Nowviskie notes the apparent paucity of women engaged in data mining, which may have causal connections with the lack of quantitative literary research on women's texts (Nowviskie, 2012). Women's writings raise questions about the representativeness of Franco Moretti's project on distant reading, and suggest that book history and bibliographic studies remain central to quantitative projects (Moretti, 2000; Trumpener, 2009).

Two innovative databases, the Orlando Project and *WomenWriters,* move beyond the traditional categories of poetry, prose fiction, and national literatures, to open national narratives to other genres and writers, and include nations other than the traditional cultural empires of France and England in transnational literary histories. My research on European and American compilations and Russian women writers uses an international database, *WomenWriters,* which maps the national and international networks of the reception of women writers throughout Europe before 1900 to join traditional humanities and digital scholarship to illustrate new narratives of European women's literary history. In its historical and geographical sweep, this project illuminates other networks beyond the cultural empires of Europe by integrating small nations, Western with Eastern Europe and Russia, and the U.S. This is also an umbrella project for smaller independent national databases that expands access to the resources and practices of digital humanities for women from diverse countries.

WomenWriters, run by New approaches to European Women Writers (NEWW, 2001-), is a unique international database stored and developed at Huygens ING of the Royal Netherlands Academy of Arts and Sciences (KNAW). *WomenWriters* uses data fields to link the relations between women and works through readers' reception, nationally and internationally in many countries large and small to fundamentally reassess the influence of women as writers as broadly as possible. The database contains over 4,000 women writers, 12,000 works, and 22,000 receptions, found in such large-scale sources as library catalogs, translations, the periodical press, and compilations, together with memoirs, letters, archives, and so on. The project was awarded a European Cooperation in Science and Technology (COST) networking grant (400,000€, 2009-13) and COST Action IS0901 "Women Writers In History" has over 120 participants from 25 European Union countries, as well as myself (in Working Group 3, on sources) and others from the United States and other non-EU countries (http://www.costwwih.net/home). In December 2012, NEWW and Huygens ING received a grant for 2013 from CLARINS-NL to upgrade the database to a Virtual Research Environment (VRE) running on REST data services, which Huygens developed to map the republic of letters (Huygens, Oxford, and Stanford). It will include faceted searches to dynamically visualize networks and trees in interactive timelines or geographical maps, and statistical analysis and charting to map reception networks and topoi. This upgrade will also establish connectivity with five other European databases.

Methodology

My sample maps for both select compilations and select Russian women writers use a preliminary VRE with new data fields that can show quantitative data geographically and over time of transnational reception of women writers, their biographies, and their texts. Rather than rely on a single approach, the combined methodologies of book history, bibliography, national expertise, and quantitative methods of *WomenWriters* maximize the potential national and transnational influence of women writers. Methodologically, quantitative work on Russian women writers is an instructive case study because, aside from expanding corpora to include women's writings from smaller nations, Franco Moretti (1998) shows that Russia is, like most nations, an importer of literature, but is exceptional in the amount it imported (over 80%).

WomenWriters has begun to test the systematic input of select contents of compilations of women writers from before 1900, beginning with eighteenth- and nineteenth-century French compilations, some of which found their way to England, Germany, and Russia. Since compilers often reference and borrow material from earlier compilations, this is an especially coherent way to track over time the presence and absence of writers together with the kinds of biographical reception material about them. These compilations can then be compared with national dictionaries, encyclopedias, and literary histories of writers on a larger scale to trace the national inclusion, exclusion, and changing reception topoi of women writers over time. I have done this selectively manually, comparing Russian

compilations of women with a handful of dictionaries, encyclopedias, and literary histories for Russian literature (Hoogenboom, 2008), and will present select data visually.

My national research area of Russia expands the reception networks for European women substantially because Russia had many women writers and was among the biggest importers of foreign literature in translation and in the original languages. Currently, *WomenWriters* contains the only database collection of Russian women writers, who were very active translators. Using compilations, we have input around 500 out of about 1,400 Russian women authors. We are also inputting the Russian reception of George Sand (1804-76), who at present is a central node in *WomenWriters* because the database began as a Dutch-French project and the director, Suzan van Dijk, is a George Sand scholar (Van Dijk, 2001). Sand was a central node in European networks of women writers and readers, especially for Russians, whose enthusiasm can be measured in the number and speed of translations, many by women, in comparison with other nations. Sand's predecessor on the international stage, Stéphanie-Félicité, Madame la Comtesse de Genlis (1746-1830), is also a node for Russia's connectedness to European literature in an earlier era. Women's hidden role as translators can slowly be made more visible in Russia and the many other nations that depended heavily on them for their reading. Thus as the database grows, it will be able to show significant, hitherto unseen, international connections and networks for such other international writers as Jeanne Leprince de Beaumont, Comtesse Dash and Ouida not only in literature, but also in, for example, pedagogy, religion, politics, and history.

Results

In anticipation of VRE maps and timelines later this year, the following links use the tree tool, the only data tool in the current version of *WomenWriters* to show an author's position as a node between predecessors and followers. In Russia, Akhmatova was a translator and publisher of a series of over 300 translated novels and Khvoshchinskaia was the most productive and respected woman writer of the second half of the nineteenth century, and the most highly paid writer by the serious literary journals after Ivan Turgenev and Lev Tolstoy in the 1870s; both women translated novels by Sand and many others.

George Sand:
http://neww.huygens.knaw.nl/treeviews/show/7

Elizaveta Nikolaevna Akhmatova:
http://neww.huygens.knaw.nl/treeviews/show/4934

Nadezhda Dmitrievna Khvoshchinskaia:
http://neww.huygens.knaw.nl/treeviews/show/199

References

Argamon, S., J.-B. Goulain, R. Horton, and M. Olsen. (2009). Vive la Différence! Text Mining Gender Difference in French Literature. *Digital Humanities Quarterly* 3.2. http://www.digitalhumanities.org/dhq/vol/3/2/000042/000042.html (accessed 3 November 2012).

Blair, A. M. (2010). *Too Much to Know: Managing Scholarly Information before the Modern Age.* New Haven, CT: Yale University Press.

Bode, K. (2012). Modeling Gender: The 'Rise and Rise' of the Australian Woman Novelist. *Digital Humanities 2012.* http://www.dh2012.uni-hamburg.de/conference/programme/abstracts/modeling-gender-the-rise-and-rise-of-the-australian-woman-novelist/ (accessed 3 November 2012).

Garside, P., J. Raven, and R. Schöwerling. (ed.) (2000).*The English Novel 1770-1829: A Bibliographical Survey of Prose Fiction Published in the British Isles.* 2 vols. New York: Oxford University Press.

Hoogenboom, H. (2013). Bibliography and National Canons: Women Writers in France, England, Germany, and Russia (1800-2010). *Comparative Literature Studies,* 50(2).

Hoogenboom, H. (2008). The Non-Canonical Canon: From Nikolai Novikov's *Historical Dictionary to Dictionary of Russian Women Writers.* In Hoogenboom, H., Nepomnyashchy, C., and Reyfman, I. (eds). *Mapping the Feminine: Russian Women and Cultural Difference.* Bloomington, IN: Slavica. 281-300.

Moretti, F. (2000). The Slaughterhouse of Literature. *Modern Language Quarterly* 61 207-27.

Nowviskie, B. (2012). What do Girls Dig? In Gold, M. K. (ed). *Debates in the Digital Humanities.* Minneapolis: University of Minnesota Press. 235-40.

Trumpener, K. (2009). Paratext and Genre System: A Response to Franco Moretti. *Critical Inquiry* 36(1): 159-71.

Van Dijk, S. (2001). *WomenWriters.* http://neww.huygens.knaw.nl/ (accessed 15 March 2013).

Almost All the Way Through — All at Once

Hoover, David L.
david.hoover@nyu.edu
New York University, United States of America

Computational stylistics over the past twenty-five years has focused mainly on the most frequent (function) words of texts. This focus has been based on the reasonable belief that very frequent words, especially function words, tend to be used in a routinized or habitual way. Such words seem unlikely to be affected by the conscious manipulations of authors and thus should be the safest words to use in authorship attribution and computational stylistics. However, there has been a recent trend of increasing the number of words for analysis, with improved results (Hoover 2001; Burrows 2002; Hoover 2007; Rybicki and Eder 2011). Recently, special attention has also been paid to individual parts of the word frequency spectrum, attention sparked by Burrows's "All the Way Through: Testing for Authorship in Different Frequency Strata" (2007), which introduces two new measures of textual difference, Zeta and Iota. Zeta focuses on words that are neither extremely frequent nor rare, and Iota focuses on relatively rare words. Both measures have been applied to a variety of texts (Craig and Kinney 2009; Hoover 2008, 2010, forthcoming). So far as I know, however, no one has suggested using the entire word spectrum all at once, and that is the focus of my proposal.

One problem with analyzing all the words is that standard statistical methods are inappropriate for rare words, as Burrows points out in his discussion of Iota (Burrows 2007: 36). However, both Zeta and Iota are based on presence/absence rather than frequency, and do not require any sophisticated statistical analysis. They are derived by dividing two authors' texts into segments of the same size and identifying "marker" words that are characteristic of the two authors. (Zeta and Iota can be used to compare any two groups of texts, but my discussion is based on a comparison of two authors for simplicity.) In his introduction of Zeta and Iota, Burrows begins with samples of text by Marvell and Waller and divides Marvell's sample into five equal segments. For Zeta, he analyzes only those words that appear in at least three of Marvell's five segments and have a maximum frequency of two in Waller's sample. For Iota, he analyzes only those words that appear in fewer than three of Marvell's segments and not at all in Waller's sample. Both measures eliminate the most frequent words, which appear in most segments of most texts.

Hugh Craig's version of Zeta (Craig and Kinney 2009), which I will modify to analyze the entire word list, can be explained more clearly by comparing Joseph Conrad and Ford Madox Ford, two authors who were involved in three problematic collaborations. First, samples of text by Ford and Conrad (twelve novels and novellas) are divided into equal-sized segments, here 3,000 words, 132 segments by Conrad and 131 by Ford. Zeta is calculated for each word by counting how many segments by each author contain the word and adding the percentage of Conrad's segments in which the word appears to the percentage of Ford's segments in which the word does not appear (with the percentages expressed as decimals). A word found in all of Conrad's segments but none of Ford's would have a Zeta score of two; with the occurrences reversed, the Zeta score would be zero. In practice, scores higher than 1.8 or lower than .2 are rare; here they range from 1.66 to .48 (the range depends on the size and number of segments, and on how different the authors are). Sorting the words on their Zeta scores identifies marker words that are consistently used by Conrad and consistently avoided by Ford, and vice versa.

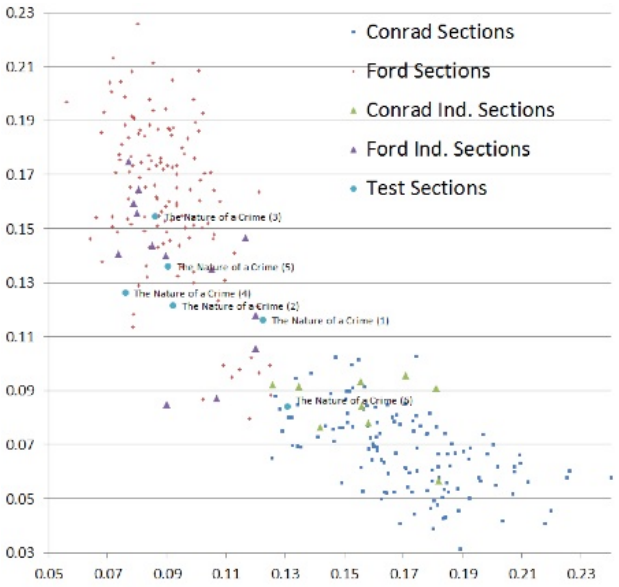


Fig. 1–
Craig Zeta Analysis of Conrad and Ford: 500 Most Distinctive Marker Words for Each

The results of a Craig Zeta analysis of Conrad and Ford are presented in Fig. 1, in which the X axis is the percentage of the types (individual word forms) in each text that are among the 500 most distinctive Conrad marker words, and the Y axis is the percentage of types that are among the 500 most distinctive Ford markers. Craig Zeta does a good, but not perfect, job of separating the segments of text by Conrad and Ford and in attributing some additional independent texts by the two authors–ones not involved in creating the lists of marker words. It also assigns all but the final segment of the collaborative *The Nature of a Crime* to Ford (most critics believe it was written almost entirely by Ford). For the Ford segments (upper left), Ford marker words account for a minimum of about 8% of the types and a maximum of about 23%, while Conrad marker words account for a minimum of about 6% and a maximum of about 13%. Conversely, about 13%-24% of the types in the Conrad segments (lower right) are Conrad marker words, but only about 3%-10% are Ford marker words. The part of

the word frequency spectrum that Zeta is capturing can be gauged by noting that the 1000 marker words here (500 for each author) appear in a range of 12 to 244 of the 263 total base segments, with frequencies ranging from 13 to 7296 and ranks ranging from 16 to 4416. In this analysis, Zeta eliminates the 15MFW, all but 8 of the 100MFW and more than 3/4 of the 200MFW.

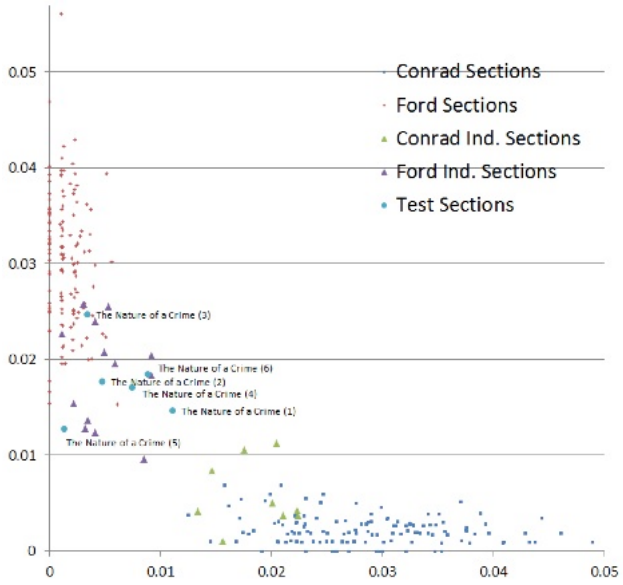


Fig. 2–
Craig Zeta Analysis of Conrad and Ford: 500 Most Distinctive Rare Markers for Each

Focusing only on the rest of the words, those appearing in 11 or fewer of the 263 base segments, results in Fig. 2, an analog of Burrows Iota, based on 500 marker words for each author with total frequencies ranging from 1 to 179 and ranks ranging from 4225 to 7797. The fact that the primary and independent texts are more clearly distinguished by these "Iota" markers than by the Zeta markers suggests that it may be useful to test the entire spectrum at once. The results of such a test are shown in Fig. 3, based on almost the full 28,177-word vocabulary of the combined texts by both authors —about 14000 marker words for each author. This analysis omits the 31 words that occur in every segment of every text because these words have a Zeta score of exactly 1, and so cannot help to distinguish the authors.[1]

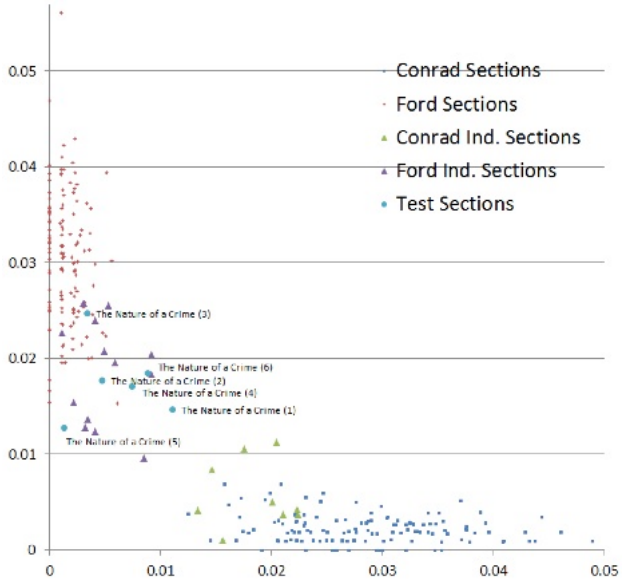


Fig. 3–
Analysis of (Almost) All the Words of Conrad and Ford: 14,000 Marker Words Each

Much more work is needed to determine how well analyzing the entire word frequency spectrum at once works on various groups of texts and authors, but the method seems promising, in spite of, or perhaps because of, the demonstration by Rybicki and Eder that different groups of texts and authors show different "sweet spots" in the word frequency spectrum (Rybicki and Eder 2011). One reason for this can be seen in Fig. 4, which shows that, in comparing Conrad and Ford, Conrad's most distinctive words tend to be found among the more frequent words (with lower ranks) than Ford's.[2] Perhaps using the entire spectrum at once can help to overcome some of the problems of using various methods with various groups of texts. At the very least it has the benefit of basing an argument about similarity and difference on (almost) all of the words of the texts — all at once.

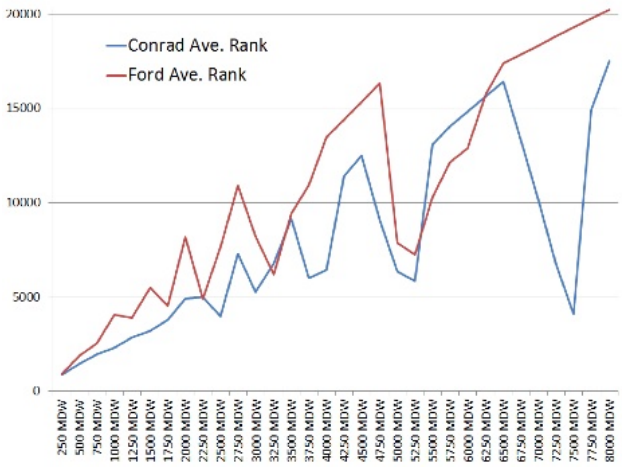


Fig. 4–
Average Ranks of the 8000 Most Distinctive Conrad and Ford Marker Words

References

Burrows, J. (2007). All the Way Through: Testing for Authorship in Different Frequency Strata, *LLC*, 22. 27-47.

Burrows, J. (2002). Delta: a Measure of Stylistic Difference and a Guide to Likely Authorship. *LLC* 17. 267-287.

Craig, H., and A. Kinney (eds) (2009). *Shakespeare, Computers, and the Mystery of Authorship*. Cambridge: Cambridge University Press.

Hoover, D. (2012). Text Analysis. In Price, K. and Siemens, R. (eds) *Literary Studies in the Digital Age*. New York: Modern Language Association.

Hoover, D. (2010). Authorial Style. In McIntyre, D. and Busse, B. (eds) *Language and Style: Essays in Honour of Mick Short*. London: Palgrave. 250-71.

Hoover, D. (2008). Searching for Style in Modern American Poetry. In Zyngier, S. et. al. (eds) *Directions in Empirical Literary Studies: Essays in Honor of Willie van Peer*. Amsterdam: John Benjamins. 211-27.

Hoover, D. (2007). Corpus Stylistics, Stylometry, and the Styles of Henry James. *Style* 41 174-203.

Hoover, D. (2001). Statistical Stylistics and Authorship Attribution: An Empirical Investigation. *LLC* 16 421-444.

Rybicki, J. and M. Eder (2011). Deeper Delta across genres and languages: do we really need the most frequent words? *LLC* 26. 315-21.

Notes

1. In analyses with equal numbers of segments, all words that occur in the same number of segments for each author, sometimes as many as 1,500-2000, also have a Zeta score of 1 and are eliminated; the potential effects of this need further investigation.

2. The X axis shows the average ranks for the 1-250 most distinctive marker words (MDW), the 251-500MDW, etc. The bizarre peaks and valleys are caused by words with radically different ranks but almost identical Zeta scores; for example, the rare word *who'd* (rank 10772) appears in 0 of 132 Conrad segments, and in 4 of 131 Ford segments (0% presence + 97% absence = a Zeta of .97), and the common word *little* (rank 58) appears in 127 Conrad segments and is absent from just 1 Ford segment (96.2% presence + .8% absence = a Zeta of .97).

The Full-Spectrum Text-Analysis Spreadsheet

Hoover, David L.
david.hoover@nyu.edu
New York University, United States of America

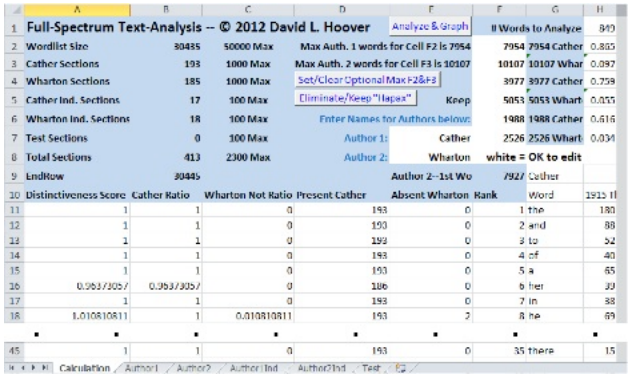
The most frequent function words have received the lion's share of attention in authorship attribution and computational stylistics, partly because they seem intuitively unlikely to be manipulated consciously by authors, and partly because analyses based on them have been quite successful. Rare words have sometimes also been studied (Baayen, H., van Halteren, H., and Tweedie. F. 1996; Holmes 1998). Burrows has recently introduced Iota, which focuses on relatively rare words, and Zeta, which focuses on words that are neither among the very most frequent words nor rare (Burrows 2007; Hoover 2007b; Hoover forthcoming; Craig and Kinney 2009). Other researchers have analyzed very large numbers of the most frequent words (Rybicki, J. and M. Eder 2011; Hoover 2007a). And Rudman has argued that "An individual's style is made up of hundreds and hundreds of markers. The more of these that can be shown to be used consistently (within the same genre and time constraints) by the author, the more that can be used in the study" (1998: 153). However, I know of no computational stylistics tool that analyzes the entire word frequency spectrum.

My Full-Spectrum Text-Analysis spreadsheet, designed to do just that, is a Microsoft Excel spreadsheet with macros, using a simple but powerful method of measuring differences between two groups of texts. It begins with sets of texts by two authors, divided into equal-sized segments, and then compares how many segments for each author contain each word, ignoring their frequencies. Any two groups can be compared, but here I describe the simplest

comparison, between two authors, Willa Cather and Edith Wharton.

The snippet from the Calculation sub-sheet of the spreadsheet in Fig. 1 clarifies how it is used (shown before the macro has completed the analysis; the buttons are explained below). In cells E7 and E8, the user enters the names of the two authors to be compared (automatically copied into columns A and G and Row 9), then enters the raw word frequencies for the segments into five sub-sheets (see the tabs at the bottom of Fig. 1). The frequencies for the segments that will be used to create the comparison between the two authors are placed in the "Author1" and "Author2" sub-sheets, with the full word list in column A of"Author1" (all segments in this analysis are 2,500 words). Optionally, independent segments by the same authors can be placed in "Author1Ind" and "Author2Ind" and used to confirm that the method correctly attributes texts not involved in creating the initial comparison. Finally, any texts to be tested for authorship are placed in "Test." (All word frequency lists must be based on the word list in "Author1.")

The macro, run by clicking the "Analyze & Graph" button, clears out old data, enters formulas, copies data from the sub-sheets into the "Calculation" sheet (columns H and following), shrinks the columns for easier reading. It also copies the word list into column G, and enters their ranks in column F (this is useful for studying where each author's characteristic words fall in the frequency spectrum). The calculations are performed in columns A-E. Column D records the number of Cather's segments that contain the word, and column E records the number of Wharton's segments that do *not* contain the word. The most frequent words typically occur in all segments and receive a neutral score of 1, but note that *her* the 6th most frequent, occurs in only 186 of the 193 Cather segments. Column B calculates the percentage of Cather's segments that contain each word; column C calculates the percentage of Wharton's segments that do *not* contain each word (both expressed as decimals). Column A sums columns B and C, producing the Distinctiveness Scores (DS). Columns H and following of row 1 show the number of different words (types) in each segment, and below them the percentage of types that are marker words for Cather or Wharton (these figures are not meaningful until the macro has finished). It sorts the words on the DS, with Cather's most distinctive marker at the top and Wharton's at the bottom, then selects Wharton's markers and re-sorts them in reverse order. The sheet can handle 50,000 words, but the full word list for these samples is 30,435 words.

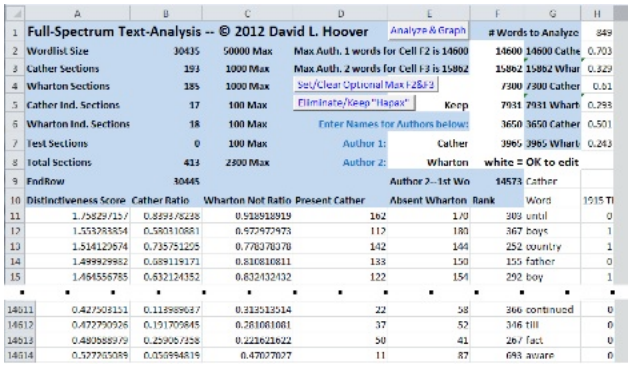


	A	B	C	D	E	F	G	H
1	Full-Spectrum Text-Analysis -- © 2012 David L. Hoover				Analyze & Graph	# Words to Analyze		849
2	Wordlist Size	30435	50000 Max	Max Auth. 1 words for Cell F2 is 7954		7954	7954 Cather	0.805
3	Cather Sections	193	1000 Max	Max Auth. 2 words for Cell F3 is 10107		10107	10107 Whar	0.097
4	Wharton Sections	185	1000 Max	Set/Clear Optional Max F2&F3		3977	3977 Cather	0.759
5	Cather Ind. Sections	17	100 Max	Eliminate/Keep "Hapax"	Keep	5053	5053 Whart	0.055
6	Wharton Ind. Sections	18	100 Max	Enter Names for Authors below:		1988	1988 Cather	0.616
7	Test Sections	0	100 Max	Author 1:	Cather	2526	2526 Whart	0.034
8	Total Sections	413	2300 Max	Author 2:	Wharton	white = OK to edit		
9	EndRow	30445			Author 2--1st Wo	7927	Cather	
10	Distinctiveness Score	Cather Ratio	Wharton Not Ratio	Present Cather	Absent Wharton	Rank	Word	1915 tl
11	1	1	0	193	0	1	the	180
12	1	1	0	193	0	2	and	88
13	1	1	0	193	0	3	to	52
14	1	1	0	193	0	4	of	40
15	1	1	0	193	0	5	a	65
16	0.9637305	0.9637305	0	186	0	6	her	99
17	1	1	0	193	0	7	in	88
18	1.010810811	1	0.010810811	193	2	8	he	69
	▪	▪	▪	▪	▪	▪	▪	▪
45	1	1	0	193	0	35	there	15

Calculation | Author1 | Author2 | Author1Ind | Author2Ind | Test

Fig. 1 —
The Full-Spectrum Text-Analysis Spreadsheet, With Data, Macro Not Finished

After the macro has finished, the spreadsheet looks like Fig. 2. Cather's most distinctive marker, *until*(row 11), is found in 162 of the 193 Cather segments and is absent from 170 of the 185 Wharton segments. Wharton's most distinctive marker, *continued* (row 14,611), is found in just 22 of the 193 Cather segments, but is absent from only 58 of the 185 Wharton segments. (Note that *till*, a nice authorial contrast to *until*, is Wharton's second most distinctive marker.) Rows 2 and 3 of columns H and following show how these markers are distributed in each segment. For example, H2-H3 shows that about 70% of the types in this segment are Cather markers, but only about 33% are Wharton markers. For Wharton's first segment (not shown), the proportions are roughly reversed: about 66% of the types are Wharton markers and about 37% are Cather markers.



	A	B	C	D	E	F	G	H
1	Full-Spectrum Text-Analysis -- © 2012 David L. Hoover				Analyze & Graph	# Words to Analyze		849
2	Wordlist Size	30435	50000 Max	Max Auth. 1 words for Cell F2 is 14600		14600	14600 Cathe	0.703
3	Cather Sections	193	1000 Max	Max Auth. 2 words for Cell F3 is 15862		15862	15862 Whar	0.329
4	Wharton Sections	185	1000 Max	Set/Clear Optional Max F2&F3		7300	7300 Cather	0.61
5	Cather Ind. Sections	17	100 Max	Eliminate/Keep "Hapax"	Keep	7931	7931 Whart	0.298
6	Wharton Ind. Sections	18	100 Max	Enter Names for Authors below:		3650	3650 Cather	0.501
7	Test Sections	0	100 Max	Author 1:	Cather	3965	3965 Whart	0.243
8	Total Sections	413	2300 Max	Author 2:	Wharton	white = OK to edit		
9	EndRow	30445			Author 2--1st Wo	14573	Cather	
10	Distinctiveness Score	Cather Ratio	Wharton Not Ratio	Present Cather	Absent Wharton	Rank	Word	1915 Tl
11	1.758297157	0.8393782383	0.918918919	162	170	309	until	0
12	1.553283854	0.580310881	0.972972973	112	180	367	boys	1
13	1.514120674	0.735751295	0.778378378	142	144	252	country	1
14	1.489929982	0.689119171	0.810810811	133	150	155	father	0
15	1.464556785	0.632124352	0.832432432	122	154	292	boy	1
	▪	▪	▪	▪	▪	▪	▪	▪
14611	0.427503151	0.118998627	0.318513514	22	58	366	continued	0
14612	0.472790926	0.191709845	0.381081081	37	52	346	till	0
14613	0.480588979	0.259067358	0.221621622	50	41	267	fact	0
14614	0.527765089	0.359994819	0.47027027	11	87	698	aware	0

Fig. 2 —
The Full-Spectrum Text-Analysis Spreadsheet, With Cather and Wharton Data

The macro also creates the scatter graph in Fig. 3. The horizontal and vertical axes record the percentage of types in each segment that are Cather and Wharton markers, respectively. Note that all the independent and test segments are correctly attributed. (I have put some texts

in “Author1Ind” and “Author2Ind” and some in “Test” for illustration and have deleted some labels to make the graph easier to read.) Although full-spectrum analysis produces excellent results for these authors and texts, more work will be needed to evaluate its general effectiveness fairly.

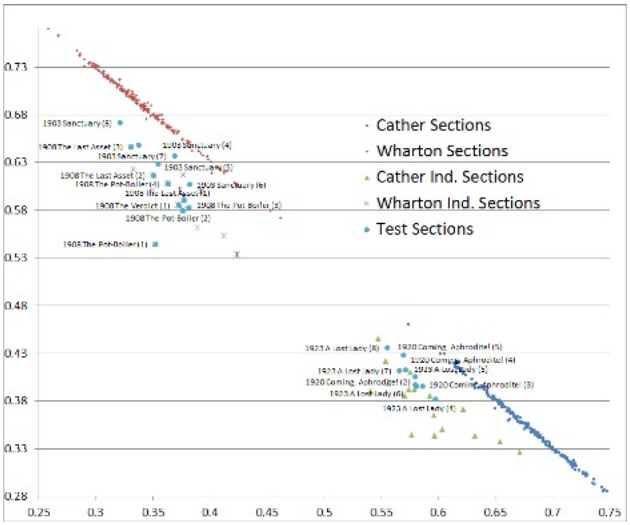


Fig. 3 —
A Full-Spectrum Text-Analysis Scatter Graph, With Cather and Wharton Data

The spreadsheet is designed to facilitate further study. I have included a button that toggles the elimination of hapax legomena. Setting this option to “Eliminate” before clicking the “Analyze & Graph” button removes these words from the analysis. The default range of analysis is full-spectrum, but the “Set/Clear Optional Max F2&F3” button toggles a limit of 500 marker words for each author, for comparison with Craig Zeta, which it then mimics. Half the maximum markers for each author are calculated in cells F4 and F5 and one fourth in F6 and F7, and three sets of results based on these three pairs of numbers appear rows 2-7 of column H and following. If these or any other numbers are pasted into cells F2 and F3, the graph automatically updates, facilitating a comparison between full-spectrum and more limited analyses.

I conclude with two more graphs, Fig. 4 showing the same analysis as Fig. 3, but without the hapax legomena, and Fig. 5 showing line graphs of the same information in Fig. 3 and Fig. 4 along with information based on the 500 most distinctive Craig Zeta markers for comparison (the sheet pastes the data on which these graphs are based to the right of the word frequency information, ready for graphing). These line graphs show less information about how various segments compare, but give a clearer picture of how many Cather and Wharton markers appear in each segment.

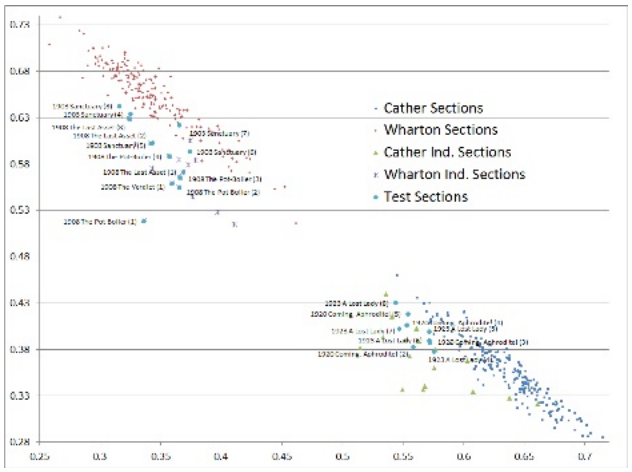


Fig. 4 —
A Full-Spectrum Text-Analysis Scatter Graph, With Cather and Wharton Data

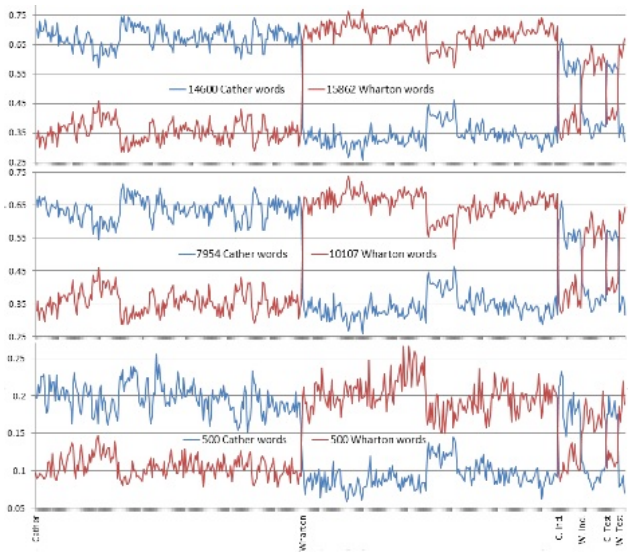


Fig. 5 —
Cather vs Wharton: Full-Spectrum, Full-Spectrum Less Hapax, 500 Markers Each

References

Burrows, J. (2007). All the Way Through: Testing for Authorship in Different Frequency Strata. *LLC* 22: 27-47.

Baayen, H., H. van Halteren, and F. Tweedie. (1996). Outside the Cave of Shadows: Using Syntactic Annotation to Enhance Authorship Attribution. *LLC* 11: 121-132.

Craig, H., and A. Kinney. (eds.) (2009). *Shakespeare, Computers, and the Mystery of Authorship.* Cambridge: Cambridge University Press.

Holmes, D. (1998). The Evolution of Stylometry in Humanities Scholarship. *LLC* 13: 111-117.

Hoover, D. (Forthcoming). "Text Analysis," in Ken Price and Ray Siemens (eds), *Literary Studies in the Digital Age.* New York: MLA.

Hoover, D. (2011). *Delta, Zeta, and Iota: An Ngrammatical Investigation* 'Language Individuation: A Symposium in Honour of John Burrows'. held July 4-8, 2011 at University of Newcastle. Australia.

Hoover, D. (2007a). Corpus Stylistics, Stylometry, and the Styles of Henry James. *Style* 41: 174-203.

Hoover, D. (2007b). Quantitative Analysis and Literary Studies. In Schreibman, S., and Siemens, R. (eds). *A Companion to Digital Literary Studies.* Oxford: Blackwell. 517-33.

Rudman, J. (1998). Non-traditional Authorship Attribution Studies in the Historia Augusta: Some Caveats. *LLC* 13: 151-57.

Rybicki, J. and M. Eder. (2011). Deeper Delta Across Genres and Languages: Do We Really Need the Most Frequent Words? *LLC* 26: 315-21.

Reading the Visual Page of Victorian Poetry

Houston, Natalie M

nhouston@uh.edu
University of Houston

Audenaert, Neal

neal.audenaert@gmail.com
Digital Archives, Research & Technology Services (DARTS)

The digitization of nineteenth-century texts offers us the opportunity of asking new research questions that could transform our historical understanding of Victorian culture. Digital access to the breadth of nineteenth-century print culture, which included books, periodicals, and newspapers published for an ever-increasing reading audience, puts pressure on traditional configurations of the literary canon, which examines only a limited number of authors and texts. As Dan Cohen asks, 'Should we be worrying that our scholarship might be anecdotally correct but comprehensively wrong? Is 1 or 10 or 100 or 1000 books an adequate sample to know the Victorians?' (Cohen). In developing *VisualPage*, a software application for the large-scale identification and analysis of the graphical elements of digitized printed books, we will enable researchers to identify unique or representative examples across very large data sets of digitized texts. Such computational analysis will reveal new ways of thinking about both the printed book and its digitized forms. This paper presents the current development of this proof-of-concept software (funded in 2012 by a Level II NEH Digital Humanities Start-Up Grant) and some findings from the analysis of our initial data set.

Large Scale Analysis and Victorian Books

As Franco Moretti notes, the traditional Victorian canon includes only a 'minimal fraction of the literary field' constituted by all the texts published in the period. He calls for new methods of analysis because 'a field this large cannot be understood by stitching together separate bits of knowledge about individual cases, because it isn't a sum of individual cases: it's a collective system' (3-4). To understand the literary field as a system requires examining what Pierre Bourdieu calls 'position-takings' — all of the actions, persons, and objects that produce a work of art and its cultural value, including those 'which, because they were part of the self-evident givens of the situation, remain unremarked and are therefore unlikely to be mentioned in contemporary accounts' (30-31). The structures of cultural value that surrounded Alfred Tennyson's *In Memoriam*, Christina Rossetti's *Goblin Market and Other Poems*, or any other volume of Victorian poetry were partly created by all the other books of poetry, including what Bourdieu calls the 'unremarked' ones that were overlooked by both Victorian critics and scholars in our own day.

Digitization offers us the possibility of expanding our study of Victorian literature to include previously 'unremarked' texts. However, most tools for large-scale research focus on the linguistic content of texts, through either syntactic or semantic analysis. But printed texts simultaneously convey meaning through both linguistic and graphic signs. As Jerome McGann suggests, 'text documents, while coded bibliographically and semantically, are all marked graphically' and a 'page of printed or scripted text should thus be understood as a certain kind of graphic interface' (138, 199). We've taken poetry as our starting point for *VisualPage* because the visual appearance of the printed page contributes to the reader's understanding of the poem's form and meaning through the conventions of line capitalization, punctuation, and indentation.

Printed poems are typically framed by the white space created by line endings, creating a distinctive visual signal of the genre on the printed page. Experienced readers evaluate the graphical codes of printed texts quickly, often subconsciously; as Johanna Drucker suggests, 'we see before we read and the recognition thus produced predisposes us to reading according to specific graphic codes before we engage with the language of the text' (242). In Victorian books of poetry, for example, rhymed lines

were frequently indented the same distance from the left margin to visually indicate the poem's form and structure. Rhyme is thus simultaneously a linguistic, poetic, and graphic feature of many Victorian books.

Scholars have long realized that 'Typographic transcriptions . . . abstract texts from the artifacts in which they are versioned and embodied' (Viscomi 29). Although full bibliographical analysis of a book is not available from a digital surrogate, digital images of a book's pages offer researchers more information about 'the interaction of its physical characteristics with its signifying strategies' than can text alone (Hayles 103). Accordingly, most scholarly digital archive projects today recognize the value of this graphical meaning and provide users access to both digitized page images and plain text versions. But until now, researchers have been limited to only what their human eyes can see or recognize in those page images. Developing tools for the large scale graphical analysis of digitized books will contribute to a broader understanding of literature's circulation, consumption, and function within Victorian culture.

Overview of the *VisualPage* Application

In order to make the visual structure of document images explicit and available to both computational processing and interactive human analysis, the *VisualPage* application is designed around three inter-related tasks. The Feature Extraction module analyzes the digitized page images in order to translate pixels into the language of visual features used to design and analyze page layout: typeface size; margin size; width, height and spacing of text lines; and more. These features are then organized in relation to bibliographic categories, such as volumes, poems, and pages, in order to enable questions such as 'how much variability is there in the length of lines in poems from two different publishers?' or 'how does the visual density of a page change for this publisher over time?' *VisualPage* is designed so that the specific set of features extracted from a collection can be changed in response to new analytical needs and new technical capabilities.

Once these features have been extracted and stored in an attribute-relation file (ARFF), the next task is to discover relationships within the data. This is the responsibility of the Pattern Recognition module. The pattern recognition module will support basic queries such as 'find all poems that use dropped capital letters' or 'find poems whose line length is in the bottom 25% of poems from this publisher.' It will also enable more sophisticated data mining based on machine-learning techniques. Simple examples include the ability to cluster documents based on a set of features such as margin width and line height or to find documents that are 'visually similar' to a set of known documents.

Finally, the Analysis module presents data visualization and exploration interfaces. This is the outward-facing portion of the application that allows scholars to interact with the documents and to harness the pattern recognition tools in order to pose new questions and discover new relationships within the collection.

During the start-up phase of this project, we are focusing software development work toward two main objectives. First, we are designing the main structure of each module and implementing an initial set of features that can be extended and enhanced through future work. Second, we are performing an initial proof-of-concept prototyping for the more technically complex components of the system. Notably, this includes:

- recognition of the low-level image features
- understanding the higher-level structure in terms of both poetry (e.g., titles, epigraphs, stanzas as they are found within a single page and across multiple pages) and page layout (margins, running heads, page numbers, footnotes, etc.)
- analyzing these structures using pattern recognition and machine learning techniques

This proof-of-concept work addresses questions about which approaches hold the most promise for scholarly research in addition to demonstrating the technical feasibility of our approach. In order to ensure that the techniques we develop are appropriate to collections beyond that which can be easily analyzed and comprehended by a single scholar, our initial data set consists of 300 single-author books of poetry published between 1860-1880, or approximately 60,000 page images.

Initial Research Findings

In presenting research findings from our initial data set of single-author books of poetry, we will focus on three main areas of research:

- identifying historical changes correlated with particular publishers in the printing of poetry during the period 1860-1880
- analyzing line indenting and line length to understand Victorian rhyme and poetic form across a varied set of authors and texts
- identifying computationally significant patterns or trends in the graphical design of Victorian books

Our *VisualPage* software enables researchers to move beyond our human capacity to view, compare, and understand only a limited number of texts at one time. Large-scale analysis of the graphical dimensions of previously 'unremarked' books offers us the possibility of understanding the cultural field of Victorian poetry in all its historical complexity.

Funding

This work was supported by the National Endowment for the Humanities [HD5156012].

References

Bourdieu, P. (1993). The Field of Cultural Production, or: The Economic World Reversed. In Johnson, R. (ed). *The Field of Cultural Production: Essays on Art and Literature*. New York: Columbia University Press.

Cohen, D. (2010). Searching for the Victorians. *Dan Cohen*, http://www.dancohen.org/2010/10/04/searching-for-the-victorians (accessed 4 October 2010).

Drucker, J. (2009). *SpecLab: Digital Aesthetics and Projects in Speculative Computing*. Chicago: University of Chicago Press.

Hayles, N. K. (2005). *My Mother was a Computer: Digital Subjects and Literary Texts*. Chicago: University of Chicago Press.

McGann, J. (2001). *Radiant Textuality: Literature after the World Wide Web*. New York: Palgrave Macmillan.

Moretti, F. (2005). *Graphs, Maps, Trees: Abstract Models for a Literary History*. London: Verso.

Viscomi, J. (2002). Digital Facsimiles: Reading the William Blake Archive. *Computers and the Humanities* 36(1). 27-48.

Coding Media History: A Digital Suite for Opening Access, Building Tools, and Analyzing Texts

Hoyt, Eric Rutledge
ehoyt@wisc.edu
University of Wisconsin-Madison

Description

This short paper seeks to introduce the Digital Humanities community to three ongoing, interrelated projects: the Media History Digital Library (an open access digital resource), Lantern (a search tool) and Coding Media History (a text mining research project). Together, these three projects aim to use digital technology to transform the field of Film and Media Studies, a discipline that has lagged English and History in the creation of high impact DH work.

I would also like to suggest that the three interrelated projects demonstrate a productive model for scaffolding work in the Digital Humanities. The three sides of this work—enabling access, building tools, and text analysis — support and enhance one another. In the space below, I will briefly address all three projects and suggest the ways they enrich one another.

In terms of enabling access, the Media History Digital Library (www.mediahistoryproject.org) has digitized over 500,000 pages of out-of-copyright periodicals relating to the histories of film, broadcasting, and recorded sound. Prior to the launch of the MHDL, scholars wrote the histories of film and television through page-by-page microfilm readings of key periodicals, such as *Moving Picture World* and *Photoplay*. By scanning these publications along with previously unavailable materials, the MHDL enables scholars to conduct research more efficiently, ask new questions, and write new histories.

The MHDL's collections are open access and built on a collaborative model. David Pierce and I lead the project, and we work closely with collectors, who loan materials, and sponsors, who pay for the scanning. The scanning is carried out by the Internet Archive (www.archive.org), which also hosts and preserves the digital files. By using the Internet Archive as a scanning vendor and provider of backend infrastructure, the MHDL follows in the tradition of other collaboratively built digital collections, including the Biodiversity Heritage Library (http://www.biodiversitylibrary.org/), Medical Heritage Library (http://www.medicalheritage.org/), International Children's Digital Library (http://en.childrenslibrary.org/), and International Music Score Library Project (http://imslp.org/).

Film and media educators at institutions around the world are already incorporating the Media History Digital Library into their teaching. In one especially creative assignment, Elizabeth Clarke is having her students at Wilfrid Laurier University in Waterloo, Ontario read the MHDL's digital editions of early cinema magazines and imagine they are the intended audience of motion picture exhibitors in the 1910s. Students are asked to design their own programs of short films and live entertainment based on what they discover inside the magazines.

The MHDL's diverse user-base encompasses students, educators, expert researchers, and casual classic movie

fans. In order to better serve all of these groups, I have been leading the development of Lantern, a software tool that is a co-production of the Media History Digital Library and UW-Madison's Department of Communication Arts. Lantern offers users the ability to perform fulltext searches across the Media History Digital Library's entire corpus. Eventually, we also hope to equip Lantern with powerful functionalities beyond search, such as topic modeling and network visualizations.

My team and I are developing Lantern through using Ruby on Rails, Python, XML, and CSS and customizing three open source technologies: Apache's Solr search engine; the University of Virginia Library's Blacklight interface; and the Internet Archive's BookReader. We are currently indexing more materials into Lantern, overhauling its graphic interface, and enhancing its speed and functionality. We anticipate publicly launching Lantern in Summer 2013. In the meantime, you may view a work-in-progress demo at http://lantern-demo.commarts.wisc.edu/

The third project I want to address is a work-in-progress called "Coding Media History: Computational Analysis of the Hollywood Trade Press." Despite the heavy reliance of film and television scholars on Variety and other industry trade papers, there has been little work that reflexively examines these sources. My research project, Coding Media History, uses computer analytics both to enrich our understanding of these key sources and destabilize the notion that we can conceive of 60 years of *Variety* as a singular "text." In pursuit of these goals, I borrow from the text mining methods (and warnings) of Stephen Rasmsay and, especially, from Andrew J. Torget, Rada Mihalcea, Jon Christensen, and Geoff McGhee's work on applying topic modeling and text mining to historical newspapers.

I have begun the process of working with a research assistant, who is marking-up the XML of the digitized publications. We will soon be able to start asking research questions over the marked-up corpus. In a 1905 issue of *Variety*, for instance, what percent of the pages were dedicated to vaudeville compared to motion pictures? How were these page allocations different in 1915, 1925, 1935, 1945, and 1955? When were radio and television introduced as their own sections? How did the buyers and amounts of advertising change over time? These are questions that I can answer by starting with the digitized magazines, adding a research assistant's tags, and finally running my own algorithms over the marked-up corpus.

One of the questions I am exploring is the extent to which the various trade papers were truly similar or different from one another. The Hollywood trade papers have an infamous reputation for publishing the exact same studio press releases. By using open source plagiarism software, we can test whether this reputation is warranted. The answers to these questions hold real stakes. Consider the case of *Motion Picture Herald,* a trade paper that proclaimed to represent the interests of independent movie theatre owners. What does it mean if we discover that *Motion Picture Herald* published 40% of the same content as the trade papers that spoke to producers and the major studios? Can we truly think of Motion Picture Herald as representing the independent theatre owners' interests?

In conclusion, the Media History Digital Library, Lantern, and Coding Media History are already making a positive intervention in the field of Film and Media Studies. I also hope that this suite of interrelated projects can serve as a useful model for scholars in other fields pursuing Digital Humanities projects. In the course of my work, I've found that being involved across a suite of activities (digitization, tool building, and text analysis) leads to better decision-making at every stage in the process. Although it's unrealistic to expect that we'll all become hybrid librarian-programmer-scholars, we need to better understand the integrated range of activities in order for the Digital Humanities to tackle bold new projects and break free of our tokenized comfort zones.

References

Hoyt, E., W. Hagenmaier, and C. Hagenmaier (2013). "Media + History + Digital + Library: An Experiment in Synthesis." *Journal of E-Media Studies* 3: forthcoming.

Ramsay, S. (2011). *Reading Machines: Toward an Algorithmic Criticism.* Urbana: University of Illinois Press.

Torget, A. J., et al. (2011). *Mapping Texts: Combining Text-Mining and Geo-Visualization To Unlock The Research Potential of Historical Newspapers.* UNT Digital Library. http://digital.library.unt.edu/ark:/67531/metadc83797/ . accessed March 9, 2013).

Yang, T.-I., A. J. Torget, and R. Mihalcea. (2011). "Topic Modeling on Historical Newspapers." *Proceedings of the Association for Computational Linguistics workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities* (June): 96-104. http://aclweb.org/anthology/W/W11/W11-15.pdf. (accessed March 9, 2013).

Reading Habits & Attitude in the Digital Environment: A Study on Dhaka University Students

Islam, Md. Anwarul
anwar81du@gmail.com
Univeristy of Dhaka

Technological advancement has made books available not only in printed format but also in electronic format. Bangladesh is currently experiencing the exponential growth of information and entertainment being created in a digital format. The government of Bangladesh declares 'Digital Bangladesh' by 2020. This steps, activities and environment are gaining importance particularly among younger people in Bangladesh (Islam & Tsuji 2011). This phenomenon may change the way people perceive about reading and how printed materials are being utilized to facilitate reading. At present this paper is the first attempt to measure reading habits and attitudes in digital environment of the students in Bangladesh. This study may trigger more such research in other developing countries.

Reading culture in Bangladesh

Based on the survey conducted Bishwa Sahitya Kendra meaning 'World Literature Center' the reading interest and habits of Bangladeshis is very low (Bishwa Sahitya Kendra, 2012). In Bangladesh, primary, secondary and higher secondary level education are completely in Bengali medium and in higher education is both in Bengali and English languages. As most of the web contents in Bangladesh are in English language, it usually reduces the reading habits in online environment. The electronic media is challenging the reading habit in the society by shifting the attention to computer, mobile phone and internet. In the last few years it has grown dramatically, although obviously from a very low base. With an estimated internet user-base of 7.5 million coming into 2012, representing a 5% user penetration by population, the local internet industry has been preparing to move into the next stage of its development (Internet World Statistics, 2012). In a recent survey by the Bangladesh Bureau of Statistics (BBS), it was found that the literacy rate is only 56.8% (CIA, 2012).

Research about reading has been approached from various possible angles and from a variety of disciplinary backgrounds including literature, social science, library and information science, information systems and more recently information and communication technologies (ICT). With the growing amount of digital information available and the increasing amount of time that people spend reading electronic media, the digital environment has begun to affect people's reading behavior. Studies on reading habits and attitude among university students has gained as much attention in recent years due to the impact of digital media made available through the internet (Liu 2005; Ramirez 2003). Several theorists in reading and literacy such as Landow (1992), Lanham (1993), O'Donnell (1998) and Murray (1997) all agree that the digital media brought through progressive development of ICT has introduced a transformative shift in reading and writing. University students have been known to be very receptive to different forms of media in their reading and writing practices. A number of scholars argue that the arrival of digital media, together with the fragmentary nature of hypertext, is threatening sustained reading (Healy 1990; Birkerts 1994). Birkerts (1994) further notes that the younger generation growing up in the digital environment lacks the ability to read deeply and to sustain a prolonged engagement in reading. Bolter (1991) states: "The shift from print to the computer does not mean the end of literacy itself, but the literacy of print, for electronic technology offers us a new kind of book and new ways to write and read". Digital media contribute to a transformative shift in reading. They also introduce a number of powerful advantages that are traditionally absent in the printed environment, such as interactivity, non-linearity, immediacy of accessing information, and the convergence of text and images, audio and video (Landow 1992; Lanham 1993; Murray 1997; Ross 2003). Abidin, Pour-Mohammadi & Choon Lean (2011) conducted a study on Malaysian Chinese university students and it is revealed that the participants prefer the electronic media when reading for leisure but prefer the printed media to pass exams.

Methodology

In this study, the researcher intends to explore the reading attitude and habits among the Dhaka University students using a survey research method.

Online survey tool

An online survey tool Kwik Surveys was used for this research. The URL to the online questionnaire was sent mainly through different SNTs using personal messaging option and group post where the university students were connected via Facebook, Twitter, etc. Also, the URL was printed and provided to the students to respond to the questionnaire.

The questionnaire

The survey requested basic demographic data regarding age, gender and academic status and also contained items regarding the reading attitudes and habits. Students were asked to provide answers of viz, how often do you read in

a week, what types of materials do they read, resources to get the reading materials, reading materials in leisure time, times spend on reading in online and so on. The survey also included 7-point Likert scale items regarding students' attitudes towards reading, preference of reading in online and manual (1 being the lowest and 7 being the highest). The data analysis was carried out using SPSS statistical analysis software.

Data analysis techniques

In order to determine influence of students' demographic characteristics on their opinions on reading attitudes and behaviour, Mann-Whitney and Kruskall-Wallis tests were carried out. Descriptive statistics were used to analyze demographic characteristics of the students in relation to their reading books in digital environment. A total 192 students responded to the online survey.

Objectives of the study

This study attempts to answer the following Major Research Questions (MRQs) as formulated below:

MRQ 1. What is the reading habit of the students in terms of the following?

(a) What type of reading material do they read?
(b) How much time do student spend on reading?
(c) Where do they get the reading material?
(d) When do they read?
(e) What do they read during leisure time?
(f) What percentage of time reading spent on reading electronic documents, browsing and scanning, and their overall experience with online reading?
(g) What are the tools they use access to electronic resources?

MRQ2. What is the students' attitude towards reading?

(a) Do they love to read book?
(b) Do they think reading is boring?
(c) Do they feel pleasure to spend money for buying book?
(d) Which medium (printed or online) do they feel comfort for reading book?
(d) What is their overall reading attitudes and other relevant issues?

MRQ3. Is there any relationship between gender, age, study level with their reading habit and attitude?

Findings

This study was conducted in an attempt to enhance our understanding about reading habits and attitudes of the university students in Bangladesh focusing on a case of a public university. In this effort, students from all the faculties were chosen as the respondents. In overall analysis that include both groups, results indicates that university students spend quite a significant amount of time reading newspapers 84.90% and second highest is the academic books 61.46%. A good number of students (50%) read also website materials. Reading has also become a major activity during their leisure time. Most of the students (57.81%) read fiction and novel in their leisure time and it is followed by newspaper 55.73%, magazine 43.75% and website materials (29.17%). The amount of time spent on reading other materials by the university students is seen as higher than their academic books and materials. This group is expected to read more due to their engagement in the academic process that requires them to read. Despite different extent in the preference among different genders and different study level, the study finds that preference for reading printed text remains strong which is 51.56%. This clearly indicates that printed media is the more used than online media due to the lack of facilities, poor online content and less availability of reading materials in online. On the other hand, students prefer reading in online environment when read short documents, casual reading and most recent information which mean score is above 4. In online environment, access to electronic resources mobile phone, laptop and desktop are the most used tools. Attitudes towards reading was high as the statements love to read books, like to read books when have free time and enjoyment from reading score was above 4. Several limitations can be found in the conduct of this study. Among them are the small sample size, the inclusion of only one university and the limited amount of variables studied. A bigger scale study need to be conducted for more reliable results, and with the inclusion of more variables such as family background, reading exposure and availability of reading materias, and variables that are related specifically with reading in the digital environment. In addition, the findings of the study should assist the university authority, especially the library and the computing department to look into service matters pertaining to accommodating the reading as well as the studying habits of the student. 24 hours of computing service may also allow students to use the internet since the day time is fully occupied with classes. This practice has been carried out in many western high academic institutions.

References

BSK (2012). Bishwa Sahitya Kendra available at: http://en.wikipedia.org/wiki/Bishwa_Sahitya_Kendra

Birkerts, S. (1994). *The Gutenberg Elegies: The Fate of Reading in an Electronic Age*. Boston: Faber and Faber.

Bolter, J. D. (1991). *Writing Spaces: The Computer, Hypertext, and the History of Writing*. Hillsdale, NJ: Laurence Erlbaum Associates.

CIA (2012). The World Factbook: Bangladesh, available at: https://www.cia.gov/library/publications/the-world-factbook/geos/bg.html .

Healy, J. M. (1990). *Endangered Minds: Why Our Children Don't Think*. New York: Simon and Schuster.

Internet World Statistics (2012). Bangladesh-Internet Markets and Forecasts, available at: http://www.internetworldstats.com/asia/bd.htm

Islam, M. A., and K. Tsuji. (2011). Bridging digital divide in Bangladesh: study on community information centers, The Electronic Library. 29 (4): 506–522

Landow, G. (1992). *Hypertext: The Convergence of Technology and Contemporary Critical Theory*. Baltimore: Johns Hopkins University Press.

Lanham, R. (1993). *The Electronic Word: Technology, Democracy, and the Arts*. Chicago: University of Chicago Press.

Liu, Z. (2005). Reading behavior in the digital environment: changes in reading behavior over the past 10 years, *Journal of Documentation*. 61(6): 700-12.

Murray, J. H. (1997). *Hamlet on the Holodeck: The Future of Narrative in Cyberspace*. Boston: MIT Press.

O'Donnell, J. J. (1998). *Avatars of the Word: From Papyrus to Cyberspace*. Cambridge, MA: Harvard University Press.

Abidin, M. J. Z, M. Pour-Mohammadi, and O. Choon Lean (2011). The reading habits of Malaysian Chinese University Students. *Journal of Studies in Education.* 1(1):E9.

Ramirez, E. (2003). "The impact of the internet on the reading practices of a university community: the case of UNAM", paper presented at the World Library and Information Congress: 69th IFLA General Conference and Council, Berlin, August 1-9.

Ross, C. S. (2003). "Reading in a digital age", available at www.camls.org/ce/ross.pdf .

Visualizing Uncertainty: How to Use the Fuzzy Data of 550 Medieval Texts?

Jänicke, Stefan

stjaenicke@informatik.uni-leipzig.de
Institut für Informatik, University of Leipzig, Germany

Wrisley, David Joseph

dw04@aub.edu.lb
Department of English, American University of Beirut, Lebanon

Our research project *Visualizing Medieval Places* brings together a computer scientist and a literary historian. We use the web-based tool *GeoTemCo* (Jänicke, 2012) to visualize thousands of place names against a focusable timeline. The resulting geospatial-temporal visualization is a way for the researcher to analyze space and time in a historical corpus of literature. The ideal user interface will allow manipulation of the visualization by (1) dynamically changing the thresholds for both visualizing and suppressing given types of uncertainty in the geospatial and temporal dimensions, and (2) adding or removing facets (e.g. particular genres, time ranges) to broaden or constrain the amount of data to be displayed. This interactivity will hopefully allow for controlled visualization of literary data, and will facilitate the formation of nuanced, supportable hypotheses about time and space in literature. Figure 1 illustrates the current user interface.

Figure 1:
The current GeoTemCo user interface, showing 636 data points from the Franco-Italian dataset. NB: The time line at the bottom does not represent the real temporal data of the project.

The data set is being built using a corpus of nearly 550 medieval French texts. Unlike English or Classics, scholars of medieval French have few electronic texts at their disposal. Furthermore, spelling variance of toponyms in medieval vernacular texts poses a significant challenge for any semi-automatic extraction. We are considering

combining our geospatial data of French places with Latin place names to build a bilingual gazetteer for use by the digital medievalist community in the future. For now, the toponyms (and their variants) are being manually harvested from a canonical reference work, the *Table des noms propres* (Flutre, 1962); they are subsequently disambiguated and geocoded. Unlike narratologically-inspired digital literary geographies such as *The Literary Atlas of Europe* (Hurni, Piatti et al., 2012) constructed on close readings of fictional and vernacular spaces, the data model for *Visualizing Medieval Places* includes only real geographical place names. It shares more with the GIS-based analysis of unstructured texts found in the Lancaster University initiative entitled *Spatial Humanities* (Gregory et al., 2012). Since Flutre's work does not fully represent the variety of textual communities and genres of medieval French, we are also extracting place names from name indices in selected critical editions. The first subset of data points from Franco-Italian literature is virtually complete. The project uses the crowd-sourced *Archives de littérature du moyen âge* (Brun, 2012) to enrich the metadata about the texts.

Using the data has proved problematic since so many aspects of it are uncertain. Situating the composition of medieval texts in a specific time and place can be at best speculative. Date formats of traditional scholarship have been represented in idiosyncratic ways (e.g. between 1095-1291, first half of the 14th century, before 1453). Likewise, the toponyms found in these works are difficult for various reasons: they are unmappable, they can refer to multiple places, or they designate ancient Greco-Latin or medieval geographical zones no longer found on the contemporary map.

The visualization of uncertainty is a hot topic in the visualization community. Despite a broad set of applications in this field, there are still no straightforward solutions for displaying multiple, overlapping kinds of uncertainty within one set of visual interfaces. Drawing upon a long list of uncertainty types (Griethe et al., 2006), a data item within our project might be said to embody two basic kinds of uncertainty. The first uncertainty is one of "lineage," by which we mean the reliability of the text source. Certainty values for lineage can simultaneously affect the representation of data items in both dimensions, the geospatial and temporal. The second uncertainty is one of "accuracy," referring to the granularity of place or time, that is, to the distinctly sized intervals in which a value can lie. Again, granularity impacts both dimensions, the geospatial (with units such as landmarks, localities, regions, countries, continents) and the temporal(years, eras, as well as upper- or lower-bounded time declarations). Unlike Rees, who primarily uses transparency to depict uncertain information (Rees, 2012), we need to investigate multiple visual metaphors that represent several dimensions of uncertainty in a clear way.

Visualizing distinct, overlapping geographic entities with different certainty values represents a major challenge for the project. Inspired by MacEachren's overview of existing methods for the geospatial (MacEachren et al., 2005), we suggest testing pairwise mixtures of color hue, texture, saturation and transparency, as well as other features such as pop-up text, backgrounds and overlapping/non-overlapping shapes to encode lineage and accuracy uncertainties. Figure 2 demonstrates how we use different shapes to encode objects with distinct geospatial accuracies.

Figure 2:
A visualization of 636 points from the Franco-Italian dataset, with non-overlapping shapes denoting different toponym types.

Although some work addresses the problem of temporal uncertainty for small datasets (Zuk et al., 2005), sufficient research on large-scale temporal uncertainty is not available. Expecting thousands of overlapping temporal values of variable granularity on the timeline, we need to create novel visualization approaches. Figure 3 illustrates one solution for dealing with aggregated temporal uncertainty where increased saturation designates a higher degree of certainty. In our short presentation, we will demonstrate the tool and some strategies for simultaneous visualization of various aspects of the data.

Figure 3:
: A timeline represented as a stacked graph with multiple aggregated uncertain temporal values (increased saturation designates increased certainty in lineage value). NB: The time line below does not represent the real temporal data of this project.

The project hopes to bring attention back to hundreds of unread works of the period (Moretti, 2005), perhaps even spawning new close readings of them based on their "interspatiality" — the common spaces that texts reference — but also to encourage students and scholars to experiment with visualizing spatial clusters and patterns

References

Brun, L. (2012). *Archives de littérature du moyen âge.* University of Ottawa http://www.arlima.net (accessed 28 October 2012).

Flutre, L.-F. (1962). *Table des noms propres avec toutes leurs variantes figurant dans les romans du moyen âge écrits en français ou en provençal et actuellement publiés ou analysés.*

Gregory, I. (2012). *Spatial Humanities: Texts, Geographic Information Systems and Places.* http://www.lancs.ac.uk/spatialhum/ (accessed 15 February 2013)

Griethe, H., H. Schumann (2006). *The Visualization of Uncertain Data: Methods and Problems.* In *Proceedings of SimVis'06*, 143-156.

Harris, R. L. (1999). *Information Graphics: A Comprehensive Illustrated Reference.*

Hurni, L., B. Piatti, et al. (2013). *A Literary Atlas of Europe — Ein Literarischer Atlas Europas.* http://www.literaturatlas.eu/ [accessed 1 March 2013].

Jänicke, S. (2012) *GeoTemCo: Comparative Visualization of Geospatial-Temporal Data* http://www.informatik.uni-leipzig.de/geotemco (accessed 29 October 2012)

MacEachren, A., A. Robinson, S. Hopper, R.M. Gardner, M. Gahegan, and E. Hetzler (2005). *Visualizing Geospatial Information Uncertainty: What We Know and What We Need to Know.* In *Cartography and Geographic Information Science* 32.3 139-160.

Moretti, F. (2005). *Graphs Maps Trees: Alternative Models for a Literary History.*

Rees, G. P. (2012). *Uncertain Date, Uncertain Place: Interpreting the History of Jewish Communities in the Byzantine Empire using GIS.* Abstract DH2012, Hamburg. http://www.dh2012.uni-hamburg.de/conference/programme/abstracts/uncertain-date-uncertain-place-interpreting-the-history-of-jewish-communities-in-the-byzantine-empire-using-gis/ (accessed 14 March 2013).

Zuk, T., S. Carpendale, and W. Glanzman (2005). Visualizing Temporal Uncertainty in 3D Virtual Reconstructions. *In Proceedings of the 6th International Symposium on Virtual Reality, Archaeology and Cultural Heritage* (VAST'05). 99-106.

A concept of data modeling for the humanities

Jannidis, Fotis

fotis.jannidis@uni-wuerzburg.de
Institut für Deutsche Philologie, University of Würzburg, Germany

Flanders, Julia

Julia_Flanders@brown.edu
Women Writers Project, Brown University, United States of America

Data modeling is one of the main tasks of digital humanists. They engage in it in creating databases, corpora, digital editions, geographical information systems, research collections, digital libraries, project-specific markup and also crosswalks for heterogenous data collections. In light of this variety of activities, however, it is interesting to note that there is no general theory of data modeling in the digital humanities. Looking at the important works that have defined digital humanities in the recent years, this discrepancy becomes even more noteworthy. On the one hand McCarty 2005 regards modeling as the very basic of humanities computing; on the other hand there is almost no general literature on data modeling in the digital humanities. An explanation for this could be the general feeling that there is no need for research on data modeling because the field of computer science has a vast amount of literature on it. But a closer look soon reveals that in that context the term 'data modeling' refers almost exclusively to database design (for an overview, see Simsion 2007). This discussion carried on by computer scientists and practitioners is detailed and elaborate, but it cannot be a substitute for research on data modeling in the humanities. The relational database – as important as it is – has not become the sole and prototypical data modeling activity in the humanities, in the way that text encoding has. On the principles of text encoding there is an important research literature (for example DeRose et al. 1990, Maler/Andaloussi 1995, or Hockey 2000) and even more literature on specific problems of the intellectual tools used by most digital humanists

nowadays, such as the problem of overlapping hierarchies (Schmidt 2010). But it is not our goal to replace one main data modeling activity with another one, but to combine the insights of those different activities and to ask, in a very preliminary way, what ideas could be part of a first outline of a general theory of data modeling in the humanities.

One interesting lesson to learn from the computer science literature on data modeling is the difference between a conceptual and a logical data model. The conceptual data model identifies and describes the entities and their relationship in the 'universe of discourse' and displays its result in a graphical notation such as an entity-relationship-diagram. The logical data model defines the tables of a database according to the underlying relational model. In data modeling related to textual material, people usually describe their results in prose and map them to a schema, but there is no graphical notation, or at least no system that is generally agreed upon in the manner of Chen's ER-diagram or the crow foot notation. This is probably less due to the lack of graphical notation than to the lack of a generally accepted conceptual model for XML (for a comprehensive overview of conceptual models for XML see Haitao Chen / Husheng Liao 2010). Besides the technical challenges for establishing such a model it is interesting to note that in the digital humanities there seems to be no real need for it. Even the tree notation proposed by Maler/ Andaloussi never caught on. So a deeper understanding is needed why we can find this two very different practices in data modeling.

It is a common feature of literature on data modeling that in order to create and evaluate a model one has to have a clear understanding of the user requirements for the data model. We note an interesting duality in this respect: on the one hand, data models serve as an interchange format for some types of users and user communities where data is typically being created and modeled with someone else's needs in mind (archives, libraries, others whom we might characterize as "altruistic modelers"). On the other hand, data models also exist whose function is to express specific research ideas in cases where data is being created to support the creator's own research needs (particularly for individual scholars and projects, whom we might characterize as "egoistic modelers"). Altruistic modelers also make assumptions what features of the digital objects are of interest for most users and in most use cases, while egoistic modelers can and will concentrate on their own needs. Thus, we have in practice very different ways of modeling: one that tries to include very different views on digital objects and aiming to establish standards (and this involves very specific processes to decide on these user needs and to connect these new models with existing traditions of modeling, for example in library science), and another that is interested mainly in expressing as exactly as possible the theoretical assumptions and research interests of one or more scholars. Often the data model for research purposes is evolving during the research process and many functions of schemas are not that important in such a context.

To conclude our discussion we want to propose a generic definition of data modeling: It refers to the activity to design a model of some real (or fictional) world segment to fulfill a specific set of user requirements using one or more of the meta models available in order to make some aspects of the data computable, to enable consistency constraints and to establish a common field of perception. Two typical forms of user requirements can be found: one which aims to define a data model for a domain in order to support data exchange and long-term use; and alternatively, one that is interested above all in modeling specific research assumptions and is often only used for a short time in a specific research context. In general a data model consists of a conceptual part which defines the data semantics, the relevant entities, their features and their relations, and a logical part which expresses a subset of the conceptual model in such a way that it is an abstract, self-contained, logical definition of the data, data operators, and so forth that together make up an abstract machine which makes the data computable.

Thus far we have been considering the general aspects of data modeling but in the final section of the paper we turn to the question of whether there is something specific to all of the data modeling activities in the humanities, something which sets them apart. Although we do not assume that this is a clear cut line, we argue that two features are of particular importance for data modeling in the humanities: first, that the objects of the data modeling activities are artifacts and many of their properties are not only man-made but intentionally created. And second, that (in contrast for example to business applications) not only do the objects of humanities research possess a long history which humanist usually respect and want to handle adequately, but in addition the research on these objects has a comparatively long history as well, and our models have to convey the complexities of this research. The implications of these features for a theory of data modeling in the humanities will have to be the topic of further research in the future.

References

Chen, H., and H. Liao. (2010). A Survey to Conceptual Modeling for XML. *In proceedings of: Computer Science and Information Technology (ICCSIT)*, 3rd IEEE International Conference on, Volume: 8.

Hockey, S. (2001). *Electronic Texts in the Humanities: Principles and Practice*. Oxford: Oxford University Press.

Maler, E., and J. El Andaloussi (1995). Developing SGML Dtds: From Text to Model to Markup. Prentice Hall.

McCarty, W. (2005). *Humanities Computing*. Houndsmill: Palgrave.

DeRose, S. J., D. G. Durand, E. Mylonas, and A. H. Renear (1990). "What is Text, Really?" *Journal of Computing in Higher Education* 2:1 3-26.

Schmidt, D. (2010). *The inadequacy of embedded markup for cultural heritage texts*. In Literary and Linguistic Computing 25(3): 337-356.

Simsion, G. (2007). *Data Modeling: Theory and Practice*. Bradley Beach: Technics Publications.

Eighteenth- and Twenty-First-Century Genres of Topical Knowledge

Jennings, Collin
crj237@nyu.edu
New York University, United States of America

Binder, Jeff
jmb783@nyu.edu
New York University, United States of America

Our paper examines the similarities and differences between two technologies for representing topical information from different historical periods. Taking the 1784 index to Adam Smith's seminal theory of political economy, *An Inquiry into the Nature and Causes of the Wealth of Nations* (1776), as a case study, we investigate the contrasting logics and aesthetics of knowledge organization apparent in the lateeighteenthcentury index and the twentyfirstcentury topic modeling algorithm known as latent dirichlet allocation. To accomplish this, we have collaborated to create a tool we call the *Networked Corpus*.[1] Inspired by the way readers mark passages that share a common theme, trope, or other feature, the *Networked Corpus* provides a computational way to navigate a corpus based on a topic model, while also facilitating comparison between this mechanical model and the humancreated index of Smith's text. Our presentation will suggest a historical lineage between modern topic modeling and the eighteenthcentury concepts of topic (which developed from the rhetorical idea of "topos") and system (which emerged in the allencompassing moral philosophy of the period) that are both exhibited in the index. We argue that the intentionally anachronistic comparison of topic modeling with the eighteenthcentury index reveals similarities and differences between what each approach counts as a salient feature of a text for the purpose of organizing and representing its topical, informational, or conceptual content.

Topic modeling is a family of statistical methods that attempt to find "latent" semantic content in texts, under the assumption that texts exhibit mixtures of "topics" that have characteristic vocabularies.[2] Topic modeling software attempts to find the topic definitions that best fit a given set of texts, while inferring which texts exhibit which "topics" based on the words that they use. This method was originally proposed as an information retrieval tool, with the goal of enabling people to search for broad themes that cannot necessarily be identified with particular words. As such, it competes with the subject index; but it differs from traditional indexing both in its assumptions and in the form the output takes. Preparing a subject index generally involves imagining what a reader might want to find — "think of the user" is the "motto" given in G. Norman Knight's classic indexing textbook — and results in a product that suits those who can describe what they are looking for.[3] Topic modeling, by contrast, is based on a *generative model* —an abstract description of the process through which texts were produced — and constructs "topics" that do not necessarily correspond to anything that can be easily described. Using the output of a topic model as an index requires that the topics be labeled, something that requires an interpretive judgment that is often very difficult to make, and that can often, as scholars using topic modeling have argued, be misleading.[4]

The first stage of our project was an attempt to create an information retrieval program that better suits this limitation of topic modeling than forms that are tailored for users who already know what they are looking for, such as the index or the search engine. We wrote a Python script that takes in a collection of texts and the output of the topic modeling program MALLET, and produces an HTML version of the corpus with interactive navigation features.[5] In addition to an index of the "topics" in the model, the output includes asterisks in the margin next to passages where there is a particularly high concentration of a given topic relative to the concentration in the text as a whole — "exemplary" passages, as we are calling them. Clicking on an asterisk summons a popup box that contains links to other "exemplary" passages for the same topic. This construct is intended to enable navigation not from a "topic" to a passage, as in an index, but from one passage to another. It encourages the user to read until they come across something interesting that has been marked with an asterisk, and then see where the links go. The networklike structure of this tool gives the topic model an exploratory function that does not depend on any prior knowledge of what topics there are, or of what, if any, significance the topics have.

The *Networked Corpus* also includes features that are intended to make topic models easier to interpret. The user has the option to "explain the relevance" of a topic, showing a box listing the words most strongly associated with that topic, the "exemplary" passages that the program found for that topic, and other texts in the corpus that also contain a high concentration of that topic. The "explain" feature also highlights all of the words in the text that arose from the selected topic according to the model, and shows the density of the topic over the course of the text as a sideways line graph that runs in the margin. This visualization gives an idea of which parts of a text contribute to its association with a given topic and which do not, providing a rich body of both positive and negative evidence by which the topic model can be interpreted.

In the second stage of our project, we are investigating how the "topics" of a topic model differ from the notion of topic or subject that was employed in the lateeighteenth-century index. Recently literary and intellectual historians including Ann M. Blair and Leah Price have described the historical development of practices of indexing, commonplacing, and generally what we might call "topical" knowledge in the seventeenth and eighteenth centuries.[6] As a lecturer in rhetoric and belles lettres and later a professor of moral philosophy, Adam Smith participated in defining the significance of those activities for the emergence of modern disciplines such as literature, history, and political economy. Upon the publication of the first edition of *The Wealth of Nations* without an index, Smith's friend and fellow university professor, Hugh Blair, encouraged him to add an index and a syllabus like the ones they used "to give in [their] college lectures" because those additions would offer "Exhibit a Scientifical View of the Whole System."[7] For Blair, like Smith, representing the knowledge a text contained in various comprehensible and manageable forms was a critial aspect of the production of new knowledge.

Our paper considers whether new epistemological units produced by digital methods may facilitate comparative examination of similar ones from earlier periods. More specifically, we track the development of topical knowledge in the eighteenth century by comparing indices and commonplace books of the period to algorithmically produced "topics." During the last year we have collaborated to create a tool we call the *Networked Corpus*. Inspired by the way readers mark passages that share a common theme, trope, or other feature, the *Networked Corpus* provides a computational way to connect topics across the entire range of a corpus. We are currently working on a project that compares the historical changes to the content and form of eighteenth-century indices and commonplace books and topic modeling output from the corpus.

This "scientifical" approach to indexing suggests a particular model of the text, in the sense that Willard McCarty uses the word "model" — a "fictional or idealized representation" that one can see the text through.[8] The models actually employed in the construction of Scottish Enlightenment texts cannot, of course, be directly observed, but as a way of investigating one of these models speculatively, we produced a special version of the *Networked Corpus* for *The Wealth of Nations* that presents the highly detailed index that appeared in the 1784 edition of the book alongside a topic model generated from the text (Figure 1). To facilitate comparison of these two constructs at a conceptual level, we transformed both the index and the topic model into something similar to marginal annotations, showing all of the index headings that reference a page on the screen when the page is displayed, along with a list of prominent topics on the page. We also used the Spearman rank correlation to find topics that tend to strongly match the pages referenced under particular index headings, and indicated the pages on which these correlations break down. Based on a reading of these points of disagreement, we contend that the topic model is able to pick up on rhetorical moves in the text that are not represented within the sort of system of concepts that the index constructs, at the cost of never being able to claim the sort of exhaustiveness that the Scottish Enlightenment writers sought.

In this presentation, we suggest that the approach we have taken in our study of The Wealth of Nations — comparing constructs from different time periods that address a problem in radically different ways — opens up a new avenue for examining both contemporary text mining models and the models that are implicit in the organization of historical texts. As McCarty has observed, modeling supports an "orientation to questioning rather than to answers, and opening up rather than glossing over the inevitable discrepancies between representation and reality."[9] Our deformation of *The Wealth of Nations* employs a statistical model not as a way of studying the text itself, but as a vantage point from which we can examine the assumptions and blind spots of another, historical model of textual organization, turning this discrepancy into something of hermeneutic use. We thus agree with Alan Liu, who encourages scholars to pursue "any mediation that produces a sense of anachronism (residual or emergent, in Raymond Williams's vocabulary) able to make us see history as a compound relation of proximity and distance between past and present."[10] With the *Networked Corpus*, we suggest a way of doing this that converts the alienness of mechanical methods of reading in comparison to older models into a productive source of tension.

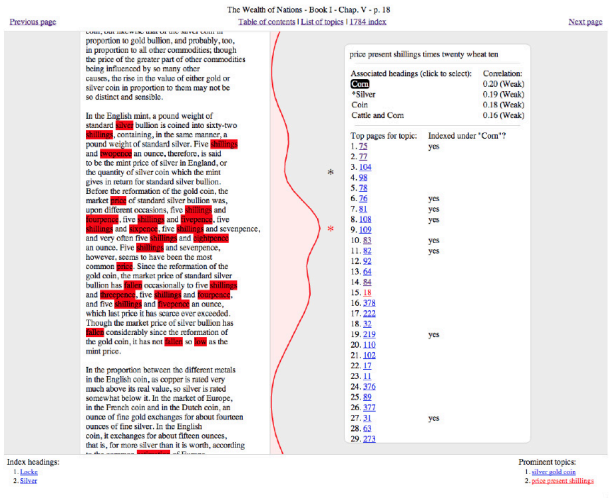


Figure 1:
Screen shot from the Networked Corpus

References

Blair, A. M. (2011). *Too Much to Know: Managing Scholarly Information Before the Modern Age.* New Haven: Yale University Press.

Blei, D. M., A. Y. Ng, and M. I. Jordan. (2003). Latent Dirichlet Allocation. *The Journal of Machine Learning Research* 3: 993–1022.

Blei, D. M. (2012). Probabilistic Topic Models. *Communications of the ACM* 55.4: 77–84.

McCallum, A. K. (2002). MALLET: A Machine Learning for Language Toolkit. http://mallet.cs.umass.edu.

McCarty, W. (2005). *Humanities Computing.* Palgrave Macmillan.

Price, L. (2003). *The Anthology and the Rise of the Novel: From Richardson to George Eliot.* Cambridge: Cambridge University Press.

Notes

1. Our tool can be found at networkedcorpus.com.
2. See D. M. Blei, "Probabilistic Topic Models," *Communications of the ACM 55*, no. 4 (2012): 77–84 for a brief review of topic modeling and D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet Allocation," *The Journal of Machine Learning Research* 3 (2003): 993–1022 for a more detailed explanation.
3. G. Norman Knight, *Indexing, the Art of* (George Allen & Unwin, 1979), 159.
4. See recent discussions of this challenge on Ted Underwood's blog post, "Visualizing Topic Models," *The Shell and the Stone,* 11 November 2012, Accessed 13 March 2013, and Ben Schmidt's blog post, "When you have a MALLET, everything looks like a nail," *Sapping Attention,* 2 November 2012, Accessed 13 March 2013.
5. Andrew Kachites McCallum, "MALLET: A Machine Learning for Language Toolkit.", 2002, http://mallet.cs.umass.edu .
6. See Ann M. Blair, *Too Much to Know: Managing Scholarly Information Before the Modern Age* (New Haven, CT: Yale University Press, 2011), 137–144 and Leah Price, *The Anthology and the Rise of the Novel: From Richardson to George Eliot* (Cambridge; New York: Cambridge University Press, 2003), 67–99.
7. Hugh Blair, "Letter to Adam Smith, 3 April 1776," *The Correspondence of Adam Smith*, ed. by Ernest Campbell Mossner and Ian Simpson Ross (Oxford: Clarendon Press: 1977), 189.
8. Willard McCarty, "Modelling," Humanities Computing (Palgrave Macmillan, 2005), 24.
9. McCarty, Humanities Computing, 38.
10. Alan Liu, Local Transcendence: Essays on Postmodern Historicism and the Database (University of Chicago, 2008), 25.

Collaborative technologies for Knowledge Socialization: the case of elBulli

Jiménez-Mavillard, Antonio
ajimene6@uwo.ca
The CulturePlex Laboratory - Western University Canada

Suárez, Juan Luis
jsuarez@uwo.ca
The CulturePlex Laboratory - Western University Canada

Introduction

Today, organizations face the crucial challenge of creating and managing knowledge in order to succeed. As part of the *Knowledge Management* process, *Knowledge Socialization* is a critical step during which the community experiences a decisive interchange of ideas. In this work, we present **a new model for *Knowledge Management*** based on the classic Nonaka and Takeuchi's one but adapted to the Web 2.0 by using wiki technologies to support *Knowledge*

Socialization, and we propose **to apply this model to the case of *elBulli*** .

elBulli, voted by industry authority *Restaurant* magazine as the best restaurant in the world in 2002 and from 2006 to 2009 (William, 2012), has now become a foundation for creativity and innovation in high cuisine. It incorporates disciplines such as technology, science, philosophy, and the arts in its research. Aware of the value of knowledge, the organization publishes its results in international conferences, books or journal articles, in a similar way to the academic process of peer review. Therefore, *elBulli* is an appropriate case to apply a *Knowledge Management* model that makes maximum use of its knowledge.

Knowledge as an asset

In recent years, the technological development and globalization have produced significant structural changes in society and the economy. New emergent industries embody a new economic reality: knowledge has become the main economic resource (Drucker, 1969). Indeed, nowadays, knowledge is the asset that generates more value in an organization, and the competitiveness of companies depends heavily on how they maintain and access their knowledge (Fensel, 2004; Davies, 2003).

Ikujiro Nonaka and Hirotaka Takeuchi propose a model whereby the creation of knowledge is a continuous and iterative process that transforms *tacit* knowledge (individual and subjective) into *explicit* (objective and shared) and vice versa (Nonaka, 1995). This process goes through four phases:

- **Socialization** is the action of sharing tacit knowledge with other individuals by means of observation, imitation and practice during collective work (for example, celebrating a meeting).
- **Externalization** is the action of making tacit knowledge explicit (for example, writing an article) in order to share it.
- **Combination** is the process of synthesizing more complex explicit knowledge from various sources of simpler explicit knowledge (for example, building a prototype).
- **Internalization** is the assimilation of explicit knowledge. It occurs when we understand explicit knowledge through experience (for example, nobody can tell you how to ride a bike, you need to practise to learn).

Knowledge at *elBulli*

elBulli is especially fascinating because of the processes from *elBulli Restaurant* to *elBulli Foundation* and the way the organization has redefined itself in the last few years: evolution and reinvention, diffusion of their practices, concepts and techniques across time and social space, etc. (William, 2012)

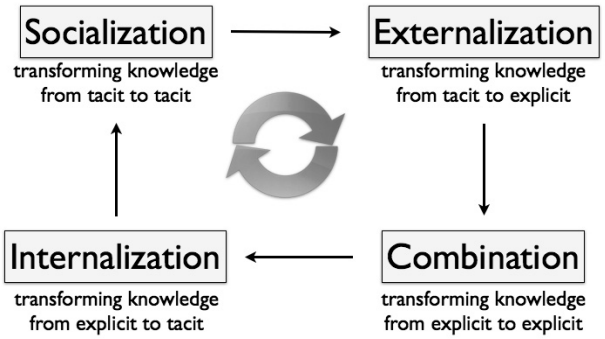


Illustration 1:
Nonaka & Takeuchi's cycle (Aranda-Corral et al., 2010)

By applying Nonaka and Takeuchi's principles, we can demonstrate that **this organization creates valuable knowledge:**

- **Socialization.** Creativity and innovation implies constant learning so that its employees attend specialized courses, gastronomic fairs, research stays, conferences, etc.
- **Externalization.** Unlike the traditional model, in which a chef would closely guard his secrets, *elBulli* documents exhaustively all its achievements and developments since 1990 (catalogues, audiovisual resources, recipes…) and it publishes its research in conferences, books or journal articles.
- **Combination.** The process of creation can be seen as an engineering process (Soler, 2007):
  1. Initial idea
  2. Use of one of the creative methods known
  3. Tests
  4. Analysis and reflection (wisdom and previous knowledge)
  5. Finish and last tests
  6. Prototype
  7. Customer's feedback
  8. Last changes
  9. Finished and catalogued dish

  The combination of existing creative methods, techniques, previous knowledge and known recipes results in new creative methods and new dishes.
- **Internalization.** Cooking is a suitable example of knowledge internalization: only by practising these new

methods and recipes, chefs can learn how to prepare new dishes.

Adrià suggests externalizing all their knowledge onto the *Bullipedia*, an online database that *"will contain every piece of gastronomic knowledge ever gathered"*. Adrià and some of the best chefs all over the world are putting in common their wisdom to agree on the content of the *Bullipedia*. Their aim is to create an encyclopedia with 15,000 articles "*where users will leave suggestions for dishes, concepts and combinations of flavours"* that will affect, in Adrià's opinion, other chefs' creativity.

This agreement can only be achieved through a sound process of socialization. According to Nonaka & Takeuchi's cycle, socialization entails a transformation from individual tacit knowledge into collective tacit knowledge. This conversion is extremely hard to carry out due to tacit knowledge being highly personal, deeply attached to individual actions in a specific context (a profession, a technology…) and composed of individual's technical skills that can only be acquired by learning and improved by experience . Hence, it is difficult to formalize and share.

Proposal

Our contribution is **a new model for *Knowledge Management*** in which we replace the disadvantaged classic *Knowledge Socialization* phase, based on few experts and unidirectional master-apprentice relationships, with a new process characterized by:

- Multidirectional socialization given in social networks, where users adopt multiple roles (they are now content *prosumers*, that is, the production and consumption of knowledge is entrusted to the users (Aranda-Corral et al., 2010)).
- Interchange of ideas through wiki technologies.

This model fits in better with the current technological and social reality.

Thus, we showcase the advantages that our model would provide in a case such as that of Adrià and his colleagues when building the *Bullipedia* from scratch, so that they really benefit from the interaction with a large community of prosumers and use wiki technologies to exchange ideas and develop a more nuanced metalanguage that really supports this initative and make it sustainable in the long term. This approach could also be exploited in other cases of collaboration such as collective software construction, project documentation or interactive learning.

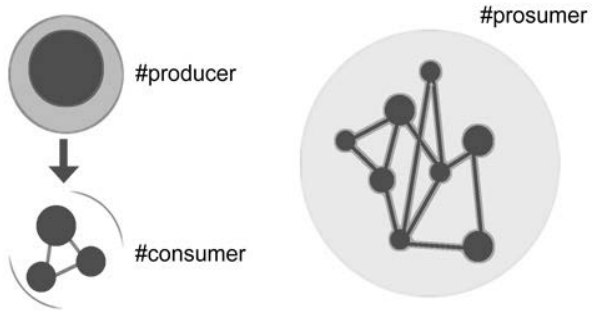


Illustration 2:
Social network based on prosumers

Theoretical framework: the power of crowds, social networks and collaborative technologies Socializing knowledge is critical to both its evolution and its usefulness (Inmon, 2008). *Knowledge Management* cycle might provide better outcomes, in terms of quality and value of knowledge, by improving the socialization process. In the case of *elBulli*, Adrià and his colleagues are agreeing on the content of the upcoming *Bullipedia*. However, they discovered that the hardest part was to find a common language, that is, the worlds of science and gastronomy may share similar processes and methodologies, but they rarely intersect.

It is at this point in which we believe things can be enhanced: turning to a big community instead of a few experts. We borrow some James Surowiecki's concepts on crowdsourcing (Adams, 2011; Schall, 2012; Sautter, 2011; Doan, 2011; Brabham, 2008). Surowiecki asserts that a large group of people is smarter than an elite few, no matter how brilliant (Surowiecki, 2004). Four conditions must be met for a crowd's collective intelligence to produce more accurate outcomes than a small group of experts (Tapscott, 2006):

1) Diversity of opinion
2) Independence of members from one another
3) Decentralization
4) A good method for aggregating opinions

We can demonstrate Surowiecki's ideas are applicable to our model. First of all, the case of *elBulli* is both especially interdisciplinary and heterogeneous enough. Adrià uses *"cuisine as a discourse in order to create a dialogue with other disciplines"*. The foundation is a mash-up of science, the arts, philosophy and technology as a *creativity-generating* universe that produces knowledge (William, 2012). It is this crossroad among various disciplines what

creates such heterogeneity and diversity of knowledge and opinions about the same topic (1).

Secondly, advance in technology have provided novel ways to socialize knowledge. Social networks have created a new reality of social interaction (Easly, 2010; Mika, 2007) that has enabled more effective ways of agreement (Mazzega, 2011). Users organized around a wiki constitute a non-hierarchical and decentralized social network whose members are independent but collaborate (Leuf, 2001) (2 y 3). According to *Complex Systems* theory, members in social networks selforganize and agreement emerges from the bottom-up as a result of their interactions (Wood, 2010; Jones-Rooy, 2010).

Finally, a wiki fosters the idea of *prosumers* collaborating on the Web as it blurs the line between the reader and the writer (Caverly, 2008). Many online communities have adopted this approach to create collective knowledge (John, 2004; Krötzsch, 2006; Ebersbach, 2006). *Wikipedia*, the great online encyclopedia, supports this proposition. Therefore, Wikis are a good method for aggregating knowldge and opinions (4).

Conclusions

We have shown that *elBulli* creates valuable knowledge by applying Nonaka and Takeuchi's principles. The organization is now creating the Bullipedia, an online database that *"will contain every piece of gastronomic knowledge ever gathered"*. Nonaka & Takeuchi's cycle can be adapted to manage that knowledge in social networks based on prosumers. In our model, we propose:

- Turning to the crowd instead of letting the Bullipedia be constructed by a small group of experts.
- Use of wiki technologies, as they have demonstrated their effectiveness for collaboration.

References

Adams, P. (2011). *Grouped: How Small Groups of Friends Are the Key to Influence on the Social Web*. Berkeley, CA, USA: New Riders.

Aranda-Corral, G., J. Borrego-Díaz, and A. Jiménez-Mavillard (2010). Social Ontology Documentation for Knowledge Externalization. *Metadata and Semantic Research: Proceedings of the 4th International Conference, MTSR 2010*. Sánchez-Alonso, S. and Athanasiadis, I. N. (eds). Berlin, Heidelberg, Germany: Springer-Verlag. CCIS 108, 137–148.

Brabham, D. C. (2008). Crowdsourcing as a Model for Problem Solving An Introduction and Cases. *Convergence: The International Journal of Research into New Media Technologies* 14.1, 75–90.

Caverly, D. C., and A. Ward (2008). Techtalk: Wikis and Collaborative Knowledge Construction. *Journal of Developmental Education* 32.2, 36-37.

Davies, J., D. Fensel, and F. Harmelen (eds.) (2003). *Towards the Semantic Web: Ontology- Driven Knowledge Management*. John Wiley & Sons.

Doan, A., R. Ramakrishnan, and A. Y. Halevy (2011). *Crowdsourcing systems on the World-Wide Web*. ACM 54.4, 86-96.

Drucker, P. (1969). *The Age of Discontinuity: Guidelines to our Changing Society*. New York: Harper & Row.

Easly, D., and J. Kleinberg. (2010). *Networks, Crowds, and Markets*. Cambridge: Cambridge University Press.

Ebersbach, A., et al. (2006). *Wiki: Web Collaboration*. Berlin, Heidelberg, Germany: Springer Science+Business Media.

Fensel, D. (2004). *Ontologies: Silver Bullet for Knowledge Management and Electronic Commerce*. 2nd ed. Heidelberg: Springer.

Inmon, W., B. O'Neil, and L. Fryman (2008) *Business Metadata: Capturing Enterprise Knowledge*. Morgan Kaufmann.

John, M. and R. Melster. (2004). Knowledge Networks – Managing Collaborative Knowledge Spaces. *Proceedings of the LSO 2004*. Melnik, G. and Holz, H. (eds). Berlin, Heidelberg, Germany: Springer-Verlag. LNCS 3096, 165–171.

Jones-Rooy, A. and S.E. Page. (2010). The Complexities of Global Systems History. *Journal of The Historical Society* 10, 345–365.

Krötzsch, M., D. Vrandecic, and M. Völkel. (2006). Semantic MediaWiki. *Proceedings of the 5th ISWC 2006*. Berlin, Heidelberg, Germany: Springer-Verlag. LNCS 4273, 935–942.

Leuf, B., and W. Cunningham. (2001). *The Wiki Way: Quick Collaboration on the Web*. Boston, USA: Addison-Wesley.

Mazzega, P., et al. (2011). A Complex-System Approach: Legal Knowledge, Ontology, Information and Networks. *Approaches to Legal Ontologies*. Sartor, G., et al. (eds), Springer Science+Business Media: Law, Governance and Technology Series 1, 117-132.

Mika, P. (2007). *Social Networks and the Semantic Web*. Barcelona, Spain: Springer.

Nonaka, I., and T. Takeuchi (1995). *The Knowledge-Creating Company: How Japanese Companies Create the Dynamics of Innovation*. Oxford University Press.

Sautter, G., and K. Böhm. (2011). High-Throughput Crowdsourcing Mechanisms for Complex Tasks. *Social*

Informatics. Datta, Anwitaman, et al. (eds). Springer: LNCS 6984, 240-254

Schall, D., and F. Skopik (2012). Social network mining of requester communities in crowdsourcing markets. *Social Network Analysis and Mining* 2.4, 329-344.

Soler, J., F. Adrià, and F. Adrià (2007). *Un día en elBulli . elBulli* Books.

Surowiecki, J. (2004). *The Wisdom of Crowds*. New York: Random House.

Tapscott, D., and A. Williams (2006). *Wikinomics: How Mass Collaboration Changes Everything*. New York: Penguin Group.

William, G. (2012). Staying creative. *Wired UK Edition*, 102-111.

Wood, A. T. (2010). Fire, Water, Earth, and Sky: Global Systems History and the Human Prospect. *Journal of The Historical Society* 10, 287–318.

Are Google's linguistic prosthesis biased towards commercially more interesting expressions? A preliminary study on the linguistic effects of autocompletion algorithms

Jobin, Anna
anna.jobin@epfl.ch
EPFL, Switzerland

Kaplan, Frederic
frederic.kaplan@epfl.ch
EPFL, Switzerland

Commodification of words

Displaying relevant ads next to non-paid relevant search results is part of Google's original, highly successful business model (Kaplan, 2011). Advertisers bid on certain keywords they want their ad associated with and pay only if Google displays [1] their ad. Which ads are displayed thus does not solely depend on the query keywords, but also on the bidding price advertisers have been willing to pay for the association of their ads to these keywords, as well as the quality score Google has attributed to their ad (Google, 2012a; Kaplan 2011; Lee 2011, p. 439). Advertising accounts for 97% of Google's revenues, which represented about 3 billion dollars per months in 2012/2011 (Singel 2011). By making advertisers bid on certain keywords to advertisers, Google has commodified words (Lee 2011).

The value of keywords change over time. Keywords are traded within Google's Second Price Auction (Edelman et al. 2007). Google's Diane Tang, creator of the Keyword Pricing Index, illustrates the different trading values of various keywords with the following examples: there are generally "very competitive keywords, like 'flowers' and 'hotels'", whereas other keywords may cost generally little or — such as "snowboarding", more expensive in winter — vary seasonally (Levy 2009). As a result, there exist different bidding strategies for advertisers (cf. D'Avanzo et al. 2011, p. 143-147 for a discussion of the most common strategies). However, Lee (2011) points out that — although online marketing literature about possible and recommended bidding strategies is abundant — "none of the studies found in advertising journals adopted a critical perspective" (p. 438).

Google-ese, Googlais and Googlisch

The fact that words have become a commodity with different monetary values and can be "bought" from Google implies that there is commodified language, which can be represented in a lexicon containing all the words and expressions which are actually "bought". The lexicon of the commodified derivate of the English language is *Google-ese.* [2] (Analogical to the English language, *Googlais* is Google's commodified derivate of the French language, *Googlisch* its German equivalent etc.; other ad-selling search engines with potentially different algorithms are associated with different lexica, leading to e.g. *Bingese*, *Bingais* and *Bingisch* for Bing.) *Google-ese* therefore consists of a certain proportion of the roughly 500"000 dictionary entries (McCrum et al. 2001), some non-dictionary English words like names and places, foreign language expressions and "non-words" such as acronyms, misspelled and mistyped words etc.

Search engine biases

The very existence of *Google-ese*, *Googlais*, *Googlisch* and the like — i.e. specific keywords bought by advertisers

and marketers — account for the company's financial success. Therefore, the “importance of pleasing the advertisers and marketers who support Google and other search engines can hardly be underestimated”, underlines Hinman (2008, p. 70). There is “potential for abuse” (p. 73 ff.), notably in the form of “subtle biases” when it comes to search results and search in general (p. 71). Since commodified language is the baseline for a search engine's revenue, there is reason to explore potential biases when it comes to the treatment of words.

Many scholars from different fields have demonstrated the existence of search engine bias empirically (e.g. Edelman 2011), analytically and (e.g. Diaz 2008) conceptually (e.g. Hinman 2008), overall adopting a critical viewpoint (cf. also Lawrence 2009, Zimmer 2009 p.7, Pariser 2011). Studies addressing search engine bias refer mostly to issues of information access, i.e. search engines as gatekeepers (cf. also Gasser 2006), as well as issues of knowledge shaping (Grimmelman 2009, Hinman 2008). But although that there is general acknowledgement of search engines' impact on access to and classification of knowledge, researchers agree that there has been little to no research focus on search engines' impact on society (Hargittai 2007, p.769; Lewandowski 2012, p.5; Spink and Zimmer 2008, p.344; Zimmer 2009, p. 516-517), let alone search engine's impact on language itself.

Linguistic Prosthesis

Certain functions of Google search are visibly impacting language by transforming the initial search query: “related searches”, for instance, associates our initial keyword to other keywords we might not have thought of; “Did you mean” suggests alternatives to our initial keyword, which was either not correctly spelled or not popular (Google 2012b); auto-completion suggests ways of finishing our initial keywords on our behalf while we are typing according to “purely algorithmic factors (including popularity of search terms)” (Google 2012c). These functions act as a mediation between our thoughts (i.e. the initially intended query) and its expression. We suggest to call them linguistic prosthesis (Kaplan 2011). If search engine biases manifest in linguistic prosthesis, the expression of our thoughts is seamlessly transformed by Google's algorithms.

Modeling of linguistic prosthesis and corresponding commercial value

To progress in the understanding of the effects of these linguistic prosthesis, we have started a systematic and periodic modeling of two important functions. The fist function A(*x) -> {s}* models the association between a string *x* (a partial word or sentence) in the search engine query field and a set of string *{s}* (autocompletion). The second function V(*s*) evaluates the suggested bidding value for a given string *s*. We would like to measure whether the value V(*s*) is influencing the suggested set of strings given by the function A(x). Both functions may, of course, vary over time, and we have to measure A(*x*) and V(*s*) as time-dependent in order to document their relation at a given time t and, potentially, their evolution and possible correlation.

One obvious difficulty is the size of the space we need to monitor and the scale of all measurements: it is nearly impossible to test all the possible x entries at regular intervals over time. However, this scalability issue is very similar to a well-documented problem in the field of optimal experiment design (Fedorov 1972), addressed by artificial intelligence researchers for at least 20 years (Schmidhuber 1991). When a space is too big to explore in its entirety and when, in addition, each trial is costly – which is exactly our situation – one needs to choose smartly what query to test. In the context of their research in open-ended learning systems, Oudeyer and Kaplan have designed a optimal experiment design algorithm that performs precisely this task (details can be found in Oudeyer et al 2007, Kaplan and Oudeyer 2009): instead of trying random configuration, the algorithm detects situations in which its predictions progress maximally, and it then chooses the input signal in order to optimize its own progress. Following this principle, the algorithm running the measurements of the functions A(x) and V(s) avoids "uninteresting" subspaces in order to focus on the actions which are most likely to bring progress. Typically, it will focus its “attention” on subspaces of query strings with significant change in return value as measured by V(s). A daily script thus selects a set Q of n queries each day based on the optimal design algorithm. This produces a set S of results suggestions. For each s of S, we re-test the Value V(s).

Our ongoing experiment, focusing on the commodified lexicon *Google-ese*, derivated from the English language, is being conducted during one year. In that timeframe we hope to elaborate — at least on certain subspaces — a sufficiently good model of the two functions and their evolution over time to test various possible correlations between the two. These models will then be made public in form of a structured corpus, enabling long-time analysis and further studies by other research groups.

This preliminary study will not permit to assess whether or not — and if so: how — Google's linguistic economy is impacting natural languages. It will, however, allow us to make first educated guesses on the linguistic effects of autocompletion algorithms and keyword bidding. In a broader context, this research is an example of how

academia can study technological "black boxes", such as search engines' algorithms, without accessing their inner workings. We believe that by their properties (i.e. enabling us to explore big, costly spaces) optimal experiment design algorithms are of great pertinence for this kind of "reverse engineering" modeling, and such research is likely to become of crucial societal relevance within the coming years.

References

Byrne, J. (2009). "Do You Speak Google-ese?" *Jodybyrne.com,* May 31 2009. http://www.jodybyrne.com/1312. (accessed October 31, 2012)

D'Avanzo, E., T. Kuflik, and A. Elia (2011). "Online Advertising Using Linguistic Knowledge." In **D'Atri, A., Ferrara, M., George, J. F., and Spagnoletti, P.** (eds), *Information Technology and Innovation Trends in Organizations.* 143–150. Physica-Verlag HD.

Diaz, A. (2008). Through the Google Goggles: Sociopolitical Bias in Search Engine Design. In Spink, A. and Zimmer, M. (eds), *Web search multidisciplinary perspectives.* Berlin: Springer.

Edelman, B. H. (2011). "Bias in Search Results?: Diagnosis and Response." *The Indian Journal of Law and Technology* 7: 16–32.

Edelman, M., M. Ostrovsky, and M. Schwarz (2007). "Internet Advertising and the Generalized Second-Price Auction: Selling Billions of Dollars Worth of Keywords." *American Economic Review* 97.1: 242–259.

Fedorov, V. V. (1972). *Theory of Optimal Experiment.* New York: Academic Press.

Gasser, U. (2006). "Regulating Search Engines: Taking Stock and Looking Ahead." *Yale Journal of Law & Technology*: 124–157.

Google. (2012a). "Quality Score — AdWords Help". http://support.google.com/adwords/bin/answer.py?hl=en&answer=2454010. (accessed October 31, 2012)

Google. (2012b). "'Did You Mean' — Web Search Help". http://support.google.com/websearch/bin/answer.py?hl=en&answer=1723. (accessed October 31, 2012)

Google. (2012c). "Autocomplete — Web Search Help". http://support.google.com/websearch/bin/answer.py?hl=en&answer=106230. (accessed October 31, 2012)

Grimmelmann, J. (2009). "The Google Dilemma." *New York Law School Law Review* 53, no. 939.

Hargittai, E. (2007). "The Social, Political, Economic, and Cultural Dimensions of Search Engines: An Introduction." *Journal of Computer-Mediated Communication* 12.3: 769–777.

Hinman, L. M. (2008). "Searching Ethics: The Role of Search Engines in the Construction and Distribution of Knowledge." In Spink, A. and M. Zimmer, (eds). *Web search multidisciplinary perspectives*. Berlin: Springer.

Kaplan, F. (2009). Quand Les Mots Valent De L'or. *Le Monde Diplomatique*, November. http://www.monde-diplomatique.fr/2011/11/KAPLAN/46925. (accessed October 31, 2012)

Lee, M. (2011). "Google Ads and the Blindspot Debate." *Media, Culture & Society* 33(3) (April 1, 2011): 433–447.

Levy, S. (2009)."Secret of Googlenomics: Data-Fueled Recipe Brews Profitability." *WIRED,* 22 2009. http://www.wired.com/culture/culturereviews/magazine/17-06/nep_googlenomics?currentPage=all. (accessed October 31, 2012)

McCrum, R. M., W. Cran, and R. MacNeil (2001). "The Story of English." *Number of Words in the English Language,*. http://hypertextbook.com/facts/2001/JohnnyLing.shtml. (accessed October 31, 2012)

Miller, D. (2012). "Google Voice Search Coming to iOS | Macworld." *Macworld,* August 8, 2012. http://www.macworld.com/article/1168078/google_voice_search_coming_to_ios.html. (last accessed: October 31, 2012)

Oudeyer, P.-Y., and F. Kaplan (2009). "Stable Kernels and Fluid Body Envelopes." *SICE Journal of Control, Measurement, and System Integration* 48.1

Oudeyer, P.-Y., F. Kaplan, and V. V. Hafner (2007). "Intrinsic Motivation Systems for Autonomous Mental Development." *IEEE Transactions on Evolutionary Computation* 11.2: 265–286.

Pariser, E. (2011). *The Filter Bubble. What the Internet Is Hiding from You.* London: Penguin Books Ltd.

Schmidhuber, J. (1991). "Curious Model-building Control Systems." 2:1458–1463. Singapore: IEEE.

Singel, R. (2011). "How Does Google Make the Big Bucks? An Infographic Answer Wired Business Wired.com." *Wired Business,* 19 2011. http://www.wired.com/business/2011/07/google-revenue-sources/. (accessed October 31, 2012)

Spink, A., and M. Zimmer (2008). "Conclusions and Further Research." In Spink, A. and Zimmer, M., *Web search multidisciplinary perspectives*. Berlin: Springer.

Zimmer, M. (2009). "Renvois of the Past, Present and Future: Hyperlinks and the Structuring of Knowledge from the Encyclopédie to Web 2.0." *New Media & Society* 11: 95–113.

Notes

1. Advertisers choose between two options: display or pay per view (PPV — by thousands) or pay per click (PPC). Differentiating the two options is not necessary at this point,

and we therefore use “display” without referring specifically to the PPV option, since PPC also requires the ad to be displayed.
2. Please note that the expression “Google-ese” has previously occurred in different contexts, notably in two different blogposts describing (1) the language entered in Google's search query field (Miller 2012); (2) machine-generated translation text by Google Translate (Byrne 2009) — none of which we are referring to here.

From database to mobile app: scholar-led development of the Heurist platform

Johnson, Ian R.
ian.johnson@sydney.edu.au
University of Sydney, Australia

At a recent meeting, the head of Intersect, the New South Wales eResearch agency, argued that, since academics are not professional software engineers, eResearch software might be prototyped by academics but development should then be turned over to professionals. I believe this IT-centric viewpoint is based on flawed assumptions about the aims of Humanities computing, and that the potential benefits in good design, documentation and QA (assuming IT-managed projects to be superior in this respect) are generally far outweighed by the loss of flexibility which requirements gathering and structured development impose. An engineering approach may be suitable for well-defined deliverables (with commensurate funding), but research-driven and lightly funded projects will generally be far more productive under the control of a Scholar Programmer (Reside 2011, Welsh 2011) who can optimise the outcomes as opportunities present and rely on the goodwill of colleagues for bug identification and testing. Such an approach does not preclude the involvement of professional programmers in the development process (beyond a certain small scale such involvement is essential) nor equal partnership with IT professionals bringing multi-disciplinary perspectives.

I will use as example our development of Heurist, a database abstraction which allows non-technical researchers to rapidly build and incrementally modify — using a (fairly) simple web interface and researcher-centric concepts — complex multi-user web database applications with many entity types and rich relationships. Effective use of Heurist depends not on technical skill but on the decomposition of research data into a clear conceptual structure — an exercise which imposes a valuable sanity check on the quality of a researcher's conceptual model — and then allows them to pass almost directly to a fully functioning database, bypassing the plumbing of implementation (including programmer-centric abstractions such as UML and application frameworks). Over 30 projects are now running in Heurist, from individual research to public resources such as the Dictionary of Sydney (http://dictionaryofsydney.org), and it has also been used in undergraduate classes. Yet Heurist has encountered opposition from software engineers who believe it could be ‘faster’ or ‘better’ if rewritten (in a variety of alternative technologies, mostly post-dating its design), ignoring the fact that it is stable, extensible and fulfils the needs of its users better than any alternative they can propose.

Over the last few years, there have been significant shifts in data modelling towards object-relational models, XML databases, triple stores and so called NoSQL databases. Digital Humanists have embraced these techniques with varying degrees of success. However these technologies lie outside the knowledge or skills of most Humanities researchers — Excel, Access and Filemaker remain the bread and butter of many who venture into the digital domain; more ambitious projects may commission or develop custom applications backended on MySQL or PostGres. Some of these (eg. Kora) evolve into capable adaptable systems, or are conceived as such from the outset (eg. Zotero and Omeka).

While Heurist uses MySQL as its backend, it uses it to build an agnostic datastore along NoSQL principles. The majority of the code is then devoted to managing a user view of the structure based on entities, attributes and relationships which requires no understanding of the underlying methods. The structure is itself stored within the database, allowing on-the-fly modifications and providing a fully internally documented database (which can be easily — one click — exported to a fully documented XML dump of both the database structure and its content for archiving or transfer to another system). A key strength of this approach is that database structure can be developed incrementally and modified throughout the life of a project as the researcher's understanding of the domain (and software) evolves, without the need for programmers or the loss of existing data.

In this paper I will review the design principles underpinning Heurist and show how its flexible and user-modifiable approach to data structure allows the database to evolve with a scholar's needs. Heurist development has been research-driven and informal, responding to the needs of the many projects that use it, while developing new capabilities as generic tools rather than project-specific additions. I will particularly focus on the advantages of a pragmatic approach to development, driven by a Scholar-

Programmer and based on evolving user needs, rapid prototyping and 'permanent beta'. I will contrast this with an 'engineering' approach based on pinning down a set of user requirements and rigorous development and testing within a framework which — even if notionally using an Agile program development methodology — discourages revision and innovation in favour of QA.

My argument will be supported by looking specifically at two mobile applications which were developed as core functions of Heurist. These applications were never envisaged in the original design of the system, but have been added with minor effort, extending the functionality of the system for all users.

First, as part of research into community engagement for the Dictionary of Sydney (dictionaryofsydney.org) we developed a mobile heritage tour application which runs off the Dictionary's underlying Heurist database. The application adopts TourML (TAP Into Museums 2012) but unlike standalone tour applications it is simply designed as a viewer with offline data caching, and can therefore run off any Heurist database containing appropriate data. No additional database design or programming was required to support data entry, storage of TourML data or the import of schemas or data, since these are already an intrinsic part of the Heurist model. With the addition of an appropriate XSL template to transform Heurist's native XML output, we also now have a web-based tour view which will work across all Heurist databases.

My second example is the handling of data schemas for the recently funded FAIMS (Federated Archaeological Information Management System) research infrastructure project funded by the Australian government's NeCTAR program. With trivial programming we were able to add a function to Heurist which exports record schemas as W3C standard XForms which can be loaded onto tablet devices for use in the field. Data is collected using these forms with Open Data Kit (Open Data Kit 2012) and synced back into the Heurist database from which the forms were exported. The field application is thus simply an extension of the Heurist database, rather than a separate application, allowing any Heurist database to become 'mobile' on a tablet — archive and library information gathering are an obvious spin-off acquired at no extra cost. This paradigm is further extended by exposing the database on the web, allowing any other Heurist database to import and reuse the schemas. Using this capability we were able to set up a schema clearinghouse for the FAIMS project and populate it with an initial set of schemas in a matter of hours. The other major components of the FAIMS project are following a formal engineering methodology and will provide useful comparative material by the time this paper is presented.

Through this paper I hope to stimulate a discussion in two areas:

First, on the appropriateness of the concept of the Scholar Programmer, picking up on recent debates on the need for Digital Humanists to embrace programming skills. I will propose an alternative concept of Scholar Analyst — a scholar who might or might not have programming skills, but mostly brings conceptual skills informed by a knowledge of what can be achieved (without necessarily being technically capable, even if time permitted). In that sense the Scholar Analyst is an Architect, who understands what different materials can achieve without being able to so much as lay a brick. I am not convinced that the Architect needs to be able to lay bricks, although they must know the limits of brick structures and of the bricklaying process.

Secondly I hope to stimulate some discussion about the dangers of development for, rather than by, the Academy. While this may not be an issue of concern to current projects with momentum, the centralisation of resources in relatively well-funded and science-dominated eResearch agencies and IT centres plays to the dominance of engineering-driven approaches and loss of capacity to develop applications aligned with Humanities research needs.

References

Date, C. J. (2004). *An Introduction to Database Systems*. Boston: Pearson/Addison Wesley. 8th edn.

Open Data Kit (2012). *Magnifying human resources through technology*. http://opendatakit.org/ Accessed 30/10/12

Reside, D. (2011). *On Ant-Lions and Scholar-Programmers*. http://mediacommons.futureofthebook.org/alt-ac/pieces/ant-lions-and-scholar-programmers Accessed 30/10/2012

TAP Into Museums (2012). *TourML Overview*. http://tapintomuseums.org/TourML Accessed 30/10/12

Welsh, T. (2011). *To Code, or Not to Code*. http://hastac.org/blogs/twelsh/code-or-not-code . Accessed 30/10/2012

The Network is Everting: the Death of Cyberspace and the Emergence of the Digital Humanities

Jones, Steven Edward
sjones1@luc.edu
Loyola University Chicago, United States of America

This paper (with slides) is drawn from a book in progress (under contract at Routledge) about an important cultural context for the digital humanities. It argues that the emergence of DH in its newly prominent forms during the past decade is closely connected to a larger shift in the collective imagination of the network. This context helps us understand the public role that DH might play in exploring multiple connections between digital and physical materialities out in the world.

"Cyberspace is everting," as William Gibson has said, turning inside out and leaking out into the physical world (2007; 2001). Cyberspace was always a notional nonspace, a consensual hallucination. a metaphor for our relationship to the global network. But it was a powerful metaphor that made a material difference in both technology and culture. In 2007, Gibson overwrote his own metaphor in *Spook Country*, thirty years after he coined the term cyberspace (1984). As he now notes, the term cyberspace is fading from use as the network *everts*, turns itself inside out. More than a literary metaphor, this eversion of cyberspace marks a profound shift in our collective understanding of the digital network—from an online world apart to a pervasive part of the world, from a transcendent virtual reality to a ubiquitous grid of data we move through every day.

We can even roughly date the shift to 2005-2006, when a number of important developments occurred: the quintessential virtual world Second Life began to decline; Nintendo's motion-control Wii was introduced and helped to usher in the era of casual gaming; so-called "Web 2.0" social networks and mobile platforms, first introduced in 2004-2005, came into their own and gained a mass user base; the Google Maps API was launched in 2005; the iPhone was previewed in 2006 and introduced in January 2007; again, Gibson's *Spook Country* was published early in 2007, its story based on augmented reality and locative art and media. More recently, this same shift in the collective imagination has been tracked by proponents of the so-called New Aesthetic (2012), which notes the increasing appearances of glitches, pixel art, and so on, as signs of "the irruption of the digital into the physical." These and other emerging expressions are part of a larger cultural change whose effects we are still experiencing, a multi-platform shift in the nature of our relation to technology, corresponding to what N. Katherine Hayles has called a fourth phase in the history of cybernetics, a phase of "mixed reality" in terms of cultural perception and material technology (148).

At about that same historical moment, the digital humanities gained public attention, emerging out of the longer tradition of humanities computing, marked by the rise of the term "digital humanities"—which was coined in 2001 but reached public consciousness and institutional weight between 2004-2007—with an emphasis on data analysis, distant reading, the maker movement, and the spatial turn, as well as a sometimes unacknowledged debt to video games and game theory.

These juxtaposed events have nothing to do with an argument for technological determinism. They're just meant to suggest that the emergence of the new digital humanities isn't an isolated academic phenomenon. The institutional and disciplinary changes are part of a larger cultural shift, a rapid cycle of emergence an convergence in technology and culture. The new DH is both a response to and a contributing cause of the ongoing eversion. Seen in context, the newer forms of supposedly practical or instrumental DH were produced in the first place by younger scholars working with a keen awareness of the developments I'm grouping under the concept of the eversion, and a sense of what these meant at the time for various technology platforms of interest to humanists (Gold). In the era of social networks, casual gaming, distributed cognition, augmented reality, the internet of things, the geospatial turn, one segment of new digital humanities work took a hands-on, practical turn, yes ("more hack, less yack"), but arguably based on theoretical insight, as a kind of deliberate rhetorical gesture—a dialectical counter-move to the still-prevailing idealisms associated with the cyberculture studies of the 1990s. Much of the practical DH work during the decade that followed was undertaken not in avoidance of theory or in pursuit of scientistic instrumentalism, but *against disembodiment*, against the ideology of cyberspace. The new DH was deliberately figured as digital not just *digitized*, moving from a perceived separation between the stuff of the humanities—books, manuscripts, artifacts, works of art—and the digital medium, to more of a mixed-reality model, characterized by two-way transactions between the two realms, crossing multiple materialities (Kirschenbaum; Svensson; Gold).

This kind of mixed-reality reciprocal interaction between data and artifacts, algorithm and world, has been effectively modeled for decades in video games. My paper will triangulate (1) the eversion of the network and the (2) rise of the new digital humanities by way of (3) the increasing role of video games in the contemporary media landscape. If possible, I'll demonstrate the games live as part of the slide presentation, as examples of interdimensional transit and play across the boundaries of different materialities, citing Fez, an independent platformer with a toggling 2D/3D game world that includes QR codes incorporated in it; Wii U games and 3DS AR Games—which experiment with augmented and mixed reality in console gaming; and Skylanders, with collectible fantasy figurines, prototyped using 3D printers, containing RFID and NFC chips that create a "Hertzian field" (Dunne) within which to translate characters into and back out of the game world. As part of establishing context, I'll also cite recent fiction about

the interdimensional relation of games and the world by Neal Stephenson, Ernest Cline, and China Mieville. In conclusion I'll suggest that games offer ways to think about the active role of DH in the ongoing eversion of the network, and about the engagement of humanities research with developments in the wider world.

References

Cline, E. (2011). *Ready Player One*. New York: Crown Publishers.

Dunne, A. (2005). *Hertzian Tales: Electronic Products, Aesthetic Experience, and Critical Design*. Cambridge, MA: MIT Press; rept. 2008.

Farman, J. (2011). *Mobile Interface Theory*. New York: Routledge.

Gibson, W. (2011). Interviewed by David Wallace-Wells in *The Paris Review*, 197 (Summer 2011), 107-149.

Gibson, W. *Neuromancer*. New York: Ace Books, 1984.

Gibson, W. *Spook Country*. New York: Putnam, 2007.

Gold, M. K. (ed). (2012). *Debates in the The Digital Humanities*. Minneapolis: University of Minnesota Press.

Gordon, E. and A. de Souza-Silva (2011). *Net Locality: Why Location Matters in a Networked World*. Boston: Wiley-Blackwell.

Greenfield, A. (2006). *Everyware: The Dawning Age of Ubiquitous Computing*. Berkeley, CA: New Riders.

Hayles, N. K. (2010). Cybernetics. In *Critical Terms for Media Studies*. Mitchell, W. J. T., and M. B. N. Hansen (eds.), Chicago: University of Chicago Press, 145-56.

Kirschenbaum, M. (2008). *Mechanisms: New Media and the Forensic Imagination*. Cambridge, MA: MIT Press.

Mieville, C. (2009). *The City and the City*. New York: Del Ray.

Sterling, B. (2005). *Shaping Things*. Cambridge, MA: MIT Press.

Stephenson, N. (2011). *Reamde*. New York: William Morrow.

Svensson, P. (2009). Humanities Computing as Digital Humanities. *DHQ* 3.3 http://www.digitalhumanities.org/dhq/vol/3/3/000065/000065.html .

A Clear Temporal GIS Viewer and Software for Discovering Irregularities in Historical GIS

Kantabutra, Vitit
vkantabu@computer.org
Idaho State University, United States of America

Owens, J.B.
jowens@computer.org
Idaho State University, United States of America

In order to use geographic information systems (GIS) to understand any aspect of human systems, researchers and teachers require a GIS capable of dealing with temporal information. Recognition of this need among historians, the humanities disciplines, and the historical social scientists extends back at least as far as the initial development of TimeMap (not maintained after 2007) in conjunction with the Electronic Cultural Atlas Initiative (ECAI) in the late 1990s. The project presented in this abstract represents a portion of a larger effort to improve spatio-temporal GIS by developing a database management system, Intentionally-Linked Entities (ILE), consisting of a data structure of direct linking via pointers, rather than storing data in tables, as in a relational database management system (RDBMS). We have taken advantage of linking information entities/objects directly to other entities/objects to create a better exploratory GIS viewer for those seeking to evaluate a large historical database.

When fully implemented the fully-linked data model such as ILE offers many advantages for spatio-temporal GIS. The model permits relationships with unlimited *arity* (number of roles) that are represented by direct links. Currently the RDBMS permits unlimited arity but relationships are not represented by direct links, whereas network/graph databases (such as the increasingly popular Neo4J (Anonymous, 2013) represent relationships with direct links but are limited in arity to 2. Additionally, network/graph databases have no means of representing entity sets, making them unsuitable for ontological inferences. High arity, direct representation of arity sets, and direct linking are all important attributes for representing and analyzing complex systems often found in the Humanities.

However, for purposes of designing a better temporal GIS exploratory viewer, the major technical innovation relates to quick, effortless searches through large amounts of data. The current standard GIS represent temporal data by means of time-stamping within an RDBMS, leading to much redundant data (Yuan and Hornsby, 2008). The RDBMS is inefficient for temporal applications because all queries require a database search, which is inefficient even with indices because temporal applications require frequent searches as the user moves from one time period to another. We have employed a direct-linking data structure

with pointers so that NO searches are required during the display of data. No special work-around in data modeling or indexing is necessary to enhance data exploration.

We used CHGIS (China Historical GIS) V. 4 (current version) time-series database as our test bed. This database is a good test bed because of the amount of data involved, especially when all seven input files are used at once. Moreover, CHGIS is the most accessible and logically designed of all current historical GIS (HGIS) databases available. Bol (Bol, 2007), director of Harvard's Center for Geographic Analysis and CHGIS founder, wrote that the CHGIS database can be used as an authoritative historical database. Therefore it should be of great interest to be able to find and hopefully correct the errors or note the irregularities. CHGIS' metadata indicates that the database is established on the datum **North American Datum of 1927** (NAD-27), and we have used the **Xian 1980, Transverse Mercator** projection, which is the correct projection, for examining the relative positions of the provincial capitals and their boundaries.

We explored some of the available options for viewing temporal GIS data by starting with Google Earth. Berman (Berman, 1999), who engineered the pioneering data model and software for CHGIS, wrote, "Now we can browse through hundreds, or thousands of years of Historical GIS objects .... Interestingly, the temporal browsing functionality is only possible using a free software application, Google Earth ..., but cannot be done with any of the major commercial GIS packages." We found Google Earth to be awkward and confusing. When used with the two time-series provincial data files of CHGIS, the display was often too messy to comprehend precisely. The labeling is very poor, with the names of the capitals showing only in Chinese characters unless we place the mouse over the placemark. Surprisingly, even though the time resolution of the CHGIS time series database is one year, Google Earth only seems to be able to display data sometimes every seven years (such as 373, 380, 387,...) and sometimes every eight years (there is a gap between 1490 and 1498). There are temporal points in the database itself that lie on years that Google Earth apparently cannot display. We attempted to set software preferences and to look in the database files, in vain, to see if there are instructions telling Google Earth to skip or not to skip years. We can only guess that the Google Earth designers skipped years in order to reduce the data handling load. Finally, when asked to animate the map by running through the years, Google Earth animates the map at a speed that is far too rapid for the eyes to comprehend, even at the slowest speed setting [1]. This is not too surprising since Google Earth was probably not created specifically for Historical GIS, and also because of the year skipping mentioned earlier.

We also tested ESRI ArcGIS Explorer on the time-series CHGIS data. Explorer behaves very strangely, however. When sliding the time slider to the right, the time did not always progress in the forward direction! For instance, starting from year 778, the progression in years for the ArcGIS Explorer was 778 - 780 - 786 - 784 - 786 (That's right, another 786!) - 864 - 804 - followed by a crash, which is something that occurs often, even on a powerful PC.

After these unsatisfactory experiences, the first author wrote his own HGIS viewer, Clearview, which displays the provincial boundaries and capital locations clearly. Clearview is also capable of displaying the data at every temporal point for which there is data. Additionally, Clearview's automated display has a widely adjustable speed that allows slow enough settings for viewing comfortably by a user. Clearview also mines the html data in the place marks to display Pinyin names, and it can be modified to display whatever the user wants.

Due to the clarity of Clearview's display, the user can clearly visualize the relative locations of provincial capitals relative to their boundaries. As soon as the user animates Clearview through time, it is immediately obvious that there are many instances of provincial capitals that are outside the respective provincial boundaries. Subsequently the first author wrote another piece of software, which goes through the entire CHGIS time-series provincial dataset and computes how many instances there are where a provincial capital lies outside the same province's boundary. This software found the following: 124 instances of the provincial capitals lying outside the same province's boundary, an irregularity (perhaps an error since normally a capital lies inside (or on) the boundary) compared to 420 instances in which the provincial capitals lies inside the same province's boundary. (An instance is a province together with a maximal period during which there are no changes in that province.) There are also three instances in which the provincial capital exists but there is no polygon for the corresponding province, which is not an error. There are also seven curious irregular instances where a province has two capitals the same year. In each case these two locations appear to be so widely separated, sometimes in different polygons. These may not errors since some are military governorships not provinces, but this fact should be in the database.

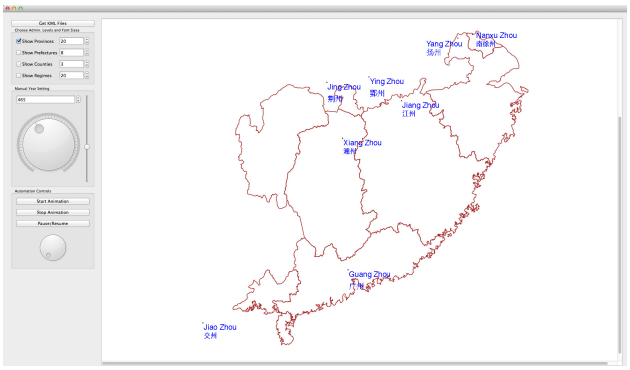
Getting Clearview to work on large amounts of data requires technical innovation. One way to achieve direct linkage from each time step in the database to all the objects to be displayed at that time step was covered in Kantabutra, et al. (2010). However, with the large amount of data in CHGIS, the number of links can be very large. Suppose a geographical object is valid for 1000 years, then the pointer scheme in Kantabutra, et al. requires 1000 links from the time objects corresponding to those 1000 years to the graphical object. For this viewer, we ameliorate the situation by using a data structure call the Tree of Time. This structure is similar to the Time Tree used for

multiresolution video (Finkelstein, et al., 1996), but used differently.

The Tree of Time is a binary tree structure (Cormen, et al, 2007). The Tree of Time is indeed a search tree, but the only searching that occurs is done during the tree construction. When Clearview is running, no searching takes place because each time-object already has pointers to the tree nodes, whose time intervals contain the time represented by the time-object. Because CHGIS intervals are one year, that is the setting for this demonstration, but the Tree of Time can be used for any time resolution in a data set.

The Tree of Time is used internally and the user doesn't have to deal with its complexity. The software for discovering data errors or irregularities also employs complex data structures in order to ensure correct and efficient enumeration of such situations.

A video about this viewer can be viewed at http://youtu.be/uZrn-hMlZXo



References

Berman, M. L. (2009). Modeling and Visualizing Historical GIS Data. In *Spatio-Temporal Workshop*. held April 2009 at Harvard University.

Bol, P. K. (2007). The China Historical Geographic Information System (CHGIS): Choices Faced, Lessons Learned. In *Conference on Historical Maps and GIS*. held August 23-24, 2007 at Nagoya University.

Cormen, T., et al. (2007). *Introduction to Algorithms*. 3rd edn. Cambridge, MA: MIT Press.

Finkelstein, A., C. E. Jacobs, and D. H. Salesin (1996). Multiresolution video. In *Proc. 23rd annual Conference on Computer Graphics and Interactive Techniques. SIGGRAPH '96.*, New York: ACM. 281-290.

Kantabutra, V., J. B. Owens, D. P. Ames, C. N. Burns, and B. Stephenson (2010). Using the Newly-created ILE DBMS to Better Represent Temporal and Historical GIS Data. In *Transactions in GIS*, **14 (1)** 39-58.

Anonymous (2013). *Neo4J, a graph database*, http://www.neo4j.org/ , modified 8 March 2013.

Yuan, M., and K. S. Hornsby (2008). *Computation and Visualization for Understanding Dynamics in Geographic Domains*. CRC Press.

Designing a graduate DH course with DH tools and methods

Kee, Kevin Bradley

kkee@brocku.ca
Brock University, St. Catharines, Ontario, Canada

Roberts, Spencer

srober26@masonlive.gmu.edu
George Mason University, Fairfax, Virginia, U.S.A.

This paper describes an experimental approach to designing and teaching an introductory digital humanities course for graduate students. In 2011 Kevin Kee was asked to create and teach a class as part of a new interdisciplinary Humanities Ph.D. program. The graduate students taking the course would be largely unfamiliar with the digital humanities.

Kee began his preparation for the course by asking several questions. The first was "what should an introductory digital humanities course attempt to accomplish"? As he searched for answers, another important question emerged: "How could digital methods be used to design and deliver the course?" In this paper, Kee and Spencer Roberts, a Research Assistant who worked with Kee, describe first their method for researching and designing the course. They then sketch the structure and content of the course that resulted from their research.

Finally, they provide examples of student responses to the material and methods covered in the course (including Roberts's perspective as a graduate student). The collected responses and their reflections on the process suggest particular ways in which future courses of this kind might be designed, implemented, and improved. Most importantly, they found that an effective way to design and teach an introductory digital humanities course is to *think about* the discipline through discussions about its topics and to *think with* the discipline by using digital tools and methods in the classroom.

Overviews of digital humanities course offerings have been conducted throughout the past fifteen years. In 1999, Willard McCarty and Matthew Kirschenbaum identified only fourteen institutions that offered courses in humanities computing. In 2006, Melissa Terras conducted another survey of digital humanities curriculum, and in 2011, Lisa Spiro undertook to collect and analyze syllabi from digital humanities courses. Of these previous surveys, Spiro's was the most comprehensive; she collected over 134 syllabi from various levels of study in the digital humanities. Although Spiro's work parallels and was helpful to that of Kee and Roberts, the latter were unaware of her project when they began, and had no way to replicate her research method. As a result, Kee and Roberts drew on the results of their own analyses while designing the course.

For Kee and Roberts' research, Roberts designed a method by which syllabi were converted into sets of data representing reading lists, assignments, assessment methods, and digital tools used. Commonly occurring items from within those sets were highlighted and identified as items deemed important by the statistical consensus of instructors represented in their sample. For example, their results showed that seven authors of digital humanities-related articles and books appeared on reading lists at a significantly higher frequency than others; the data also showed that other instructors found these authors most useful. Kee drew on these results when deciding on readings for his course list.

Topics covered in the course included text encoding and markup, distant reading, building, mapping, modelling and simulating, playing and gaming, teaching, and collaborating. Each of these was paired with a practical application, usually drawn from a modified version of William Turkel's "Method". For example, students learned the theory of text markup and were asked to create pages on the course wiki using the basic wiki standard markup. Franco Moretti's theory of distant reading made more sense for students once they experimented with text mining and analysis tools such as Voyant. Kee's assessment strategy required students to use blogs and Twitter to comment on the theories and tools they encountered; Kee also encouraged them to participate in scholarly discourse that occurs on the Web. Although most of the students studied history, Kee aimed to create an environment in which the digital humanities were understood as both theory and practice that could be incorporated into any humanities discipline.

Because Kee and Roberts hoped to learn from the experiences of graduate students new to the digital humanities, Kee built feedback mechanisms into the course assignments, and asked students to reflect on the course before and after completion. Nearly all of the students were challenged by the dual responsibility of learning theory and skills simultaneously. Although some students were relieved to finish experimenting, others were pleased with their progress and the opportunities for future research. One student commented, "Not only do I now have some new tools to use while I'm doing research… I'm also more open-minded towards using them in the first place and really trying to engage with them, rather than brushing them off." While students readily adopted some of the tools, such as Zotero and Evernote, they found more complex tools such as DevonThink or Voyant required a level of commitment and time they did not want to make. In short, these students were not willing to commit to a new, digital research method at a time when they were simultaneously taking graduate courses rooted in conventional research methods. For some students, however, patience led to late or accidental discoveries that improved their methods; in at least one case, a student who was skeptical throughout the course became an enthusiastic supporter of digital methods and now avidly attends DH conferences and events. At the conclusion of the course, most students were open to the various theories and approaches used in the digital humanities, and were enthusiastic about trying new tools and experimenting with new methods that might improve their research and scholarship.

From the outset of the project, Kee and Roberts understood that they were asking questions for which there were several feasible answers. Some graduate level digital humanities courses focus on topics within the digital humanities; others primarily train students to develop digital skills using computational tools. Kee's approach was to combine these two approaches into one course that provided opportunity for theoretical discussion while also showcasing practical applications, so that students could see the potential beneDits of digital humanities methods without having to master sophisticated tools. The research method used to build the course syllabus employed the same theories and tools that were later discussed in the course, creating an iterative loop through which student feedback and developments in the discipline can be incorporated into future versions. Already there are new tools to improve the collection and analysis of digital humanities syllabi, and new methods being explored by instructors. Through the experimental approach described in this paper, Kee

and Roberts have found that *thinking about* and *thinking with* the discipline, a method that many digital scholars employ in their research, is also an effective way to design a course, and appeals to students who are new to the discipline, fostering enthusiasm for its use in their own often conventional humanities scholarship. The authors hope that this approach contributes to the growing conversation about teaching digital humanities, while also reflecting and adapting to the dynamic topics within the field.

Stylometry and the Complex Authorship in Hildegard of Bingen's Oeuvre

Kestemont, Mike
mike.kestemont@ua.ac.be
University of Antwerp, Belgium

Moens, Sara
sara.moens@ugent.be
Ghent University, Belgium

Deploige, Jeroen
jeroen.deploige@ugent.be
Ghent University, Belgium

Hildegard of Bingen (1098–1179) is one of the most influential female authors of the Middle Ages (Newman, 1998). The numerous texts attributed to this prophetess are versatile, with topics ranging from her visionary works, over medical issues, to music. Recent decades have seen an increase in the scholarly interest in Hildegard's works and persona. From the point of view of computational stylistics, the oeuvre attributed to Hildegard is fascinating, because of the exceptionally complex authorship underlying it. Hildegard dictated her texts to secretaries in Latin, a language of which she did not master all grammatical subtleties (Ferrante, 1998; Deploige, 1998). She therefore allowed her scribes to correct her spelling and grammar. A number of manuscripts survive, produced under Hildegard's supervision. Fig. 1 shows a detail of fol. 77 in the most important manuscript of Hildegard's *Liber operum divinorum* (Ghent, University Library, MS 241) with a sample of the numerous adjustments made by Hildegard's scribes. From these it is evident that multiple scribes have polished Hildegard's original texts (Derolez, 1972; Derolez & Dronke, 1996). Many of Hildegard's assistants are known by name, including Volmar, Ludwig of Saint Eucharius and Godfrey of Disibodenberg (Ferrante, 1998).

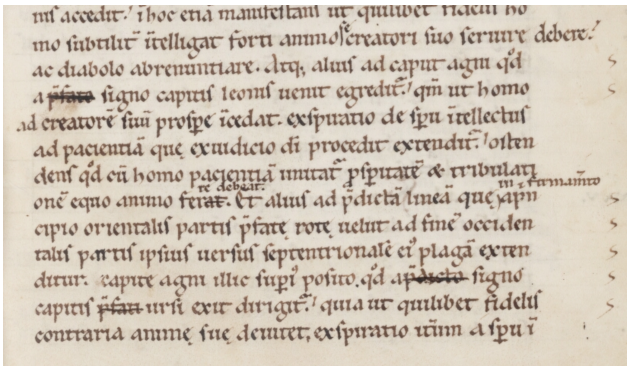


Fig. 1

The extent to which these copy-editors have interfered with Hildegard's style has been the subject of intense scholarly debate. Especially Hildegard's last collaborator, Guibert of Gembloux (Moens, 2010), seems to have heavily reworked her works during his secretaryship. Whereas her other scribes were only allowed to make superficial linguistic changes, Hildegard would have permitted Guibert to render her language stylistically more elegant. This is at least what can be deduced from the *Visio ad Guibertum missa*. Deploige and Moens are currently finalizing an edition (fortcoming in Brepols's *Corpus Christianorum*) of the *Visio ad Guibertum missa* (MISSA) and the lesser known *Visio de sancto Martino* (MART), both of which Hildegard allegedly authored during Guibert's secretaryship. Guibert's interventions in these texts nevertheless seem so far-reaching that one could wonder to what extent they are still attributable to Hildegard (Newman, 1987).

In this paper we will focus on Guibert's role. At the same time, this case study carries wider relevance for stylometry, especially because medieval Latin prose has been rarely studied in the stylometric community. Stylometrists typically focus on high-frequency function words: as opposed to content words, these function words only carry a bleak, grammatical meaning (Pennebaker, 2011). In the case of Hildegard, it were exactly these grammatical words that were often corrected by her scribes. It would be interesting to assess whether Hildegard's oeuvre is still a stylistically coherent corpus that is distinguishable from that of contemporary authors or from the stylistically more ornate body of works which she published during Guibert's secretaryship.

We researched a corpus provided by Brepols's *Corpus Christianorum Library & Knowledge Centre*, including works by Hildegard, Guibert and Bernard of Clairvaux. Medieval Latin, like other historical languages,

is characterized by spelling variation. We therefore had to normalize the orthography in the corpus using lemmatization. We have trained the *Morfette* lemmatizer (Chrupala et al. 2008) on the annotated *Index Thomisticus Treebank* (Passarotti & Dell'Orletta 2010). All experiments described below were carried out on the lemmatized corpus, some using the script suite 'Stylometry for R' (Eder & Rybicki, 2011). All content words and personal pronouns were manually "culled" from the set of most frequent lemmas. All analyses were thus restricted to a set of 65 lemmas of high-frequency function words.

A first analysis (Fig. 2) was performed on the *epistolaria* in the corpus, using *Principal Components Analysis* (PCA; Kestemont 2012). This analysis excludes the letters by Hildegard which were probably reworked by Guibert. The resulting scatterplot yields an excellent authorial separation (samples of 10,000 lemmas), with Guibert's samples (GUI_EP) clustering in the top-right corner, Hildegard's in the left part (HILD_EPNG) and Bernard's in the lower-right area (BERN_EP). The component loadings (lightgrey) draw attention to Hildegard's extravagant use of the preposition *in* ('in') or Guibert's exorbitant use of *et* ('and'). The latter remark is in line with earlier observations by philologists, noting Guibert's clear preference of *et* to the alternative conjunctions *ac* or enclitic *–que*, esp. in compound sentences (Derolez, 1972). To test how long samples from these authors should be for a correct attribution, we carried out a leave-one-out validation using Burrows's Delta for different samples sizes (Fig. 3). Samples sizes of 2,000 lemmas or more generally lead to reliable attributions (> 95% accuracy).

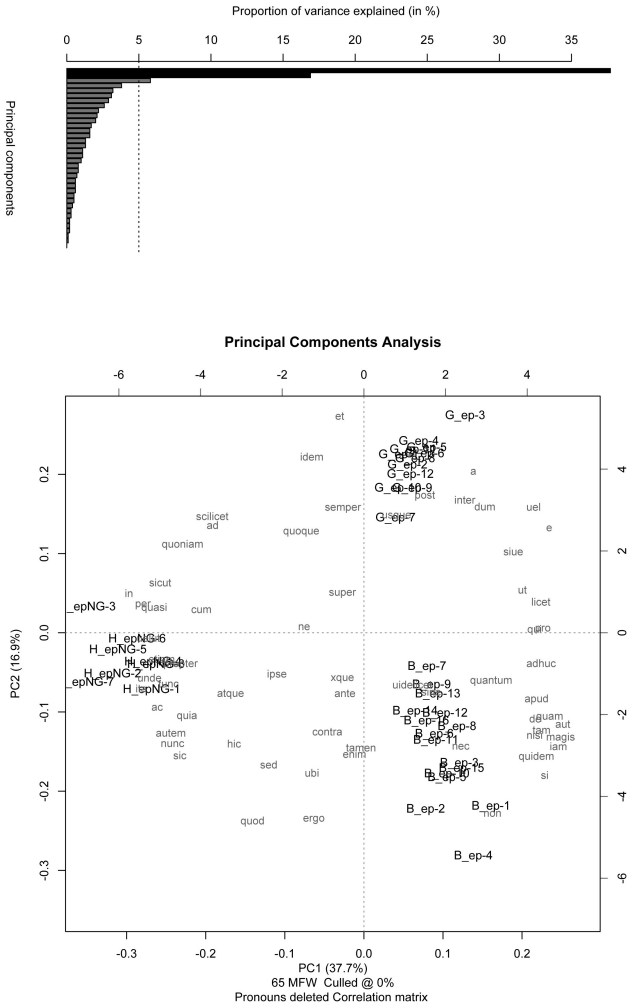


Fig.2

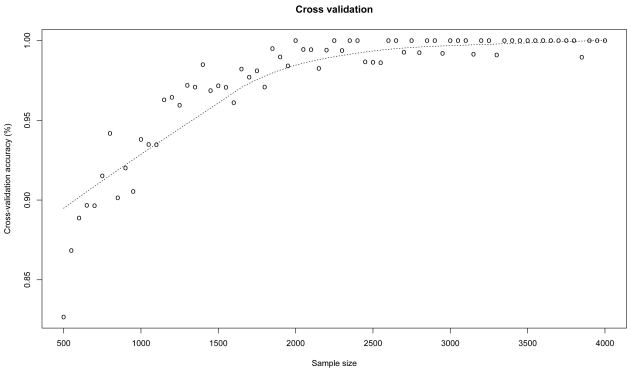


Fig. 3

In Fig. 4 we have focused on the differences in writing style between Hildegard and Guibert. In this PCA, we have included Guibert's letters (GUI_EP), Hildegard's letters during Guibert's secretaryship (HILD_EPG), as well as Hildegard's letters unrelated to Guibert (HILD_EPNG). There is a firm horizontal separation between both

authors. Apart from the *et-in* opposition, we also see a marked contrast with respect to the use of *e* ('out') or *cum* ('with'). The samples from letters that result from their "collaboration" (HILD_EPG) cluster in the lower-left corner: they appear to be on Hildegard's side of the stylistic spectrum, but interestingly, their position is rather isolated from the other Hildegard samples. This result is reminiscent of the Synergy Hypothesis, suggesting that texts resulting from collaboration can display a style markedly different from that of the (average of the) collaborating authors (Pennebaker 2011).

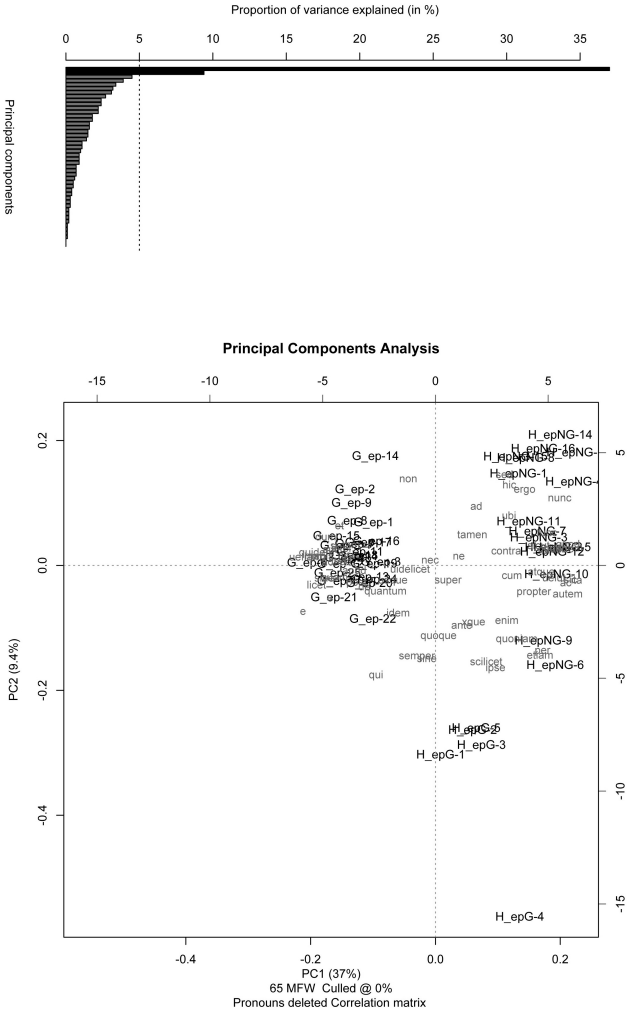


Fig. 4

These results demonstrate the general validity of the stylometric approach for the present corpus. It would therefore be interesting to assess how stylometric methods would react to the authorship in the two texts which were our point of departure. Fig. 5 shows the results of a PCA that confronts the *epistolaria* (Figs. 2 & 3) with the text pair of dubious authorial signature (MART and MISSA, prefixed "D(UBIOUS)".

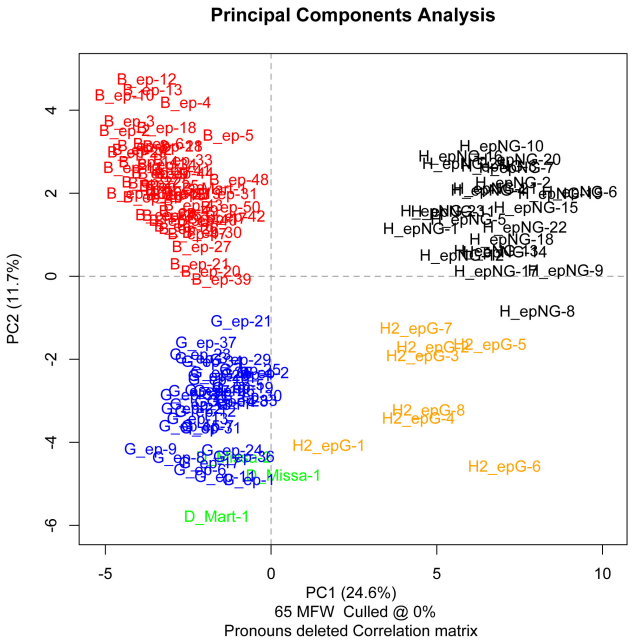


Fig. 5

Fig. 5 offers a clear-cut confirmation of the previously voiced doubts concerning Hildegard's authorship for MART and MISSA (sample size=3,304). If indeed anything of Hildegard's authorial style was ever present in these texts, Guibert seems to have reworked them to such an extent that style-oriented computational procedures are far more inclined to attribute the texts to Guibert than to Hildegard. This result thus yields a quantitative affirmation of the opinion asserted in previous Hildegard scholarship about Guibert's stylistic reworkings of Hildegard's oeuvre (Klaes, 1993; Newman, 1987), as well as fresh evidence regarding the Synergy hypothesis with respect to collaborative authorship in general (cf. the "co-authored" H2_EP samples in Figs. 4 & 5).

The results reported in this paper obviously only scratch the tip of the iceberg in Hildegard scholarship. We nevertheless hope to have illustrated the huge potential of stylometric methods in dealing with medieval Latin text collections and similar historical corpora. Quantitative techniques can be used as a refreshing means to falsify or strengthen hypotheses formulated in traditional scholarship (e.g. Guibert's stylistic influence). Future research will have to consider in what other respects stylometry could advance the traditional Hildegard scholarship, e.g. by isolating the more subtle linguistic influence of her other scribes on her works.

References

Burrows, J. (2002). 'Delta': a measure of stylistic difference and a guide to likely authorship, *Literary and Linguistic Computing*, **17 (3)**. 267–87.

Chrupala, G. et al. (2008). 'Learning Morphology with Morfette'. in *Proceedings of LREC 2008*. held in Marrakech, Morocco. 2362–7.

Deploige, J. (1998). *In nomine femineo indocta. Kennisprofiel en ideologie van Hildegard van Bingen (1098-1179)*. Hilversum: Verloren.

Derolez, A. (1972). The genesis of Hildegard of Bingen's *Liber divinorum operum*. The codicological evidence. In Gumbert, J. P. et al. (eds), *Litterae Textuales. Essays Presented to Gerard I. Lieftinck. II: Texts & Manuscripts*. Amsterdam: Van Ghent. 23–33.

Derolez, A. and P. Dronke (eds.)(1996). *Hildegardis Bingensis Liber Divinorum Operum*. Turnhout: Brepols.

Eder, M. and J. Rybicki (2011). Stylometry with R. In *Digital Humanities 2011. Conference Abstracts*. Stanford. 308–11.

Ferrante, J. (1998), *Scribe quae vides et audis*. Hildegard, Her Language, and Her Secretaries. In Townsend, D. *et al*. (eds), *The Tongue of the Fathers. Gender and Ideology in Twelfth-Century Latin*. Philadelphia: University of Pennsylvania Press. 102–35.

Kestemont, M. (2012). Stylometry for Medieval Authorship Studies. An Application to Rhyme Words, *Digital Philology: A Journal of Medieval Cultures*, **1 (1)**: 42-72.

Klaes, M. (1993), *Vita sanctae Hildegardis*. Turnhout: Brepols.

Moens, S. (2010). Twelfth-century epistolary language of friendship reconsidered. The case of Guibert of Gembloux, *Revue belge de philologie et d'histoire*, **88 (4)**. 983–1017.

Newman, B. (1987). *Sister of wisdom. St. Hildegard's theology of the feminine*. Los Angeles: University of California Press.

Newman, B. (1998). *Voice of the Living Light: Hildegard of Bingen and Her World*. Los Angeles: University of California Press.

Passarotti, M., and F. Dell'Orletta (2010). Improvements in Parsing the *Index Thomisticus* Treebank. Revision, Combination and a Feature Model for Medieval Latin. In *Proceedings of LREC 2010*. Valetta, Malta. 1694–71.

Pennebaker, J. (2011). *The Secret Life of Pronouns. What Our Words Say About Us*. New York: Bloomsbury.

Word-level Language Identification in "The Chymistry of Isaac Newton"

King, Levi
leviking@indiana.edu
Indiana University, United States of America

Kübler, Sandra
skuebler@indiana.edu
Indiana University, United States of America

Hooper, Wallace
whooper@indiana.edu
Indiana University, United States of America

Introduction

Language Identification is the task of determining the language of short text snippets, much shorter than for e.g., text classification. In Computational Linguistics (CL), language identification is generally considered a solved problem—but these methods assume that a text is monolingual, and at least 100 characters long. Furthermore, such methods cannot be used for multilingual texts in which the author switches between languages within a sentence, as in the "Chymistry of Isaac Newton" (Walsh and Hooper 2012), a collection of 119 alchemical manuscripts written by Newton over a 30-40 year period beginning in the mid-1660s. The team behind The Chymistry of Isaac Newton Project at Indiana University has transcribed these manuscripts and is publishing a digital scholarly edition at www.chymistry.org . Attempts to automatically analyze this corpus, even with basic levels like POS markup and lemmatization, are difficult because Newton frequently switches between English, Latin, and French within a paragraph or sentence, as shown in the following sentence: "The short lived & despicable plant [[LAT Paronychia folio Rutaceo [[ENG infused in beer, doth wonders in curing the kings evill." For this reason, we developed a new method for automatically identifying the language for single words rather than for complete texts. This method requires more information because the classification is finer-grained than standard methods, which have access to more text.

There is an additional complication because seventeenth-century English and French allowed many spelling variations, unlike Latin, which was fairly standardized.

We first train and test the method on the corpus itself. However, since this corpus is rather small for methods developed in CL, we also investigate whether the method can use either current texts or texts written by Newton's contemporaries. While this approach increases the amount of training data, it is unclear whether the additional data is useful given that all these additional Newton-era and modern texts are monolingual, and that the modern English texts will fail to exhibit the large variations in spelling that we see in Newton's manuscripts. Our experiments show that using Newton's own texts reaches the highest accuracy of close to 90%, but using modern text results only in a moderate decrease of 2% points.

Language Identification on the Document Level

All previous work in language identification assumes that each text to be identified is written in a single language. For this task, naïve approaches are often utilized with high success. The simplest methods use the presence of language-specific characters in a text to identify the language. Another method uses lists of the most common words of a language (Johnson 1993). Then, the text is classified based on which set of common words occurs most frequently.

Cavnar and Trenkle (1994) use the same method with relative frequencies of n-grams rather than words and reach an accuracy of 99.8% given texts with at least 400 n-grams.

Our work is based on work by Beesley (1988) and Mandl et al. (2006), who also extract n-grams. Beesley determines language identity for a whole text of any size by comparing probabilities of bigrams and characters of the individual words for each candidate language and labeling the text as the language most probable for the most words. Mandl et al. use n-grams to determine switch points between languages. Recent approaches use more sophisticated methods, such as vector-space models (Prager 1999) or multiple linear regression (Murthi and Kumar 2006). However, those approaches are difficult to use on the word level.

The Data Source

The Newton Alchemical Corpus. The Newton alchemical corpus comprises approximately 850,000 words, drawn from a three-language lexicon of 23,000 unique wordforms. Newton frequently alternates between English, Latin, and French. The collection contains documents written exclusively in either English or Latin. These documents were used as training data for our approach.

For both English and Latin, texts of approximately 70,000 words were used as training data. Additionally, a list of words was extracted from each monolingual training set and used as a lexicon for that language. French only occurs in the multilingual documents and much more rarely than English or Latin in these documents. Since there are no documents written exclusively in French, no French training data was available from this source.

Non-alphabetic elements (e.g., punctuation and numbers) are automatically labeled as non-words. Additionally, the texts include recipes, calculations and figures, and thus contain a large number of alphabetical variables and labels. These items are not relevant for language identification and present potential obstacles for automatic approaches. Thus, they were excluded from training/testing. Any string of letters not containing a vowel was determined to be a non-word.

For testing, we selected six texts (126,000 words) that contained a high degree of switches between languages and annotated them manually for the languages used. Three texts (20,000 words) were used for optimizing parameters, and three more texts (106,000 words) for testing. Note that the test set does not contain any French words.

Other texts: For English texts from Newton's era, we used excerpts of Francis Bacon's *The New Atlantis* (1627) and *Essayes or Counsels, Civill and Morall* (1625) and Robert Boyle's *The Sceptical Chemist* (1661) and *Experiments and Considerations Touching Colours* (1664). Newton-era Latin texts were excerpted from Rene Descartes' *Meditationes de prima philosophia* (1641), Benedict de Spinoza's *Ethica* (1677), and Carl Von Linne's Species *Plantarum* (1753).

The modern day training set for English was extracted from *The Los Angeles Times* and *The Washington Post* stories from 2006.

Word-Based Language Identification: The Newton Corpus

Our approach assumes that a particular document to be identified contains one or more of the languages used in the corpus: English, Latin, and French.

We automatically segment the texts into words, extract all n-grams per word, and calculate the relative frequencies of the n-grams in each language (normalized for capitalization). Figure 1 illustrates this process for the Latin word "ignis".

	$	i	g	n	i	s	#		Eng. RF	Lat. RF
1	$i								**0.01448**	0.01330
2		ig							0.00180	**0.00308**
3			gn						0.00025	**0.00196**
4				ni					0.00137	**0.00665**
5					is				0.00843	**0.01137**
6						s#			0.02129	**0.02476**
								Average RF:	0.04764	**0.061142**

Figure 1:
Extraction of bigrams (left) and comparison of relative frequencies. $ and # mark word boundaries.

We determine a language score by averaging over all n-gram probabilities of a word. Since there is no training data for French, we use only English and Latin for training, with a threshold: First, the scores for English and Latin are determined. If neither the English nor the Latin probability exceeds a pre-determined threshold, the word is determined to be French. This corresponds to the intuition that if the n-grams of a word are rare in both English and Latin, then that word is unlikely to be from those languages but from a different language. The final decision also takes the language label of the previous word into account. If the current word is in the lexicon of the language of the previous word, the current word is tagged as that language. If the word is not in the lexicon, we consider the language identity probabilities of the previous word by adding a proportion of that probability to the probability that the current word is English, and do the same for Latin. This decision captures the tendency of words to belong to the same language as the words in the immediate context, while allowing for the possibility of switches. At the beginning of a sentence, the threshold is higher than between words.

Performance on the current language identification task is defined as accuracy: the percentage of words in the test texts (excluding non-words) with correct language labels.

Ultimately, we found 5-grams to be the best performing setting.

Training set:	Accuracy
Newton Eng/Lat	89.84%

The results in table 1 show that we reach an accuracy of 89.84%. This is lower than the results reported for language identification on full documents, but the task is more difficult. The word misclassified most often is a genuinely ambiguous word, “in”. In general, the words most frequently misclassified are short (2-3 characters).

Word-Based Language Identification: Using Other Corpora for Training

Since the training set from the Newton corpus is rather small, we also investigated using either training texts from Newton’s era, or modern corpora. As no modern Latin is available, we used the Newton-era Latin texts.

Training set:	Accuracy
Newton	89.84%
Newton + Newton-era texts	89.28%
Newton-era texts	87.85%
Modern texts	87.11%

The results in table 2 show that using the small set of texts by Newton gives the highest accuracy. Adding Newton-era texts does not result in the expected increase in accuracy. Instead, accuracy decreases minimally from 89.84% to 89.28%. Using only Newton-era texts decreases accuracy by approximately 2%. Using modern texts also results in a small decrease in accuracy. However, our method does not suffer much from using modern texts, which suggests that the information about character differences between languages does not heavily depend on the changes in spelling.

Conclusion

We presented a novel method for identifying language on individual words in multilingual texts. We have shown that the method reaches an accuracy of 89.84% when trained on monolingual texts from the same author. However, if no such texts are available, other texts from the same era, or even current texts can be used with only a minor degradation in performance.

References

Beesley, R. K. (1988). Language Identifier: A Computer Program for Automatic Natural-Language Identification of

On-Line Text. *Languages at Crossroads: Proceedings of the 29th Annual Conference of the American Translators Association:* 47-54.

Cavnar, B. W., and J. M. Trenkle. (1994). N-Gram-Based Text Categorization. *Proceedings of Third Annual Symposium on Document Analysis and Information Retrieval*: 161-75.

Johnson, S. (1993). Solving the Problem of Language Identification. Technical Report. School of Computer Studies, University of Leeds.

Mandl, T., M. Shramko, O. Tartakovski, and C. Womser-Hacker (2006). Language Identification in Multi-lingual Text Documents. *Proceedings of the 11th International Conference on Applications of Natural Language to Information Systems* (NLDB 2006), Klagenfurt, Austria. Springer Lecture Notes in Computer Science 3999:153-163.

Murthy, K. N., and G. B. Kumar. (2006). Language Identification from Small Text Samples. *Journal of Quantitative Linguistics* 13: 57-80.

Walsh, J. A. and W. E. Hooper. (2012). The Liberty of Invention: Alchemical Discourse and Information Technology Standardization. *Literary and Linguistic Computing* 27: 55-79.

"Shall These Bits Live?" Towards a Digital Forensics Research Agenda for Digital Humanities with the BitCurator Project

Kirschenbaum, Matthew
mgk@umd.edu
MITH Maryland Institute for Technology in the Humanities, University of Maryland, United States of America

Lee, Cal
callee@email.unc.edu
School of Information and Library Science, University of North Carolina at Chapel Hill, United States of America

Woods, Kam
kamwoods@email.unc.edu
School of Information and Library Science, University of North Carolina at Chapel Hill, United States of America

Chassanoff, Alex
achass@email.unc.edu
School of Information and Library Science, University of North Carolina at Chapel Hill, United States of America

Olsen, Porter
polsen@umd.edu
MITH Maryland Institute for Technology in the Humanities, University of Maryland, United States of America

Mithra, Sunitha
sunithanc@yahoo.com
School of Information and Library Science, University of North Carolina at Chapel Hill, United States of America

In the title of an important early essay on the intersection of bibliographical method and literary interpretation, Jerome McGann (1988) poses the forensically inflected question: "Shall These Bones Live?" Several decades of subsequent scholarly activity have demonstrated the value of embedding the technical and material concerns of bibliographers and textual critics into the pursuits of mainstream literary studies. Though the rise of born-digital materials—documents and other media types that are created by computer and which circulate primarily or exclusively in "virtual" form—would seem to challenge the relevance of a bibliographic approach to textual studies. In fact, as a number of recent scholars have argued, digital texts themselves exhibit distinct material properties. (Kirschenbaum, 2008; Galey, 2012; Piper, 2012)

This paper seeks to enhance the discussion of digital materiality by exploring in a focused way the specific relevance of digital forensics to digital humanities research. Digital forensics is an applied practice involving the "preservation, identification, extraction, documentation, and interpretation" of digital data as legal evidence (Kruse & Heiser, 2002). We will present examples of all of these activities in the context of the BitCurator project, now concluding its first successful phase of development with funding from the Andrew W. Mellon Foundation. We will demonstrate that digital forensics offers researchers an advanced community of practice around the acquisition, processing, management, and analysis of born-digital objects, and we will suggest that as humanities scholars increasingly engaged with objects of born-digital cultural heritage, digital forensics will emerge as a significant modality within digital humanities. Finally, we will argue that digital forensics offers an especially fruitful area for

collaborative research and project development between digital humanities and the digital archives community.

The BitCurator project is a joint effort led by the School of Information and Library Science at the University of North Carolina at Chapel Hill (SILS) and the Maryland Institute for Technology in the Humanities (MITH) to develop an environment for collecting professionals that incorporates the functionality of many existing digital forensics tools. There are, however, two fundamental needs for collecting institutions that are not addressed by software designed for the digital forensics industry: incorporation into the workflow of a colleting institution's ingest and collection management environments, and provision of public access to the data. BitCurator provides an environment in which users can create forensically-packaged disk images, perform a variety of triage tasks on objects within file-systems, extract and repackage metadata, and redact sensitive information from digital materials.

The BitCurator Environment is a fully functioning Linux system built on Ubuntu 12 that has been customized to meet the needs of librarians and archivists, and can be run as a stand-alone operating system, a virtual machine, or as a set of individual tools (see Figure 1). Development has been informed by consultations with two groups: a Professional Expert Panel (PEP) of individuals who are at various levels of implementing digital forensics tools and methods in their collecting institution contexts, and a Development Advisory Group (DAG) of individuals who have significant experience with development of software. (The membership in these two groups represents many of the key researchers and practitioners currently involved in the archival processing of born-digital materials.) The BitCurator environment incorporates a number of useful digital forensics tools that can easily be integrated into digital curation workflows. A sampling of those tools, all of which are available as public domain or open-source (General Public License) software, includes:

- Guymager: A tool for creating disk images in one of three commonly used disk image formats (dd, E01, and AFF).
- Custom Nautilus scripts: A collection of enhancements to Ubuntu's default file browser that allow users to quickly generate checksums, identify file types, and safely mount drives, among other tasks.
- The Sleuth Kit: A digital investigation platform.
- fiwalk: An open source tool for processing disk images, producing Digital Forensics XML and human readable metadata on file system structure and contents
- bulk_extractor: A program that extracts information —including Personally Identifying Information— from disk images without parsing the file-system. bulk_extractor generates reports on the information in both human and machine readable formats, and includes a GUI front-end, Bulk Extractor Viewer.
- sdhash 2.x – A tool to evaluate file similarity using similarity digests.
- Ghex: An open source hex editor that allows users to view a file's bitstream in hexadecimal format.

In addition, the BitCurator team is in the process of building Python-based reporting tools that reprocess and provide visualizations based on the output of forensics tools that produce Digital Forensics XML; these tools are currently distributed separately via GitHub, and are being integrated into the environment as the project progresses.

The relevance of these tools for humanities research comes into focus when we consider that disk images are fast becoming indispensable units of analysis for scholars seeking to understand the primary sources of digital cultural heritage. A disk image is a bit-accurate copy of the raw media to obtain what amounts to a binary facsimile (or "snapshot") of every signal or inscription recorded on an original piece of source media. As such, it is the gold standard in both the legal forensics community (where investigators routinely conduct their analysis on an authenticated disk image as opposed to the original storage device) and in archival processing (Woods, Lee, & Garfinkel, 2011). Because disk images preserve a record of both file-level metadata as well as the actual physical traces of data recorded on the surface of the media, they are essential for reconstructing the "original order" of digital records, i.e. correct chronologies of file creation and manipulation (which can be obtained through techniques ranging from file system analysis to digital stratigraphy) (Xie, 2011; Woods and Lee, 2012). A disk image also allows an investigator or researcher access to unallocated or even potentially damaged portions of the original media, creating the possibility of restoring fragments of files that would be unrecoverable after normal copy processes (Kirschenbaum et al., 2009). Finally, a disk image can be used as the basis of an emulated experience allowing a researcher to recreate most aspects of the original operating system and computing environment. All of these activities have clear parallels in traditional textual studies and bibliography. Specialists in those areas routinely seek to understand and describe the relationships among sets of primary source documents, identify and enumerate the distinguishing features of individual documents and texts (using those findings to facilitate inquiries ranging from paleography to constructing version histories), and editing or curating primary source materials for presentation in a variety of formats and settings. A disk image thus manifests a strong analogy to the primacy of material/documentary evidence in more traditional forms of bibliographic analysis, enabling a researcher to analyze numerous aspects of a

born-digital object relevant to scholarly concerns. As such, we predict that disk images will become increasingly familiar to scholars and researchers working on late twentieth- and twenty-first century history, culture, literature, and the arts, as more and more leading figures consign electronic records to archives who are processing them in environments like BitCurator (AIMS).

While BitCurator's primary user community consists of librarians and archivists engaged in the processing of digital materials in a variety of institutional settings, its functionality is also useful to those digital humanities scholars with direct access to born-digital materials (Carroll, 2011; Schuessler, 2012). Digital humanities thus has an opportunity to develop a robust research agenda in conjunction with the digital archives community that would ensure that scholarship in areas like contemporary literature, recent world history, digital culture, politics and government, and the arts proceeds on a sound technological footing, using tools and best practices designed to ensure the stability and reliability (and accessibility) of the born-digital cultural record. While the digital archives and forensics communities have developed mature tools around file system metadata and the extraction of personal identifying information, for example, they lack the digital humanities community's experience in analyzing complex text corpora. Data extracted from disk images can be redirected to tools and assets ranging from geo-spatial visualization to topics modeling and other forms of textual analytics. The open and extensible nature of the BitCurator environment affords the potential for such collaborations, and so the paper concludes with some use case scenarios in this regard.

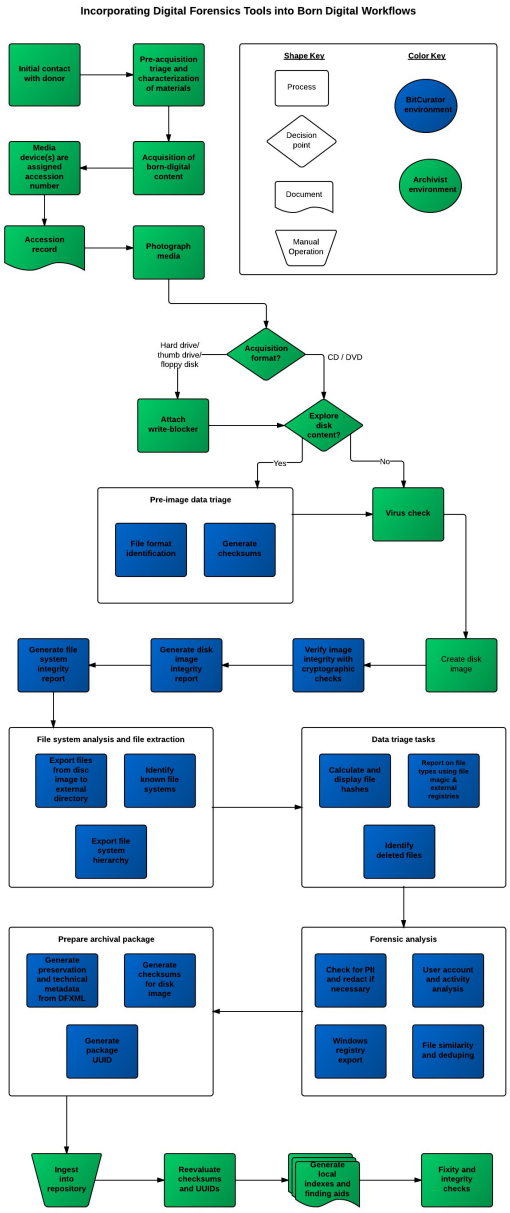


Figure 1. A model BitCurator-supported workflow.

References

AIMS Working Group (2012). *AIMS Born-Digital Collections: An Inter-Institutional Model for Stewardship*. http://www2.lib.virginia.edu/aims/whitepaper/AIMS_final.pdf

Carroll, L., E. L. Farr, P. Hornsby, and B. Ranker (2011). A comprehensive approach to born-digital archives. *Archivaria* 72. 61-92.

Galey, A. (2012). The enkindling reciter: E-Books in the bibliographical imagination. *Book History* 15. 210-47.

Kirschenbaum, M., E. L. Farr, K. M. Kraus, N. Nelson, C. S. Peters, and G. Redwine (2009). Digital materiality: Preserving access to computers as complete environments. in *Proceedings of the Sixth International Confereence on Digital Preservation (iPRES)*. held October 5-6 2009 at California Digital Library, San Francisco. 113-120. http://escholarship.org/uc/it em/7d3465vg#

Kirschenbaum, M. (2008). *Mechanisms: New Media and the Forensic Imagination*. Cambridge, MA: MIT Press.

Kruse, W. G. and J. G. Heiser (2002). *Computer Forensics: Incident Response Essentials*. Boston: Addison-Wesley.

McGann, J. J. (1988). *The beauty of inflections: Literary investigations in historical method and theory*. Oxford: Clarendon Press.

Piper, A. (2012). *Book was there: Reading in electronic times*. Chicago: University of Chicago Press.

Schuessler, J. (2012) Tale of the floppy disks: How Jonathan Larson created *RENT*. *The New York Times* February 1. http://artsbeat.blogs.nytimes.com/2012/02/01/tale-of-the-floppy-disks-how-jonathan-larson-created-rent/

Woods, K., and C. A. Lee (2012). Acquisition and processing of disk images to further archival goals. In *Proceedings of Archiving 2012* (147-152). Springfield, VA: Society for Imaging Science and Technology. Retrieved from: http://ils.unc.edu/callee/archiving-2012-woods-lee.pdf

Woods, K., C. A. Lee, and S. Garfinkel (2011). 'Extending digital repository architectures to support disk image preservation and access'. in *Proceedings of the 11th Annual International ACM/IEEE Joint Conference on Digital Libraries* held at Association for Computing Machinery. New York: 57-66. http://dx.doi.org/10.1145/1998076.1998088.

Xie, S. L. (2011). Building foundations for digital records forensics: A comparative study of the concept of reproduction in digital records management and digital forensics. *American Archivist*, **74 (2)** 576-599.

Simulation of the Complex System of Cultural Interaction

Kretzschmar, William

kretzsch@uga.edu
Department of English, University of Georgia, United States of America

Juuso, Ilkka

ilkka.juuso@ee.oulu.fi
Faculty of Engineering, University of Oulu, Finland

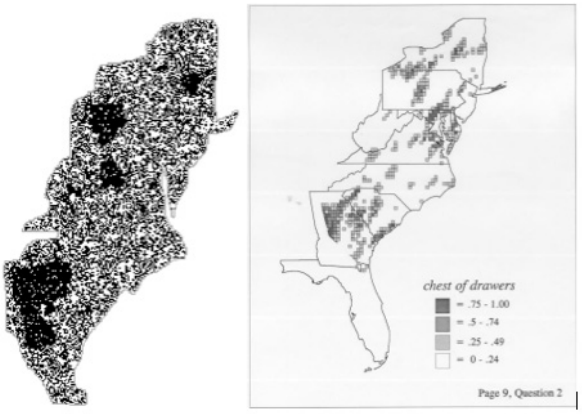
Bailey, C. Thomas

ctbailey@uga.edu
Institute of Artificial Intelligence, University of Georgia, United States of America

The crucial issue for space and time in language and cultural study is modeling diffusion, how characteristics spread spatially over time. Diffusion creates the regional and social patterns in language that we all perceive around us at any particular moment in time, and the same is true, though perhaps we notice it less frequently, of other cultural practices such as food ways, architectural styles, clothing styles, folk and cultivated narrative and literary modes, and social roles and relationships. The process of diffusion certainly occurs as a result of cultural interaction — to use language as prime example, massive numbers of people talking (and more recently writing) to each other. The new science of "complex systems" shows that order emerges from such systems by means of self-organization: particular variants come to be more or less frequent among different groups of people or types of discourse (the same nonlinear curve has a different order of variants at every scale of analysis), and variant frequency comes to mark identity of the different regional and social groups. The process of diffusion corresponds to the frequency differences that emerge from our choices of what to say and write, in the buzz and hum of our daily human interactions with each other. However, we cannot observe diffusion directly because it has never been feasible to collect the time series data required to do so. Computer simulation is the only way that we can model diffusion as the adaptive aspect of complex systems in speech and culture (see Gilbert, Miller and Page, Wolfram). This paper describes the construction and implementation of a GIS-aware cellular automaton for use as a multidimensional simulation for speech. Simulations begin with seeding of characteristics across the live cells of a large matrix, say, with linguistic variants seeded along the coast to model original settlement, or with variants seeded across the survey area of the Linguistic Atlas Project (LAP). Throughout thousands of iterations that correspond to the daily interaction of speakers across time, we observe the distributional patterns in the variants as they emerge.

The key feature of this simulation, one not yet attempted to our knowledge, is that we validate it with respect to the emergent "clustered" linguistic distributions known to occur in the LAP, not the actual patterns but instead the sort of pattern that emerges in every elicitation. We know that we

have achieved a successful simulation if the result, after hundreds of iterations, is a complex system that shows the clustered distributions we expect in Atlas feature maps (as below; CA at 1000 iterations at left, DE map at right).



In work to date we have demonstrated that such stable clusters do emerge in the simulation as the result of restricted rule sets with a random component included.

When social information is associated with each cell, there is an additional opportunity for weighting the decisions about feature adoption and maintenance. In the LAMSAS matrix of 9000+ cells (after creation of non-active boundaries), 1162 cells have metadata available from the LAMSAS survey. We have interpolated metadata for the empty cells, using the primary social characteristics of urban/rural community type, and informant type (an index showing three levels of education and social integration). These characteristics can be applied to empty cells with respect to the nearest neighbors with metadata in the cellular automaton, with respect also to the overall proportions of the characteristics in the survey as a whole, and also with a small random component. We can execute the interpolation on demand, so that we achieve somewhat different but still valid interpolations for testing the simulation.

Given a matrix with social information available for every cell, we apply social weight to the decision for the presence of the feature of interest in every cell of the matrix, based on the rules assigned for the cellular automaton. The decision to adopt or retain the feature of interest is made by rule according to a certain number of similar neighbors (i.e., adopt the feature of interest if 2, 3, or 4 neighbors have it; retain the feature if 5, 6, 7, or 8 neighbors have it). The proximity of first-order neighbors is the primary characteristic for the decision; in GIS applications it is common to invoke the inverse square law by which a second-order element, twice as far away, has only 25% of the influence of an immediately proximate first-order neighbor. Our method considers social characteristics as second-order phenomena, and assigns a small negative weight of .2 (really, it could be any decimal 0> < .25) to each socially dissimilar neighbor, when a socially-similar neighbor receives the full value of 1 for a first-order relationship. Practical implementation of this policy is to award 1 point to a similar neighbor, and .8 point to a dissimilar neighbor, considering a single social characteristic. When we assign weights to more than one social variable at a time, the cumulative influence of the weight plus the random decision component should not exceed the same .25 as a maximum weight. This means that two social variables might be assigned weights of .1 each, if they were judged to be equally influential, or of, say, .15 and .05 if one social variable was thought to be significantly more important than the other. Overall, social weighting conditions the likelihood for adoption and maintenance of any cultural feature. We will illustrate the effects of including social weighting within a cellular automaton that is essentially proximity driven.

In this way we believe that we are breaking new ground in simulation of cultural interactions as complex systems. The study of speech as a complex system addresses language as an aspect of culture that emerges from human interaction. We believe that successful simulation of speech in cultural interaction as a complex system can suggest how other aspects of the humanities, such as sites or artifacts or styles in archaeology, can diffuse and change across space and time.

References

Gilbert, N. (2008). *Agent-Based Models.* Thousand Oaks, CA: Sage.

Gilbert, N., and K. Troitzsch (2005). Simulation for the Social Scientist. Maidenhead: Open University Press.

Linguistic Atlas Project. http://www.lap.uga.edu

Miller, J., and S. Page (2007). *Complex Adaptive Systems: An Introduction to Computational Models of Social Life.* Princeton: Princeton University Press.

Wolfram, S. (2002). *A New Kind of Science.* Champaign, IL: Wolfram Media.

Agents for Actors: A Digital Humanities framework for distributed microservices for text linking and visualization

Küster, Marc Wilhelm
kuester@fh-worms.de
FH Worms, Worms

The Scene

A raucous party on the eve of the Battle of Bosworth field in 1485, wine flows by the gallon, the table bends under the weight of the meat and breadcrumbs fly across the table as Richard III rises painfully to address his loyal peers:

> Now is the summer of our sweet content,
> Made overcast winter by these Tudor clouds.
> And I that am not shaped for black-faced war,
> I that am rudely cast and want true majesty,
> Am forced to fight, To set sweet England free.
> I pray to Heaven we fare well,
> And all who fight us go to Hell
>
> (transcription of the audio recording (British Broadcasting Corporation 2009), checked against (Curtis 2001)).

Thus starts Rowan Atkinson's fabulously funny alternate history of Britain, the History of the Black Adder (Atkinson et al. 1983), one of the hallmark BBC sitcoms of the 1980s.

Fast forward through the show, right to the closing credits: "Written by Richard Curtis and Rowan Atkinson with additional dialogue by William Shakespeare". And, indeed, even Richard III's short speech contains at least three near-verbatim citations from Shakespeare's play of the same name (and actually a fourth and a fifth, but more on that later):

> Now is the Winter of our Discontent
> Made glorious Summer by this Son of Yorke: […]
> But I, that am not shap'd for sportiue trickes, […]
> I, that am Rudely stampt, and want loues Maiesty, […]
> And that so lamely and vnfashionable,
> That dogges barke at me, as I halt by them[1]

All would be good if the citations were indeed verbatim, but they are not — the wit of Richard's speech and indeed of much of Blackadder's attraction depends on twisting citations, creating tension between original and new wordings. Blackadder like much of postmodern British Sitcom consciously sees history as a "cluttered patchwork of questionable stories which have been re-written, re-evaluated and ridiculed" ((Roberts 2012), pos. 26, Kindle edition), using allusions precisely to enhance the sense of a subjective, patchwork-like history.

Enter Agents for Actors

Agents for Actors (AfA) (https://github.com/mwkuster/agents-for-actors) is an experimental, LGPL-licensed Open Source "framework for distributed microservices for text linking and visualization" that the author has developed to calculate precisely the types of "twisted" citations that we are seeing. It takes some inspiration from W. Artes' Bachelor thesis (Artes 2012) on Similarity Search (cf. also (Hedges et al. 2012)), that the author has supervised, especially on the choice of NGram models for comparison, but is an independent implementation that deviates in the way the NGram model is built (cf. below). AfA is linked to TextGrid's Text-Text-Image-Link Editor (Selig, Küster, and Conner 2012).

Afa identifies allusions between texts and their presumed sources and gives exact provenance information as XPointers (Küster et al. 2011). As an additional spin we make the comparison of Black Adder's modern orthography transcript against the original-spelling First Folio, transcribed by Trevor Howard-Hill (http://ota.ox.ac.uk/id/3014).

AfA is extendable to other similarity models and measures and could be adapted for new visualization frontends. Given the computational complexity it embraces multicore architectures and parallelizes computations wherever possible. AfA situates itself between the macro-vision of big data digital humanities (e.g. (Jockers 2012) and the forthcoming (Jockers 2013)) and the micro-vision of the classical, manually encoded critical apparatus.

Implementation

AfA is implemented in the functional language Clojure (Hickey 2010) (Emerick, Carper, and Grand 2012; Halloway and Bedra 2012) on top of the Java Virtual Machine (JVM). Clojure thrives on immutable functional data structures (Okasaki 1999) for heavily multithreaded applications. Mutable operations are largely under control of Clojure's Software Transactional Memory (STM).

AfA uses two Clojure paradigms to parallelize activities:

- Futures: non-blocking threads that are transparently managed to parallelize calculations
- Agents: asynchronous, used to interact with the visualization layer

The code basis itself consists of a number of individual functions for interacting with files, calculating similarity measures, handling visualization etc. They can be regrouped flexibly.

Handling XML

AfA currently expects both source and target to be encoded in XML. Both therefore contain parts such as the TEI header, cast lists or stage directions that without filtering generate noise when searching for references. Still, we need to preserve the exact pointers to source and target fragments in the underlying XML files to guarantee traceability.

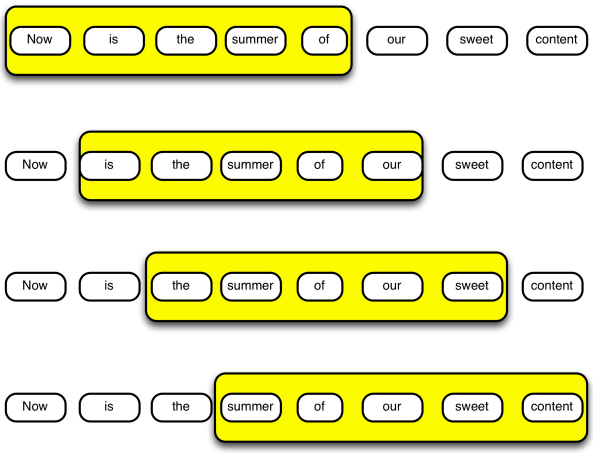
With mutable data structures this objective would be difficult to attain without costly copying operations of XML structures in memory. The immutable abstraction of Zippers (Huet 1997) offers a solution by having the XML structures only once in memory, the algorithm operating on pointers to individual elements ("locations"). Furthermore, unlike e.g. XML DOMs Zippers are generic abstractions for tree structures, not only XML. AfA can hence be adapted to any form of structured data sources.

Measuring similarity

The AfA framework allows for multiple models and measures. At is simplest, the comparison is done using NGram models (cf. (Manning and Schuetze 1999)), that are used for fuzzy text comparison, e.g. in (Kestemont, Daelemans, and De Pauw 2010) and (Bernholz and Pytlik Zillig 2011), not to mention in Google-style big-data (Google 2010).

The tests presented in this paper are done with a variation of the NGram model, combining NGrams of words, that move in a slider over the text, creating chunks of size C, with NGrams of letters for the actual comparison of those chunks.

In the following example C=5. This way, the phrase "Now is the summer of our sweet content" into four chunks of five words each:



Each of these chunks is compared with a chunk from the supposed source material, built with the same algorithm, so that the totality of compared chunks is the Cartesian product NxM, N being the number of chunks in the source and M that in the target. Each of these pairs of chunks is compared using an NGram of `ngram-count` characters to smoothen over differences in spelling. The respective confidence level for a hit is calculated using the maximum dice-coefficient (cf. (Manning and Schuetze 1999), table 8.7) applied to this combination of chunks: `(apply max (map (fn [[t1 t2]] (dice-coefficient (ngrams ngram-count t1) (ngrams ngram-count t2))) (for [chunk1 chunk-seq1 chunk2 chunk-seq2] [chunk1 chunk2])))` The algorithm works for other measures returning similarities normalized to [0,1].

Applying this admittedly computationally expensive algorithm to Richard's speech with C=6, N=4 and a minimum confidence level of 0.65 identifies in Shakespeare's complete First Folio precisely the expected links (XPointers referring to http://ota.ox.ac.uk/id/3014):

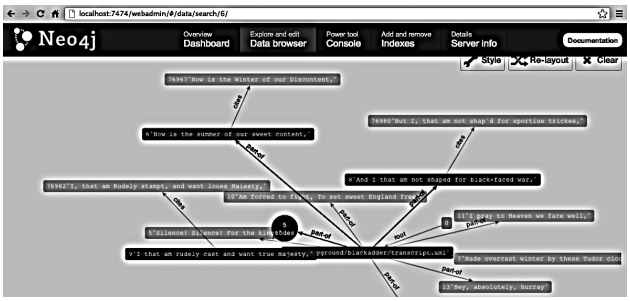
Blackadder	First Folio	Confidence level
Now is the summer of our sweet content,	Now is the Winter of our Discontent, xpointer(/TEI[1]/text[1]/group[1]/text[23]/body[1]/div[1]/ab[1]/text()[1])	0.74
And I that am not shaped for black-faced war,	But I, that am not shap'd for sportiue trickes, xpointer(/TEI[1]/text[1]/group[1]/text[23]/body[1]/div[1]/ab[1]/text()[14])	0.85
I that am rudely cast and want true majesty,	I, that am Rudely stampt, and want loues Maiesty, xpointer(/TEI[1]/text[1]/body[1]/sp[2]/l[4]/text()[1])	0.65

The algorithm cannot identify the fourth allusion, though, that contrasts the concepts of "overcast winter" with "glorious Summer".

Dressing up

For visualization AfA uses Neo4j, an increasingly popular Open Source Non-SQL graph database (Neo4j.org 2012). Neo4j centres around two Topic Map (ISO/IEC 2002) like concepts, nodes and relationships. Both have unique identifiers and can have arbitrary properties besides. Relationships must have a start and an end node.

With neocons (Klishin 2012) Neo4j has an intuitive Clojure interface that permits to store and query the graph. Neo4j's admin interface also has one of the more innovative graph UIs available in an off-the-shelf database.



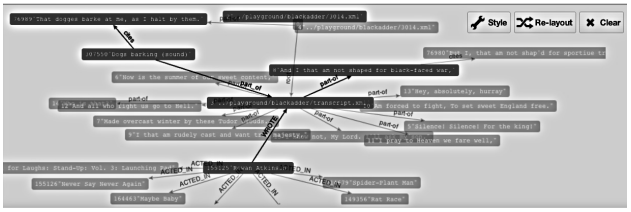
Intertextuality and a bit of theory

This screenshot shows the individual phrases of Richard's speech in Blackadder with their links to their source ("part-of") and the phrases they cite ("cites"), giving a good visualization of the intertextuality (Kristeva 1969) (Barthes and Marty 1980; Barthes 1968) (Allen 2011) that animates much of Atkinson's comedy.

Note that AfA does not "interprete" the links. Statistical methods rarely have an obvious interpretation, their epistemological openness and continuity being an asset for postmodern texts in a world of "stylistic and discursive heterogeneity without a norm" (Jameson 1991). If anything, Genette's theory of *hypertextualité* (Genette 1982) might give us an adequate terminological apparatus.

Outlook

Three of the citations in Richard's speech AfA has identified for us, the fourth we have discussed — but there is a fifth. Films are more than dialogue; intertextuality can just as well be construed without words. When Atkinson's Richard is halfway through his speech we hear a dog barking — Richard is so ugly that "*dogges barke at me*". Also, there is open linked data out there that situates Blackadder in context, here linking manually to Freebase movie data with some of Atkinson's other films (for handling Freebase data cf. (Redmond and Wilson 2012), chapter Neo4j/Big data):



In the end an approach focussed on a narrow understanding of intertextuality will not suffice for audiovisual media; it must evolve into a network including knowledge and symbols (Peirce 1998), contextual, textual and non-textual. Now we can only manually add nodes establish these links, but there may be another untold history ahead.

References

Allen, G. (2011). *Intertextuality*. 2nd edn. Abdingdon: Routledge.

Artes, W. (2012). In Marc Wilhelm Küster, (ed). "Konzeption Und Entwicklung Eines Plug-in Basierenden SOAP-Services Für Die Ähnlichkeitsanalyse Im Text-Text-Link-Editor." Worms: University of Applied Sciences Worms.

Atkinson, R., B. Blessed, E. Gray, R. East, T. McInnerny, and R. Robinson (1983). In Shardlow, M. *The Black Adder*. http://www.bbc.co.uk/programmes/p006cx8w .

Barthes, R. (1968). "La Mort De l'Auteur." *Manteia* (4e trimestre).

Barthes, R., and E. Marty (1980). *Oeuvres Complètes: 1974-1980*.

Bernholz, C. D., and B. L. Pytlik Zillig (2011). "Comparing Nearly Identical Treaty Texts: a Note on the *Treaty of Fort Laramie with Sioux*, Etc., 1851 And Levenshtein's Edit Distance Metric.." *Llc* 26.1: 5-16. doi:10.1093/llc/fqq016.

British Broadcasting Corporation. (2009). *The Black Adder*. Unabridged. AudioGO Ltd.

Curtis, R.. (2001). *Blackadder: the Whole Damn Dynasty: 1485-1917*. Re-issue. Penguin Books, Limited (UK).

Emerick, C., B. Carper, and C. Grand (2012). "Clojure Programming — Practical LISP for the Java World." *O'Reilly 2012*.

Gérard, G. (1982). *Palimpsestes. La Littérature Au Second Degré*. Paris: Éditions du seuil.

Google. (2010). *Google Books NGram Viewer*.

Halloway, S., and A. Bedra (2012). *Programming Clojure*. Second Edition. Pragmatic Bookshelf.

Hedges, M., A. Jordanous, S. Dunn, C. Roueche, M.W. Küster, T. Selig, M. Bittorf, and W. Artes (2012). "New Models for Collaborative Textual Scholarship." In, 1-6. doi:10.1109/DEST.2012.6227933.

Hickey, R. (2010). "Clojure."

Huet, G. (1997). "The Zipper." *Journal of Functional Programming* 7 (5) (September).

ISO/IEC. (2002). *ISO/IEC 13250: Information Technology — SGML Applications — Topic Maps*. {ISO}.

Jameson, F. (1991). *Postmodernism or the Cultural Logic of Late Capitalism*. Durham: Duke University Press.

Jockers, M. L. (2012). "Computing and Visualizing the 19th-Century Literary Genome." In Hamburg. http://www.dh2012.uni-hamburg.de/conference/programme/abstracts/computing-and-visualizing-the-19th-century-literary-genome/ .

Jockers, M. L. (2013). *Macroanalysis: Digital Methods and Literary History (Topics in the Digital Humanities)*. 1st edn. University of Illinois Press. http://www.matthewjockers.net/macroanalysisbook/ .

Kestemont, M., W. Daelemans, and G. De Pauw (2010). "Weigh Your Words - Memory-Based Lemmatization for Middle Dutch.." *Llc* 25.3 287-301. doi:10.1093/llc/fqq011.

Klishin, M. (2012). "Neocons." *Clojureneo4j.Info*. Accessed October 29. http://clojureneo4j.info/ .

Kristeva, J. (1969). . Éditions du seuil.

Küster, M. W., C. Ludwig, Y. Al-hajj, and T. Selig (2011). "TextGrid Provenance Tools for Digital Humanities Ecosystems." In, 317-323. doi:10.1109/DEST.2011.5936615.

Manning, C. D., and H. Schuetze (1999). *Foundations of Statistical Natural Language Processing*. 1st edn. Cambridge, MA: MIT Press.

Neo4j.org. (2012). "Neo4j." *Neo4j.org*. http://neo4j.org/ .

Okasaki, C. (1999). *Purely Functional Data Structures*. Cambridge University Press.

Peirce, C. S. (1998). "What Is a Sign?." In *The Essential Peirce. 1893-1913*, ed. Peirce Edition Project. 2 Bloomington.

Redmond, E., and J. R. Wilson (2012). *Seven Databases in Seven Weeks: a Guide to Modern Databases and the NoSQL Movement*. Pragmatic Bookshelf.

Roberts, J. F. (2012). *The True History of the Black Adder: the Unadulterated Tale of the Creation of a Comedy Legend*. Preface Publishing.

Selig, T., M.W. Küster, and E. Conner (2012). "Semantically Connecting Text Fragments — Text-Text-Link-Editor (Poster)." In, ed. Jan Christoph Meister, 518—520. Hamburg. http://www.dh2012.uni-hamburg.de/wp-content/uploads/2012/07/HamburgUP_dh2012_BoA.pdf .

Notes

1. The Tragedy of Richard the Third, cited after Shakespeare's First Folio as transcribed in http://ota.ox.ac.uk/id/3014

XML-Print: Addressing Challenges for Scholarly Typesetting

Küster, Marc Wilhem
kuester@fh-worms.de
Worms University of Applied Sciences, Germany

Selig, Thomas
selig@fh-worms.de
Worms University of Applied Sciences, Germany

Georgieff, Lukas
lukas.georgieff@hotmail.com
Worms University of Applied Sciences, Germany

Sievers, Martin
sievers@uni-trier.de
Trier Center for Digital Humanities (Kompetenzzentrum), Germany

Bittorf, Michael
bittorf@fh-worms.de
Worms University of Applied Sciences, Germany

1 Introduction

At last year's DH conference, we presented our Open Source project *XML-Print live*(Sievers, et al., 2012). The audience's responses gave a strong impulse for the second project stage to extend and fine-tune the software and improved its technical architecture. The latest build is available for download from http://www.xmlprint.eu/.

While the former talk focused on the infrastructure and general ideas of XML-Print, this paper delivers insight into the challenges of the project's development process, which are paradigmatic for the dual nature of the challenges faced by many DH development projects. This includes

issues already solved as well as concepts for open ones. The following aspects will be discussed:

1. Typesetting critical editions is one of the major use cases for *XML-Print*. Apart from "standard" typesetting features we have implemented support for manging an arbitrary number of apparatuses (cf. Section 2.1).
2. The typesetting engine is written in the functional programming language F#, which — among other advantages — offers powerful parallelization techniques. However, we must use appropriate programming routines for that. The ideas behind parallelizing the typesetting are described in section 2.2.
3. Today's typesetting world is global, meaning support for many writing systems including bidirectional scripts. Section 2.3 reports about the concepts for bidirectional typesetting within *XML-Print.*
4. Last, but not least software development needs sophisticated test concepts to recognize problems before new versions are published. This is important not only for obvious build or compilation errors, but especially for typesetting problems. The image-based test approach is presented in Section 2.4.

2 Challenges

2.1 Critical Apparatus

Typesetting a single footnote apparatus is not sufficient for scholarly purposes. We need a robust and generic algorithm to place any number of apparatuses to be grouped together on a page.

The user manages this grouping process by creating an appropriate mapping between a layout format and XML elements. The output position of the footnote apparatuses on a page is determined by the order given in the footnote dialogue (cf. Figure 1). This information is coded into the XSL-FO+ [1] file as an XML attribute to the footnote element, e.g.`<footnote place="0" fn-type="footnote">.` [2].

The typesetting engine sorts the footnotes corresponding to their place attribute and calculates the space needed on a concrete output page, including spacing between apparatuses. All this can already be done using *XML-Print*. An example is illustrated in Figure 2.

However, when typesetting critical editions, footnote apparatuses alone are not sufficient. There has to be a possibility to refer to the line number of a lemma or to use a user-defined reference schema as well. That reference is used instead of the common footnote sign. The typesetting engine must furnish the necessary references, which is only available at runtime and calculated in the output routines.

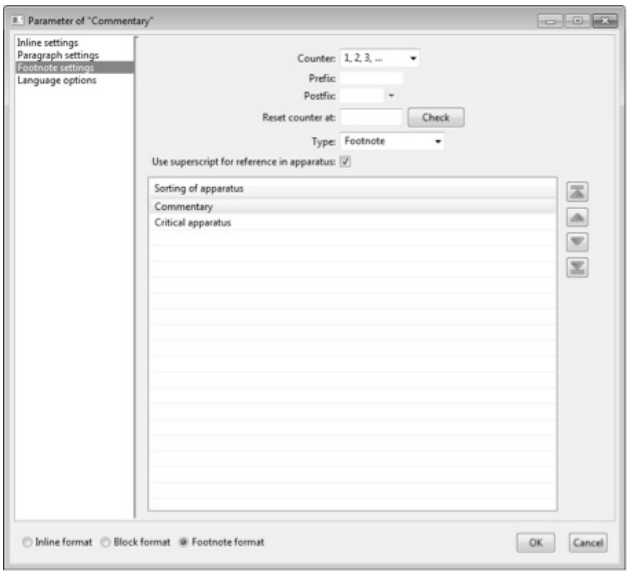


Figure 1:
The grouping and sorting of footnote apparatuses is done with the footnote dialog from the graphical user interface.

Erste deutsche Kunstausstellung in Neuyork.[40]

In zwei kleinen nüchternen Zimmern im obersten Stock eines alten Hauses an der 47. Straße zwischen Fünfter und Madison Avenue hat sich der deutschen Kunst ein Heim aufgetan. In dem einen Raum gibt es gute deutsche Bücher und eine kleine Kunstausstellung, die den überlieferten Anschauungen entspricht. In dem andern herrschen die allerjüngsten expressionistischen Maler. In der Mitte auf dem Ehrenplatz Kandinsky und an dem Rest der Wände eine Phantasie in abgerissenen Straßenbahnkarten, Brot- und Kartoffelmarken von Schwitters, eine Flußlandschaft von Karl Mense[j41], die sich

j Karl Mense] *eigentlich: Carlo Mense*

38 monatlich erscheinenden Zeitschrift „Panismus...zusammengefunden haben.] Bisher gar nichts gefunden, auch nicht zu dem Typ. Panismus von Pan-> griech. Hirtengott, der der Lebensfreude, Musik und Tanz zugetan ist... Keine Zeitschrift mit diesem Titel gefunden und auch den Typ nicht...

39 Alfred Ranft] (1895 – 1978), Lehrer? RECHERCHE BERLIN

40 Erste deutsche Kunstausstellung in Neuyork.] Es handelt sich um die Ausstellung der 1920 in New York von Katherine S. Dreier, Marcel Duchamp und Man Ray gegründeten Société Anonyme im Nov./Dez. 1920. Die Vereinigung setzte sich dafür ein, durch Ausstellungen und Projekte die europäische Avantgarde-Kunst in Amerika bekannt zu machen und einen internationalen künstlerischen Austausch zu initiieren (vgl. Ausst.-Kat. New Haven 2006). Die Ausstellung der Société war KS' erste in den USA; in den folgenden Jahren wurden seine Werke regelmäßig auf deren Schauen gezeigt. -> Kein KT im CR, welche Werke von KS wurden gezeigt?

41 eine Flußlandschaft von Karl Mense] Es handelt sich wahrscheinlich um das Gemälde Flusslandschaft mit Dampfer (The River Wuppe) (1913) des expressionistischen und neu-sachlichen Künstlers Carlo Mense (Rheine/Westfahlen 1886 – 1965 Königswinter), welches Katherine S. Dreier 1920 aus Herwarth ↗Waldens ↗Sturm-Sammlung erwarb (vgl. Ausst.-Kat. Köln 1993, S. 173; Ausst.-Kat. New Haven 2006, S. 104).

Figure 2:
Example of two footnote apparatuses of a letter from the "Kurt-Schwitters-Briefe" (Wie Kritik zur Kunst wird). For a real critical apparatus references to line numbers of lemmata have to be implemented.

2.2 Parallelizing the Typesetting

Implementing a typesetting system with all its characteristics and subtleties is a chal-lenge concerning two technical aspects: efficiency and memory consumption. Both aspects can be satisfied by the data structures and algorithms used in the implementation. We need to parallelize the *XML-Print* typesetting engine as the rendering process is resource intensive. In particular, when pre-rendered pages have to be adapted to changes and need final rendering, the process is very time-consuming. This is the case if *XML-Print* typesets remaining footnotes from previous pages. The content of the current typeset page must be balanced with the contents of the remaining and the current footnotes. The described problem can be solved in mainly two ways: directly improving the typesetting of footnotes or/and improving the overall rendering process.

Improving the typesetting of footnotes can directly be achieved by a more sophisticated algorithm. That means not by balancing the contents line by line but by using binary cuts instead, i.e. starting with *lines/1* , *lines/2* , *lines/4* , …, *lines/lines* until the footnote and the current page's contents are in balance. Unfortunately, this solution covers only one of several performance problems.

Therefore, improving the overall rendering process must be the final goal. The parallelization efforts are based on the dynamic programming idea by Bellman (1954). Thus, the rendering process is separated in typesetting single sequences, i.e. single XML fragments. For that the input XML file is read and then split into fragments representing such sequences. These are the basic elements in the XSL-FO (Anders 2006) structure that can be processed independently from other data. Afterwards, these sequences are typeset in parallel, are then concatenated and finally written into a single document.

Implementing the introduced parallelization mechanism requires the following characteristics:

- Resource-efficient XML processing, especially when reading the file and dividing it into fragments
- Isolated, side-effect free and generic rendering algorithms for separated sequences.

Indeed, F# as a functional programming language offers the possibilities to address these requirements.

2.3 Bidirectional Text

Like many other typeset editors (e.g. Texmaker, TeXstudio and TeXShop for LaTeX), *XML-Print* must support bidirectional (BiDi) texts. The implementation of this feature is part of the actual project phase, because we got many requests from scholars and projects to support BiDi. One of those projects is "RIR — Relationen im Raum" (https://dev2.dariah.eu/wiki/display/RIRPUB/RiR), which uses XML as an exchange format with "epidat", a database of Jewish epigraphy (epidat — epigraphische Datenbank). The typesetting engine of *XML-Print* shall be used to automatically generate PDF documents with mixed Hebrew and German inscriptions and synoptically typeset German translations.

The Unicode standard provides the Unicode Bidirectional Algorithm (UBA) (Davis) to display BiDi texts correctly making it the industry standard for displaying such texts. The algorithm is well known and extensively tested. XML-Print therefore will base its support for Hebrew and Arabic texts on the UBA.

2.4 Testing and Quality Assurance

Testing is undoubtedly an important factor for any software project, but assuring the quality of a typesetting engine provides some additional challenges. As pointed out earlier, algorithms addressing different functions such as footnotes or bidirectional text are already very complex. Orchestrating multiple algorithms to produce a single document reaches a level of complexity, for which no developer can ultimately consider all side-effects and interoperability problems. Moreover, problems caused by errors in the source code often have minimal visual effects to the result so that a human tester cannot easily assess errors.

Therefore we have developed an automated system for functional testing. These functional tests are based on reference documents of varying complexity. For each test run, these documents are generated anew using the current version of the typesetting engine. In the next step each page of each document is split into regions. All regions are compared to the corresponding regions of the reference document pixel by pixel. This pixel comparison is performed using the Sikuli engine (Sikuli Script), which, in addition to comparing screen images, can be easily extended to display the proper documents and to simultaneously flip the pages of these documents. If any mismatch is detected between the generated and the reference document, a difference-image is created containing the differing section. Further, this image is enhanced with fragments of both documents to make it easier for a developer to recognize the problem (cf. Figure 3).

Figure 3:
Example of a pixel comparison of two documents.

3 Summary and Outlook

XML-Print tackles challenges on a number of levels:

Handling the requirements of advanced scholarly typesetting, notably the handling of multiple apparatuses and bidirectional texts
Responding to opportunities offered by heavily parallel hardware
Guaranteeing consistent quality and absence of regressions in a very much visual domain

As we have seen, the project responds to these challenges by

Creating a user interface that guides users through the complexity of advanced requirements
Opting based on F# for a functional design of the typesetting engine that can easily be parallelized
Developing testing techniques to automatically compare typesetting results in PDF to manually validated example documents

XML-Print's version 1.0 release is already a viable option for simple typesetting needs. However, much remains to be done during the remaining project duration. In addition to fine-tuning the existing implementation for scholarly apparatuses and bidirectional texts, *XML-Print* still has to support multiple columns, synoptic typesetting, running headers and footers, interactive corrections and manual overrides to the automatic results, e.g. to manually set line breaks or hyphenation. New requirements and changes in the overall of features also result from direct user feedback, e.g. regarding horizontal alignments of phrases in addition to classical synoptic editions. These requests are themselves sign of an increasing and encouraging take-up of *XML-Print* amongst early adopters, keen to introduce the software into their projects.

References

Sievers, M., Burch, T., Küster, M. W., Moulin, C., Rapp, A., Schwarz, R., Gan, Y. (2012). *XML-Print: an Ergonomic Typesetting System for Complex Text Structures.* In Hamburg, S. *Digital Humanities. Conference Abstracts.* 375–379.

Wie Kritik zur Kunst wird. Project website. http://www.avl.uni-wuppertal.de/forschung/projekte/wie-kritik-zu-kunst-wird.html [accessed 4 Mar 2013].

Bellman, R. (1954). *Dynamic Programming and Modern Control Theory*. Princeton, New Jersey.

Anders, B. (2006). Extensible Stylesheet Language (XSL) Version 1.1. *W3C Recommendation*. http://www.w3.org/TR/xsl11/ (accessed 4 Mar 2013).

Relationen im Raum — Visualisierung topographischer Klein(st)strukturen. Project website. https://dev2.dariah.eu/wiki/display/RIRPUB/RiR (accessed 4 Mar 2013).

epidat — epigraphische Datenbank. Project website.

Davis, M. Unicode Bidirectional Algorithm. http://www.unicode.org/reports/tr9/ (accessed 4 Mar 2013).

http://www.sikuli.org/ (accessed 4 Mar 2013).

Notes

1. XSL-FO+ is an extension of the XSL-FO standard [4] to meet the requirements of scholars in the Humanities. It has been especially designed for XML-Print.
2. This is a prominent example for extending the XSL-FO standard, which does not offer anything for using apparatuses yet.

Representing Materiality in a Digital Archive: Death Comes for the Archbishop as a Case Study

Lavin, Matthew
matthew-lavin@unl.edu
University of Nebraska-Lincoln, United States of America

A central concern of digital humanities has been how satisfactorily a digital transcription or facsimile represents its object of study. Dino Buzzetti, noting that "every form of text representation entails the implicit or explicit assumption of a model," has stressed the importance of a clearly defined digital text model to define a threshold for digital representation and critical study.[1] Sarah Werner, in a related turn, has asked what happens if "we move away from reading text to studying the physical characteristics of text, characteristics that can reveal important information about the content of the text and the cultural and historical creation of the artifact."[2] Werner is particularly concerned with large-scale digitization projects' inability to represent works with physical features integral to their interpretation. Andrew Jewell and Amanda Gailey, in their introduction to the journal *Scholarly Editing*, echo her concern with "quick

and dirty automated methods" that "digitize vast quantities of texts" but invariably create "shortcomings in metadata, accuracy, representation of compositional and publication complexities, and annotation."[3] Integral to all of these interventions is a distinction between the material form and linguistic content of print and manuscript material, and a desire to create digital archives that bring audiences closer to both.[4]

Two prominent initiatives, FRBR and TEI, exemplify efforts to address these kinds of concerns. The first has its origins in the International Federation of Library Associations and Institutions (IFLA), which authored a study in 1990 "to delineate in clearly defined terms the functions performed by the bibliographic record with respect to various media, various applications, and various user needs."[5] The result of this study was a report on Functional Requirements for Bibliographic Records (FRBR), released in 1997 and updated as recently as 2009. The Text Encoding Initiative (TEI) evolved from previous efforts to make texts machine readable through standardized markup practices. TEI states as its "chief deliverable" a set of guidelines "to represent all kinds of textual material for online research and teaching."[6] Particular communities within TEI such as the Manuscript Description Task Force, the Physical Bibliography Work Group, and the Work group on Genetic Editions have established specialized approaches for the markup of particular bibliographical and book historical data.[7] TEI and FRBR share a vested interest in the responsibility of representation. Whereas TEI markup represents a mix of linguistic representation and bibliographic information, FRBR attempts to create hierarchies to differentiate record-level bibliographical attributes.[8]

I am developing a small-scale mark-up and metadata approach that reflects the strengths of TEI and FRBR.[9] The strength of such an approach would be its applicability to items with noteworthy physical and/or bibliographical features. I have created the structure for a relational database that integrates FRBR-inspired metadata with a collection of digital texts, which will eventually include digital facsimiles and transcriptions. My paper will discuss my continuing project for the University of Nebraska-Lincoln Center for Digital Research in the Humanities to provide users with a dynamic, visually-rich, and critically nuanced history of Willa Cather's *Death Comes for the Archbishop* (1927) as a set of different material objects in multiple forms, including but not limited to manuscripts, notable editions, notebooks, translations, and interviews. My project has the particular goal of advancing knowledge about the creation, production, distribution, and reception of *Death Comes for the Archbishop*. It is also a test case in creating digital representations of print culture artifacts, textual variances, and bibliographical relationships among items. Central to my presentation will be a discussion of questions of form and content a project like this one raises:

1. What is the minimum baseline for representing the materiality of digital facsimiles and transcriptions? What is the optimum standard?
2. How successfully have efforts such as TEI and FRBR offered digital text models for different kinds of materials?
3. Do we need a better understanding of potentially significant bibliographical lines of inquiry in order to make these decisions?
4. Is materiality essentially a cataloging/records management problem, a mark-up problem, both, or neither?
5. To what extent will advances in interoperability improve book historical and digital humanities scholarship?
6. What is the optimum relationship between large-scale digitization efforts and small scale projects of scholarly interest?

My presentation will engage with these questions and report on the challenges I have encountered in this process and explain some of the decisions associated with the project. I will compare and contrast my work with some of the approaches to book historical questions taken by significant digital projects, including but not limited to the Modernist Journal Project, Radical Scatters, The Walt Whitman Archive, and The Digital Scriptorium, and Folger Digital Texts. I will also compare my approach to other experiments in FRBRization.

Notes

1. **Buzzetti, D.**, Digital Representation and the Text Model *New Literary History*. 33 (1). 61. Buzzetti differentiates between the "form of the information's expression"—the data representation—and the "form of the information's content"—the "data model" (pp. 65-66). Markup has aspects of both. His central point rests on the conclusion that "the correct use of markup and the adequacy of digital representation presuppose … recourse to a suitable text model" (p. 84).
2. **Werner, S.** (2012). *Where Material Book Culture Meets Digital Humanities.* 'Geographies of Desire Conference'. held 27-28 April 2012 at University of Maryland, College Park. http://sarahwerner.net/blog/index.php/2012/04/where-material-book-culture-meets-digital-humanities/
3. **Gailey, A. and Jewell, A.**(2012). Editors' Introduction to the First Issue of *Scholarly Editing: The Annual of*

the Association for Documentary Editing. 33. http://www.scholarlyediting.org/2012/essays/essay.v33intro.html
4. **McGann, J. J.**(1991). *The Textual Condition*. Princeton: Princeton University Press). 16. In Jerome J. McGann's terms, interpreting a textual object requires delineation between its linguistic and bibliographical codes.
5. **International Federation of Library Associations and Institutions.** (1997). *Functional Requirements for Bibliographic Records, Final Report.* As amended and corrected through February 2009, p. 2.
6. TEI: Text Encoding Initiative Home Page, http://www.tei-c.org/index.xml
7. Bibliographical data, manuscripts' physical layout, and text topography fall under what McGann classifies as bibliographical codes.
8. TEI markup describes linguistic text structure and physical object properties with hierarchizing them per se. The proportion of each category differs from case to case, as TEI offers different guidelines for different types of texts, as well as a range of established customizations and recommended practices for establishing new customizations. See http://www.tei-c.org/Guidelines/Customization/ .
9. FRBR is a bibliographical ontology but does not have a prescribed implementation scheme. Recognizing its potential, several projects have attempted to "FRBRize" their catalogues with varying degrees of success. The Resources Description and Access (RDA) standard is influenced by FRBR groupings. Perseus has a "FRBR-inspired" catalogue. Indiana University has piloted "Variations as a Testbed for the FRBR Conceptual Model" as a digital project. For an initial analysis of compatibilities between TEI and FRBR, see Kevin S. Hawkins, FRBR Group 1 Entities and the TEI Guidelines TEI Annual Members Meeting, held 6–8 November 2008, London http://www.ultraslavonic.info/preprints/20081102.pdf

Lexomics: Integrating the research and teaching spaces

LeBlanc, Mark D.
mleblanc@wheatoncollege.edu
Wheaton College (Norton, MA) United States of America

Drout, Michael
mdrout@wheatoncollege.edu
Wheaton College (Norton, MA) United States of America

Kahn, Michael
mkahn@wheatoncollege.edu
Wheaton College (Norton, MA) United States of America

Herbert, Alicia
herbert_alicia@wheatoncollege.edu
Wheaton College (Norton, MA) United States of America

Neal, Richard
neal_richard@wheatoncollege.edu
Wheaton College (Norton, MA) United States of America

Integrating research and teaching is exciting, time intensive, and a prescription for energizing faculty and students. We present outcomes of a six-year effort in multidisciplinary collaboration centered on the digital humanities as experienced in our teaching and research. Rooted in a set of "connected" courses between English and Computer Science (LeBlanc, et al. 2010) and three summers of NEH-funded research, our Lexomics Research Group has developed a modest set of web-based applications for scholars of digitized texts. We report here on the iterative development of the open-source toolset, how scholars both in and outside our group have used these tools to make significant discoveries, and perhaps most important how our research and teaching collaborations introduce a spirit of experimentation to the digital humanities.

Our current website is both a repository for our tool set as well as an evangelistic platform and teaching resource: http://lexomics.wheatoncollege.edu. We continue to develop online tools for three independent, but logically connected functions that lead scholars through the steps needed for performing hierarchical cluster analyses of texts and/or sections of texts. At this point, our cluster analysis tools are more narrowly focused than other toolsets, *c.f.* Voyant Tools and the data-intensive flow execution environment of Meandre. Our *scrubber* tool (PHP, CSS) accepts texts in multiple formats (.txt, .html, .docx) and handles preprocessing steps including stripping tags, removing stop words, and applying lemma lists. A second tool, *diviText* (ExtJS, PHP), accepts the output from *scrubber,* cuts texts into "chunks" in one of three ways (fixed size chunks, a specific number of chunks, and/or by manually selecting locations between words for chunk breaks), computes word counts within each chunk, and allows users to merge chunks. The latter functionality has proved valuable for generating "virtual manuscripts", that is, joining sections from different manuscripts. A third tool, *treeView* (PHP, R) accepts output from *diviText,* performs a number of variants of hierarchical cluster analysis, and returns a dendrogram plot in .pdf or phyloXML format.

Based on feedback from scholars who are using our tools, the website now provides video and written tutorials to help new users get started. These tutorials have been especially valuable for introducing these tools to our undergraduates. In the spirit of evangelizing, our website offers a series of "best practices" videos, discussions and step-by-step diagrams that shred insight to the process of how textual analysis at this level of detail can lead to rich new questions. The instructional videos include "The Story of Daniel", a discussion of one of our initial successes when using the tools where we showed that lexomic methods can accurately characterize the structure and relationships of texts that are already known, for example, identifying *Genesis B* within the Old English *Genesis* and the section of *Daniel* that is paralleled in *Azarias* (Drout, et al. 2011). Other videos include: "How to Read a Dendrogram", "How to Create a Dendrogram", "How to Read a Ribbon Diagram", "Lexomics for Comparison", and "Lexomics for Source Detection". A much longer video, "Editions and Manuscripts," addresses the challenges of choosing between different kinds of editions that may exist for a text that is found in multiple forms.

We have made what we think are significant discoveries in a number of spaces, including *Beowulf,* the poems of Cynewulf, Anglo-Saxon prose, a few Old Norse sagas, and Modern English texts including the Harlem Renaissance play *Mule Bone* (by Zora Neale Hurston and Langston Hughes). Lexomics is both an excellent first step to augment traditional scholarship as well as a rich source of deep analysis.

For example, previous lexomic analysis of several Old English poems suggests that there is a connection between dendrograms with an isolated, single leaf and poems that have an external source for one subsection of the poem different from the source or sources of the main body of the poem. We find in the dendrogram of *Daniel* a single-leaf clade corresponding to lines 299–455 of the poem. This section includes parts of *Daniel* that have external Latin sources that are different from the source of the rest of the poem (the Latin Bible). Similarly, in the Anglo-Saxon poem *Christ III*, a single-leaf clade that represents lines 1350–1510 has its source in Sermon 57 of *Cæsarius of Arles* (lines 1379–1498), and a single-leaf clade in *Genesis A* (lines 1079–1256) is associated with the genealogical lists from Adam to Noah that give the lineages of both Cain and Seth (lines 1055–1252), material that, for at least some of its content, must have a source different from the biblical text. These relationships were already known to scholars, but our investigation of the Old English poem *Guthlac A* resolved a century-long critical controversy by demonstrating that a key section of this poem (when demons drag Guthlac to the mouth of hell) has a different proximate source than the rest of the poem and that *Guthlac A* therefore must have been composed after a separately circulating text similar in content to Vercelli Homily 23 (Downey *et al.,* 2012).

The toolset, instructional materials, and publications are obvious deliverables from our efforts. Yet, we submit that our collaborative experiences with faculty and undergraduate students are even more exciting and provide a significant use-case of how scholarship in the humanities is evolving from the stereotypical solitary scholar to a paradigm of community, collaboration, and experimentation (*cf.* Unsworth, 1997). In our recent NEH- and locally-funded summer experience, humanities faculty in particular were pleasantly surprised with the intellectual environment that emerged. We got a glimpse of what it must have been like to work at a place like Bell Labs when they were making daily discoveries. This kind of collaborative, fast-moving research is unfortunately largely unknown in the humanities.

So how to continue our own momentum as well as replicate a spirit of experimentation for others? Earhart (2010) rightly notes that "digital projects remain rare, often the product of tenacious participants rather than a *supportive academic environment*"(emphasis added). We submit that faculty (not administrators nor technologists in the library) are the prime drivers and change must begin with our syllabi. Robust working relationships in the lab are strongest after students have already applied new modes of thinking in the classroom; for example, the importance of exposing undergraduate humanities students to computational thinking: problem decomposition, algorithmic thinking, and the success *and failures* of experimentation. And we need not overplay the lab metaphor. Our image of the digital humanities lab need not include beakers and soapstone benches, rather, the "new lab" is a room filled with scholars from multiple disciplines and a whiteboard.

Even if we had discovered nothing during our past summers in the lab, the intellectual thrill of the research group would have been a major accomplishment that these students (and we faculty) will never forget. But in fact we made discoveries, so many that there were days when participating faculty got none of their own work done because we were so busy bouncing from student to student seeing what they had found. Most critically, the experience continues to shape the way we share our disciplines with new cohorts of students. The solitary scholar still has a role to be sure, but that is no longer sufficient for the multidisciplinary demands and rewards to be gained from collaborations in the digital humanities: in our teaching, to our research, and back again.

lexomics — The term was originally coined by Betsey Dexter Dyer and first appeared in *Genome Technology* (2002). Since then ''lexomics'' has appeared on the internet and in some publications without attribution. Some of these appearances could be independent inventions of the term.

References

Downey, S., M. Drout, M. Kahn, and M. D. LeBlanc (2012). 'Books Tell Us': Lexomic and Traditional Evidence for the Sources of Guthlac A. *Modern Philology* 110: 1-29.

Drout, M., M. Kahn, M. D. LeBlanc, and C. Nelson (2011). Of Dendrogrammatology: Lexomic Methods for Analyzing Relationships among Old English Poems, *Journal of English and Germanic Philology* 110: 301–36.

Drout, M., M. D. LeBlanc, and M. Kahn (2011-2013). "Lexomic Tools and Methods for Textual Analysis: Providing Deep Access to Digitized Texts." National Endowment for the Humanities–NEH PR-50112011.

Earhart, A. (2010). Challenging Gaps: Redesigning Collaboration in the Digital Humanities. In Earhart and Jewell (eds),*The American Literature Scholar in the Digital Age* Ann Arbor: University of Michigan Press. http://hdl.handle.net/2027/spo.9362034.0001.001.

Genome Technology (2002). "In the News." in *Genome Technology* 1(27), November 1, 2002.

LeBlanc, M. D., M. Gousie, and T. Armstrong (2010). *Connecting Across Campus.* 'Proceedings of the 41st SIGCSE Technical Symposium on Computer Science Education' held in Milwaukee, WI.

LeBlanc, M. D., M. Drout, and M. Kahn (2008-2010). "Pattern Recognition through Computational Stylistics: Old English and Beyond." National Endowment for the Humanities–NEH HD-50300-08.

Lexomics Research Group. http://lexomics.wheatoncollege.edu

Meandre http://seasr.org/meandre/

Unsworth, J. (1997). Documenting the Reinvention of Text: The Importance of Failure. *Journal of Electronic Publishing,* 3:2. http://dx.doi.org/10.3998/3336451.0003.201.

Voyant Tools. http://voyant-tools.org/

Automatic annotation of linguistic 2D and Kinect recordings with the Media Query Language for Elan

Lenkiewicz, Anna
anna.lenkiewicz@mpi.nl
Max Planck Institute for Psycholinguistics, The Netherlands

Drude, Sebastian
sebastian.drude@mpi.nl
Max Planck Institute for Psycholinguistics, The Netherlands

Abstract

Research in body language with use of gesture recognition and speech analysis has gained much attention in the recent times, influencing disciplines related to image and speech processing.

This study aims to design the Media Query Language (MQL) (Lenkiewicz, et al. 2012) combined with the Linguistic Media Query Interface (LMQI) for Elan (Wittenburg, et al. 2006). The system integrated with the new achievements in audio-video recognition will allow querying media files with predefined gesture phases (or motion primitives) and speech characteristics as well as combinations of both. For the purpose of this work the predefined motions and speech characteristics are called *patterns* for atomic elements and *actions* for a sequence of *patterns*. The main assumption is that a user-customized library of patterns and actions and automated media annotation with LMQI will reduce annotation time, hence decreasing costs of creation of annotated corpora. Increase of the number of annotated data should influence the speed and number of possible research in disciplines in which human multimodal interaction is a subject of interest and where annotated corpora are required.

Introduction

The development in the area of audio-visual-recording devices leads to increase of the number of high performance hardware, enabling studies based on media recordings. In order to analyze a recording, the corpus needs to be annotated. The ideal solution would be an automated annotation system, which is a challenge for software developers. The algorithms need to not only retrieve points of interest from 2D and 3D recordings, but also to interpret them and to allow users to add extra interpretation, depending on the subject of study.

The work on the MQL and the LMQI is trying to meet the expectations of researchers. The system design assumes no previous knowledge of any query or programing language, nor query interface. The query syntax is similar to the syntax of a natural language with assumption that the data output goes into the Elan tier.

The main requirement of the MQL is that the data received from the recording contains time information, allowing alignment of the tier with the recording. The assumption is that any software retrieving information

form the recording may be integrated with the MQL. At the current stage, the algorithms integrated with the system are the recognizers (Lenkiewicz, et al. 2011) delivering time aligned coordinates from 2D video recordings.

The LMQI was built to simplify the work with the MQL.

The Media Query Language

Recently several automated annotation tools and techniques for deriving metadata (Hansen, et al. 2007; Park, et al. 2007; Chia-Han, et al. 2001; Crestani, et al. 2004; Rui Peng, et al. 2010) have been developed. The study often concentrates on joining syntactic and semantic levels of analyzed recordings. The work on the MQL is based on the premise that in order to lead research, the researchers themselves need to decide whenever a given phenomenon carries semantic meaning. To meet this requirements, there has to be a tool able to formulate a query describing elements of a tested theory, and use it on recordings obtained during experiments.

The structure and the syntax of the MQL are already defined. To build a compiler for the language, "SableCC compiler — compiler" has been chosen. According with the SableCC (Gagnon and Hendren 1998) requirements, the syntax is written using context free grammar rules and a parser is created, which is a Look-Ahead LR (1) (LALR) with one token of look ahead (Puntambekar 2010). SableCC was chosen as a tool to implement the language as it separates syntax and semantic actions of the new created language, shortening development time and significantly simplifying changes, thus improving maintainability of the system.

The hardware used to obtain the linguistic recording determines the software, which may be used to identify the human body parts in 2D or 3D space, further determining the number of elements that may be described. The MQL is designed in such a way that modification of parts of the syntax is possible and relatively easy. Adding new data source algorithms and new body or speech tokens is possible; the only requirement is that obtained data needs to provide:

- body token: spatio-temporal information of points of interest (2D or 3D coordinates of new detected body part in the time domain)
- speech token: the data relevant to the element in a time domain.

The interaction with the MQL is done thought the LMQI.

Thanks to its expressiveness, the MQL allows identifying movement and speech characteristics. The work with the MQL starts from "building" *patterns*. The MQL allows creating universal patterns out of motion and speech primitives; including elements such as e.g.: left and right hand, head, joins retrieved by Kinect(like neck, elbow), eye(s), mouth, preparation, stroke, fingers, loudness, peak, range, utterance, prosodic unit and silence, etc. In the level of pattern creation, each of them can be specified accordingly:

- motion: direction of movement, angle, speed, relation of body parts and distance between them(example left hand > 20 pixels from eye; left hand <50 pixels from right hand; LH stroke 30 pixels from mouth), duration, and one body part and/or other body part described by mentioned descriptors
- speech: range, mean, duration and behaviour of sound wave (e.g. falling, level, raising with numerical description).

The choice of such elements was done after research carried out two fields:

- body language study with focus on gestures, mimic and sign language
- the study of available and promised tools for human movement detection in 2D and 3D recordings. Only open source software was taken in consideration.

On the level of *action*, the user will be given the possibility to "advance" created *patterns* and query them on the recording (in case when atomic element of human behaviour is the subject of interest) or join more than one pattern in a set. The set of *patterns* can also be describe with more specific conditions:

- General: one can assign a person detected in the recording and the action will be found only if done by the person; relation between patterns can be described in detail (example: **pattern1 after min 20 ms pattern2**), duration of the whole action, etc.,
- For motion: speed, place of the gesture in a gesture space.

To query the recordings, the *action* should be used. The query has a form `QUERY {ANNOTATE A.Hand_up by Carl.Smith to (tier1,childTier) WRITE "This is annotation" direction duration;}` where only `ANNOTATE A.Hand_up to (tier1);` is the obligatory part and the other descriptors are optional.

Linguistic Media Query Interface

The LMQI is an interface allowing working with the MQL. At the current stage of the development characteristics of the LMQI are:

- Window of the query environment is divided into parts displaying a main query window, a library preview, an info text, an syntax error tracker, and the MQL options panel
- The MQL options panel contains fields for:
  - Inserting into main query window fixed parts of the code.
  - Selection of the source data (indication of hardware, possibility of selection of new/additional data capturing algorithm).
  - Selection of format in time domain (frames, milliseconds, seconds or minutes).
  - Place where new created library of elements needs to be saved and the place from which existing should be added to the library preview.
  - Advanced options allowing to add to the MQL syntax new tokens.

At the current phase of development the system checks the syntax of the language and advises correct tokens in case of syntax errors.

For pattern matching simple algorithm for numerical interval matching with a tolerance for match is used. The tolerance may be changed by the user and specified for each single query.

Future Work

Currently the research is concentrated around new methods of data retrieval and new ways of data matching. The Kinect data retrieving algorithms and available software are under implementation and integration. For speech recognition the usage of Praat is considered. The Hidden Markov Model is studied as an option for the pattern-matching algorithm.

Conclusions

Although the research and the algorithms is in its early phase the development of the MQL and LMQI may change the way humanities researcher may carry their work on media resources. The system can find it usage in:

- so-called motor theory of speech perception, co-speech gesture
- sign language (place of gesture in body space and in relation to other body elements, speed, etc.)
- language acquisition studies
- variation in speech and gesture

The information conveyed by gesture can be in a visuo-spatial form even when the speaker's message is not visuo-spatial, therefore the interface could be used by non linguistic researchers in order to simplify research like:

- emotional state: Recognizing Human Emotions from Body Movement and Gesture Dynamics (Castellano 2007)
- teaching: Video Annotation Tools Technologies to Scaffold, Structure, and Transform Teacher Reflection (Rich and Hannafin 2009),
- events monitoring Automatic Annotation of Humans in Surveillance Video (Miyamori 2003) .

Acknowledgement(s)

This research has received support from the EU 7th Framework Program under a Marie Curie ITN, project CLARA.

References

Lenkiewicz, A., M. Lis, and P. Lenkiewicz (2012). Linguistic concepts described with Media Query Language for automated annotation. in *Digital Humanities 2012*. Hamburg.

Wittenburg, P., H. Brugman, A. Russel, A. Klassmann, and H. Sloetjes (2006). ELAN: a Professional Framework for Multimodality Research. in *The International Conference on Language Resources and Evaluation (LREC)*.. GENOA — ITALY.

Lenkiewicz, P., P. Wittenburg, O. Schreer, S. Masneri, D. Schneider, and S. Tschöpel (2011). Application of audio and video processing methods for language research. in *Supporting Digital Humanities 2011 [SDH 2011]*.. Copenhagen, Denmark.

Hansen, D.M., et al. (2007). Automatic Annotation of Humans in Surveillance Video, in *Proceedings of the Fourth Canadian Conference on Computer and Robot Vision*, IEEE Computer Society. 473-480.

Park, K.-W., et al. (2007). OLYVIA: Ontology-based Automatic Video Annotation and Summarization System Using Semantic Inference Rules, in *Proceedings of the Third International Conference on Semantics, Knowledge and Grid*, IEEE Computer Society. 170-175.

Chia-Han, L., and A.L.P. Chen (2001). *Motion event derivation and query language for video databases.* Vol.

4315, Bellingham, WA, ETATS-UNIS: Society of Photo-Optical Instrumentation Engineers. 208-218.

Crestani, F., J. Vegas, and P. D. L. Fuente (2004). *A graphical user interface for the retrieval of hierarchically structured documents.* Inf. Process. Manage. 40(2): 269-289.

Rui Peng, A.J. Aved, Kien A. Hua, (2010). *Real-Time Query Processing on Live Videos in Networks of Distributed Cameras.* IJITN. 2(1): 27-48.

Gagnon, E. M. and L. J. Hendren.(1998). *SableCC, an Object-Oriented Compiler Framework.* in *Proceedings of the Technology of Object-Oriented Languages and Systems.*. IEEE Computer Society.

Puntambekar, A. A. (2010).*Compiler Design*: Technical Publications.

Castellano, G., S. D. Villalba, and A. Camurri (2007). *Recognising Human Emotions from Body Movement and Gesture Dynamics,* in *Proceedings of the 2nd international conference on Affective Computing and Intelligent Interaction*, Springer-Verlag: Lisbon, Portugal. 71-82.

Rich, P. J., and M. Hannafin.(2009). *Video Annotation Tools: Technologies to Scaffold, Structure, and Transform Teacher Reflection.* Journal of Teacher Education. 60(1): 52-67.

Miyamori, H. (2003). *Automatic annotation of tennis action for content-based retrieval by integrated audio and visual information,* in *Proceedings of the 2nd international conference on Image and video retrieval*, Urbana-Champaign, IL, USA: Springer-Verlag. 331-341.

Document classification based on what is there and what should be there

Levy, Noga
nogaor@gmail.com
Tel-Aviv University, Israel

Wolf, Lior
wolf@cs.tau.ac.il
Tel-Aviv University, Israel

Stokes, Peter
peter.stokes@kcl.ac.uk
King's College London, UK

Introduction

Some of the key questions in paleography are those of classification, namely trying to ascertain when and where a given manuscript was written, and — if possible — by whom. Paleographers bring many skills and tools to bear on these questions in what is often a complicated and laborious task requiring reference to paleographic, linguistic and archaeological data, among others. Because it is difficult to quantify the degree of certainty in the final readings and assessments, or even to articulate the arguments underlying these readings, experts have begun to develop computer-based methods for paleographic research in which the description of the various findings is made explicit (Ciula, 2005; Stokes, 2008; Aussems and Brink, 2009; Hofmeister et al., 2009).

Some of these computer-based approaches involve little or no human intervention. However, others require manual selection of regions in the image or manual recording of descriptors, that is, of features in the handwriting which are considered significant (Ciula, 2005; Stokes, 2008). Evaluating the significance of the features can be improved using statistical analysis (Levy et al., 2012). Such manual selection raises a key challenge in any system of descriptors, namely that of attribute repeatability among documents of the same category. Would two different people necessarily record the same descriptors for a given sample of writing? Surely some significant features would then be overlooked? If so then what are the implications, both for the accuracy of the results and for the perceived "objectivity" of the method. A descriptor that is marked as existing in a document is likely to exist; however, a descriptor might be unmarked due to an omission or simply because it is not present in the part of the manuscript that is available for inspection. Moreover, even very discriminative descriptors (those which are very important for distinguishing date, location or scribe) might not be present where expected due to scribal variance within the same location and date.

In order to overcome this challenge, we suggest a new statistical tool that allows us to hypothesize which attributes should be turned on — in other words, which attributes are likely to have been omitted due to the limits of selection — and then to perform classification on the augmented data. Our results demonstrate that this tool is effective in computer-based document classification.

Overview and results

A dataset consisting of scribal hands in English Vernacular minuscule, ca. 990 – ca. 1035, is used, where "scribal hand" here refers to a single stint or block of writing by one person (Stokes, 2005; Stokes et al., 2013).

These samples are spread across some 198 manuscripts and range from the main text of the book to later additions and notes or glosses between the lines or in the margins; they therefore can include anything from hundreds of pages to just one or two words.

The hands were described using 289 descriptors (Stokes, 2008), where each descriptor indicates whether a certain letter-form is present; more precisely, whether a grapheme (or group of similar graphemes) written as specific allograph(s) appear in the manuscript, as well as forms of certain parts of letters such as ascenders, descenders, and pen-angle. Every sample of handwriting is described by its known or predicted place of writing (where possible) and the estimated range of dates of writing. The date and localisation is based on external evidence wherever possible, or otherwise by an expert assessment of paleographical judgment (Stokes, 2005).

We focus on the samples whose place of writing is unknown but there is an educated guess to their origin, and try to verify their assumed place of writing. Overall, there are 67 such samples. The samples for which the place of writing is known, totaling 120 documents, serve as the training set. There are seven categories, such as Canterbury, Sherborne, and Worcester.

The baseline classifier we employ is the popular Support Vector Machine (SVM) classifier (Cortes and Vapnik, 1995) since it is known to be robust and to provide results that are often very close to the best obtainable, and because it outperformed other classifiers which we tried such as adaBoost (Freund and Schapire, 1995) and classification trees (Breiman et al., 1984). For each location-based category a model is learned by considering all documents which are known to belong to this location as the positive training set, and all other labeled documents as the negative training set.

Given a handwriting of an unknown origin, we apply all location-based models and compare the model producing the highest classification score with the assessment of the human paleographer (PAS). The obtained accuracy is 36%.

Next, building on our intuition that an unmarked descriptor might actually be present in a given handwriting, we employ a matrix-based method often used for imputing missing data. The method approximates the observation matrix (in our case, the size equals the number of descriptors times the number of documents) as a low-rank matrix using Singular Value Decomposition (SVD). The missing elements are taken directly from the corresponding elements of the approximation (Hastie et al., 1999). Retraining and employing SVM to the obtained descriptor vectors yields only a slight improvement in performance to 37%.

The imputed values are real-valued. We aim to choose an appropriate cut-off threshold for each of the categories. To this end, for every class, we rank all documents by the classification score obtained by the specific SVM model. Then, for each descriptor, we ask what would be the threshold least likely to occur by chance (see "Technical details" below).

Applying all the per-descriptor thresholds, a new set of binary exists/does-not-exist representations is obtained for each handwriting, and SVM-based classification is applied as before. This new method shows a remarkable increase in performance, and 49% of the documents are classified correctly. Examples of augmented representations are presented in Fig.1.

To further illustrate the effectivness of the method, we consider not just the first classification provided by the system, but the top three. SVM on the original descriptor vectors provides the correct answer as one of the top-three classes 78% of the time. Using the SVD based imputation method, the performance remains 78%. Finally, using the new method, the performance improves to 84%.

Technical details

The underlying method compares two ranked lists and returns the pair of thresholds which are the least likely to occur by chance. In our case, one list is a list of classification scores for a specific category, and the other contains imputed scores for a given descriptor. Both ranked lists are of the same length – n – which is the number of handwritings.

Let x and y be two vectors in $\mathbb{R}^n$. Applying a threshold to either vector divides the elements of this vector into two groups. A natural association between x and y would capture whether there exist thresholds such that the sets of obtained indices significantly overlap.

The hypergeometric distribution $f(k;n,i,j)$ captures the probability of obtaining a certain intersection size k between two sets X and Y of given sizes $i:=|X|$, and $j:=|Y|$, where the elements of the two sets are drawn randomly from the set $1...n$: $P(|X \cap Y| = k) = f(k;n,i,j)$.

To evaluate the statistical significance of a certain intersection size, we consider the probability of obtaining an intersection at least as large by random drawing from $1...n$ two sets of sizes i and j. To that end we employ the hypergeometric cumulative distribution function $F(k;n,i,j)$, which measures the probability of obtaining an intersection size of up to k: $F(k;n,i,j) = \sum_{c=0}^{k} f(c;n,i,j)$. The statistical significance we consider (probability of an intersection size of at least k) is therefore given by the tail probability: $G(k;n,i,j) = 1 - F(k-1;n,i,j)$.

Given a vector $x \in \mathbb{R}^n$ of unique values, there are $n + 1$ possible threshold-based subsets of the indices $1...n$, i.e., sets X such that for every $p \in X$, $x_q < x_p$ implies $q \in X$. Each

such subset is uniquely identified by its size. Denote these subsets by $X_0, X_1, \ldots, X_n$ such that $|X_i| = i$.

Considering also the vector y, ordered in a similar manner and giving rise to the ordered subsets of indices $Y_0, \ldots, Y_n$. Let $I \in \mathbb{R}^{n \times n}$ be the matrix such that $I_{i,j} = |X^i \cap Y^j|$.

We define the matrix P where $P_{i,j}$ is the probability of obtaining an intersection size of at least $I_{i,j}$ for sets of sizes i and j, when randomly drawing indices from $1 \ldots n$: $P_{i,j} = G(I_{i,j},n,i,j)$.

We seek thresholds whose values produce the minimal value of P, i.e., they produce the subsets of sizes I and j for which the following minimum is obtained: $min_{i,j} P_{i,j}$.

For n documents, a naive computation of the matrix I requires $O(n^3)$. This can be improved to $O(n^2)$ by considering the lists of indices obtained by sorting x and y.

Let C be the matrix defined such that $C_{i,j} = 1$ if the jth sorted index of y is in the first i sorted indices of x. C can be computed from $x,y \in \mathbb{R}^n$ in time and storage complexity of $O(n^2)$. The following lemma shows that I can be computed from C in a similar time complexity by performing cumulative sum over the rows of C.

Lemma 1. *For every $x,y \in \mathbb{R}^n$, and for C and I as above, $I_{i,j} = \Sigma^k_{c=0} C_{i,k}$.*

Once I is computed, P is readily evaluated based on the hypergeometric cumulative distribution function. An efficient algorithm is given in (Berkopec, 2007), which has as many iterations as $min(n - i, n - j)$. Using the identity $F(k;n,i,j)=1 - F(n - k - 1;n,m - i,j)$ (Riordan, 1968), the number of iterations can be further reduced to $min(n - i, n - j, i, j)$. Still, considering that $P_{i,j}$ is evaluated for all $i=1 \ldots n$ and $j=1 \ldots n$ this is computationally demanding for large n.

The following lemma can be used to reduce the number of evaluations of the hypergeometric cumulative distribution function. It states that by examining the elements of the matrix C around the location $i, j,$ we are able to determine whether $P_{i,j}$ can potentially obtain the minimal value out of all elements of P.

Lemma 2. *Given any vectors x and y, let C, I, and P be defined as above, then if $P_{i,j}$ is a minimal value of the matrix P the following two conditions hold: (a) $C_{i,j} = 1$; and (b) $j < n \Rightarrow C_{i,j+1} = 0$.*

Experimentally it is found that using lemma 2, between 75% and 85% of the entries of the matrix P need not be computed, where the larger n is, the higher the ratio of discarded entries.

Discussion

Descriptor-based approaches are a key component in shifting paleography from an authoritative discipline to an evidence-based one in which expert rulings can be explained. In an evidence-based approach, decisions should be based on descriptors in the manuscript which can be readily verified by other experts. It should be noted that the ability to rely on concrete evidence does not mean that classification accuracy is improved. The classification of the authoritative expert who is free from the need to explain herself would probably be at least as accurate, if not very much more. Thus, in order to achieve high levels of performance, it is crucial to have accurate decision rules and models on top of the descriptors.

It is also worth observing that none of these systems are truly objective. The premise of the approach taken here is that different people will inevitably make different decisions when selecting and recording descriptors: that the input data in any system is necessarily the result of selection and human decisions with everything that this entails. Indeed, the method outlined in this paper relies on an initial set of descriptors which have themselves been selected by experts, and so any bias in that original selection will necessarily be reflected in the descriptors which it predicts. Nevertheless, it does help to reduce the degree of variation when different people are entering data into a system, as normally happens in large projects in the Digital Humanities. As well as improving classification, it can also suggest descriptors that have been overlooked, and so project members who are entering the data can then go and check their work. In this respect the method applies much more widely than simply to paleography, since the problem of consistency in selection across a team is widespread.

Building on the observation that unmarked descriptors are occasionally missing for the "wrong" reasons, we are able to improve classification accuracy significantly. The method relies on several underlying assumptions that should be considered. First, by means of the low-rank approximation, the prediction of the missing descriptors is based on past correlations between the various descriptors. Therefore, a unique configuration of descriptors would be augmented to become a more conventional one, possibly losing valuable information. Second, by means of examining the correlations between descriptors and class memberships, our method assumes that the descriptors are discriminative. As a future direction we can apply our method more selectively, only to descriptors that appear (on the training data) to be informative.

References

Aussems, J. F. A., and A. Brink (2009). Digital palaeography. In Rehbein et al., 293–308.

Berkopec, A. (2007). Hyperquick algorithm for discrete hypergeometric distribution. *Journal of Discrete Algorithms*, 5(2):341–347.

Breiman, L., J. Friedman, R. Olshen, and C. Stone (1984). Classification and Regression Trees. Monterey, CA: Wadsworth and Brooks.

Ciula, A. (2005). Digital palaeography: Using the digital representation of medieval script to support palaeographic analysis. *Digital Medievalist*, 1(1).

Cortes, C., and V. Vapnik (1995). Support-vector networks. *Machine Learning*, 20(3):273–297.

Freund, Y., and R. Schapire (1995). A desicion-theoretic generalization of on-line learning and an application to boosting. In *Computational learning theory ,* 23–37). Springer Berlin/Heidelberg.

Hastie, T., R. Tibshirani, G. Sherlock, M. Eisen, P. Brown, and D. Botstein (1999). Imputing missing data for gene expression arrays. Technical report, Division of Biostatistics, Stanford University.

Hofmeister, W., A. Hofmeister-Winter, and G. Thallinger (2009). Forschung am rande des palographischen zweifels: Die edv-basierte erfassung individueller schriftzge im projekt damals. In Rehbein et al. (eds.), 261–92.

Levy, N., L. Wolf, N. Dershowitz, and P. Stokes (2012). Estimating the distinctiveness of graphemes and allographs in palaeographic classification. *Digital Humanities* (DH), 2012.

Rehbein, M., P. Sahle, and T. Schaßan (eds.) (2009). Kodikologie und Palographie im Digitalen Zeitalter - Codicology and Palaeography in the Digital Age. *Schriften des Instituts fr Dokumentologie und Editorik.* Books on Demand, Norderstedt.

Riordan, J. (1968). Combinatorial identities. *Wiley series in probability and mathematical statistics*. Wiley, New York [u.a.].

Stokes, P. (2005). English vernacular script ca 990 – ca 1035. Unpublished Ph.D. dissertation.

Stokes, P. (2007-2008). Palaeography and image-processing: some solutions and problems. *Digital Medievalist*, 3.

Stokes, P. A., S. Brookes, J. M. Vieira, S. Hugel, P. Caton, B. Caballero, R. Mullett, and N. Jakeman (2011-2013). Digipal: Digital resource and database of palaeography, manuscripts and diplomatic.

Notes

1. LW was supported by a personal research grant from Google Inc.

2. The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7) under grant agreement no. 263751.

Toward a Noisier Digital Humanities

Lingold, Mary Caton

marycaton@gmail.com
SoundBox Project, PhD Lab in Digital Knowledge, Duke University

Mueller, Darren

darren.mueller@gmail.com
SoundBox Project, PhD Lab in Digital Knowledge, Duke University

Trettien, Whitney

whitney.trettien@duke.edu
SoundBox Project, PhD Lab in Digital Knowledge, Duke University

Soundbox, a graduate student-led project funded by the new PhD Lab in Digital Knowledge, part of the Franklin Humanities Institute at Duke University, explores new ways of incorporating sound into digital scholarly productions. It does so primarily by hosting a series of "provocations" — that is, essays, experiments and events both digital and local — that challenge those who experience them to imagine a noisier form of scholarship. In this presentation, we will discuss the theory behind Soundbox's approach to Digital Humanities, as well as the transformative impact that de-centralized graduate co-learning structures have had on it.

Unlike many other Digital Humanities projects, Soundbox 1) has been conceived and implemented entirely by graduate students, and 2) does not intend to produce a tool but generate creative spaces that provoke new expressive forms. These spaces are by their nature hybrid, layering physical rooms (exhibit halls, museums, street corners) with digital connectivity in order to extend and expand what we imagine as possible fora for scholarly productions. They are also inherently multidisciplinary and inter-institutional, bringing together artists, audio engineers and scholars from universities, museums, public libraries and commercial studios. By fostering and encouraging these creative partnerships, Soundbox perceives its role primarily as facilitating and archiving sonic interventions in research and scholarship. That is, support for ideas comes before questions of technical implementation. While not

all provocations may be fully implementable, we see the imaginative potential of digital media as one of its most transformative contributions to the humanities — perhaps more transformative than the production of platforms and tools that contain scholarship within certain product-oriented formats.

In addition to discussing the current and future work of Soundbox, this paper uses our experiences with Soundbox as a case study in how to incorporate Digital Humanities into the graduate curriculum. At an October 2012 meeting of the Scholarly Communication Institute (SCI) on "Rethinking Humanities Graduate Education," Katina Rogers discussed results from a recent SCI survey on career preparation in humanities graduate programs, focusing on alt-ac training. While the sheer number of respondents proved the viability of and interest in alt-ac careers for recent PhDs, the survey's data showed that there continues to be a significant gap between the expectations of students entering PhD programs in the humanities and the realities of the academic job market (Rogers 2012). Moreover, even those students who had successfully negotiated into alt-ac positions reported feeling under-trained and ill-prepared by their departments for the often collaborative managerial skills required in non-teaching positions. As one respondent noted, "by far my most valuable experiences were the jobs I held while in grad school (which I kept hidden from my advisers)."

The final report from the SCI's meeting emphasized the need for centers to "work in concert with humanities departments to develop pilots of innovative research and pedagogy modes," pointing to Duke's PhD Lab in Digital Knowledge as a model (Rumsey 2012). As active graduate student members of the PhD Lab co-steering one of its current projects, Soundbox, we offer our honest, practical assessment of the promises and perils of such a model. On the one hand, it is difficult to underestimate the value of extra-curricular, extra-departmental collaborative learning environments. More than technical skills or the burden of more classroom instruction, students need flexible, unstructured spaces for learning from and collaborating with each other within and across institutions. Small seed grants and other untethered forms of funding, such as that which Soundbox has received through the PhD Lab at Duke, go a long way toward providing these opportunities by encouraging students to self-organize reading groups and field trips, or by allowing individual students to attend week-long gatherings like DHSI. On the other hand, by dissociating such activities from degree requirements set out by the student's home department, these valuable learning opportunities become additional burdens that potentially lengthen the time to degree. Extra-departmental centers need to work closely with departments to ensure that the activities students pursue in these centers contribute in meaningful, tangible ways to completing their coursework, their comprehensive exams and their dissertations. Similarly, departments need to inscribe the value of students' work in these centers into their degree requirements, especially for students who articulate a desire to pursue alt-ac careers.

Collaboration is a messy, even *noisy* process. As we cross disciplines and departments, the Soundbox collaboration has made visible for us the histories embedded within the institutional structures that define our roles within the university, and our progress toward earning a degree. It has also forced us to confront the cacophony of conversation in a way that puts pressure on what "interdiscipinarity" really means. In this way, the provocatively noisy product of Soundbox is inextricable from the processes of our collaboration. This paper addresses both of these topics.

References

Rogers, K. (2012). Outside the Pipeline: From Anecdote to Data. *Scholarly Communication Institute blog.* http://www.scholarslab.org/scholarly-communication-institute/outside-the-pipeline-from-anecdote-to-data/ (accessed 5 November 2012).

Rumsey, A. S. Rethinking Humanities Graduate Education. Final report of the Scholarly Communication Institute meeting, October 22-23, 2012. Maryland Institute for Technology in the Humanities. http://uvasci.org/wp-content/uploads/2012/09/final-report.pdf

Visualizing Centuries: Data Visualization and the Comédie-Française Registers Project

Lipshin, Jason
lipshin@mit.edu
Massachusetts Institute of Technology, United States of America

Fendt, Kurt
fendt@mit.edu
Massachusetts Institute of Technology, United States of America

Ravel, Jeffrey
ravel@mit.edu

Massachusetts Institute of Technology, United States of America

Zhang, Jia
zhangjia@mit.edu
Massachusetts Institute of Technology, United States of America

I. Overview

From 1680 to the present day, the Comédie-Française (CF), France's national theater troupe, has kept daily records of its repertory, box office receipts, and expenses, as well as additional information on set and costume design, actors' roles, and other matters. This wealth of information is a vital resource for theater scholars, literary historians, and those interested in the political, social, and cultural history of France more generally. Students of French theater and literature have long been interested in the performance practices and institutional functioning of the troupe that held a monopoly on the public performance of works by Molière, Racine, Corneille, and Voltaire in Paris before 1789. Historians have come to realize that the public theaters of Paris, especially the CF, vividly reflected the mounting social and political tensions of the time. Because the CF combined the rituals and concerns of the court, the ideas of the *philosophes*, and the everyday actions of working class Parisians in the same space, some have argued that the troupe played a central role in the negotiation of French national culture. And yet, the workings of the CF have been difficult to analyze. Access to the CF's archives in Paris has historically been extremely limited and the sheer volume of information quickly overwhelms traditional humanities research methods.

As an international collaboration between Hyperstudio — MIT's digital humanities research lab, MIT's department of history, and the Bibliotheque-Musee de la Comédie-Française, the Comédie-Française Registers Project (CFRP) seeks to provide access to and new ways to analyze such culturally significant data. In addition to creating an online database containing each daily receipt register of the Comédie-Française from 1680 to 1793, the CFRP also features a suite of interactive search and data visualizations tools, which allow for both filtering and complex analysis of information according to a set of parameters. Being able to apply different parameters, filter data, and see the results dynamically generated within a set of visualizations allows scholars to discover patterns and ask new kinds of research questions not possible without the tool. In this short paper, we will discuss the relationship between original sources, tool creation, and new research questions arising from visualization, in order to ask how quantitative analysis at unprecedented "levels of abstraction" (Witmore, 2012) might contribute to existing methodologies for historical research.

II. Current Implementation and Research Questions

Funded by grants from the Office of the Dean of MIT's School of Humanities, Arts, and Social Sciences and a number of other internal MIT sources, the CFRP has already made substantial progress on a number of levels. We have created a database of high quality digital copies of thirteen seasons of the daily registers (1780-1793) and created a number of search and interactive visualization tools based on this initial data set. Our custom faceted browser allows users to filter and view archival documents according to a number of parameters, including play title, author, genre, year, number of tickets sold, ticket price, location of seats within the theater, and whether a particular showing was a premiere, a first run, revival, or revue. This search tool is also directly integrated with a range of data visualization tools, which are manipulated either by changing parameters in the faceted browser or the visualization tools directly. These visualizations encompass layered histograms, heat maps, parallel axis graphs, and flexibly scaled timelines.

In experimenting with potential combinations of such parameters through the faceted browser, our team has already generated a number of initial research questions which we believe could reveal telling patterns through data visualization. Rather than simply using data visualization to demonstrate pre-existing conclusions, our methodology emphasizes the process of discovery in research, allowing interesting questions and patterns to emerge from the scholar's dynamic interaction with the faceted browser tool:

- Like most theatrical spaces of the time, the physical structure of the theater where the CF performed was highly stratified according to class, with tickets in the *loges* costing significantly more than those in the *parterre.* Each register contains information on the kinds of tickets sold for each performance, along with a diagram of the theatrical space. Thus, pairing the number of tickets sold in each seating section with the title of the play performed could reveal the popularity of certain titles or genres with particular demographics. We are also currently developing an interactive version of the theater diagram which would allow the user to toggle between different titles or individual performances, and see the relative socioeconomic makeup of specific audiences through a map of the physical space.

- Given that the CF registers range from 1680 to 1793, researchers can use the CFRP tools to study the effect of important political and cultural events like the French Revolution or the death of a king on the popularity of certain genres and authors, as well as general ticket sales. In terms of the latter question, experiments with our parallel axis graph have revealed large gaps in ticket sales in mid 1774, reflecting the death of King Louis XV.
- Researchers could also use this data to track how the popularity of emerging plays interacted with more established works. Thus, visualizations of the data could begin to ask such questions as: Were the great seventeenth-century tragedians Racine and Corneille diminished by the success of Voltaire's tragedies in the eighteenth century? Did the success of Voltairean themes indicate a new public sensibility in the Age of Enlightenment? Did audiences in the decades leading up to the Revolution prefer comic playwrights or tragic ones, and what might this tell us about pre-Revolutionary political culture?

According to the French theater historian Christian Biet, simply having the ability to answer such macro-level questions about the Comédie-Française could have a "transformational impact on the study of eighteenth-century French theatre." But we also believe that our project will be of interest to those outside this specialized field. Because CFRP's approach emphasizes dynamic interaction with visualization tools as a means to generate new research questions, we believe that our project could provide new insights and approaches to any scholar interested in the use of data analysis for historical research.

III. Conclusion and Future Directions

These major questions will receive new and more precise answers as we continue to create visualization tools for the analysis of CFRP data. As more scholars begin to explore this rich database, we believe that other interesting lines of inquiry will emerge, prompting us to tweak existing tools and create entirely new ones which support the needs of these domain experts. However, while we believe that simply having the technical ability to search through large amounts of documents at different levels of granularity, compare parameters, and control scales of analysis in data visualization is important in generating new scholarly insights, we also believe that such large-scale data crunching must be put into conversation with historians' existing practices of close reading. Following Dan Cohen's call in a recent essay on Google's n-grams to not fetishize the macroscopic in data visualization, the CFRP toolset is meant to facilitate a scholarly process of *toggling* between macro- and micro- scales: moving seamlessly "from distant reading to close reading, from the bird's eye view to the actual texts" (Cohen, 2010). By allowing scholars the ability to control the levels and parameters with which they engage with archival documents, the CFRP helps generate new questions for historians, while also relying on their expert knowledge to discern patterns and anomalies in the data. The principal justification for the CFRP, therefore, is that it will not only make data contained in the registers immediately accessible to anyone without access to the original sources in Paris, but also permit users to interpret data in ways which merge quantitative and qualitative approaches.

References

Witmore, M. (2012). Text: A Massively Addressable Object. In Gold, M. K. (ed.), *Debates in the Digital Humanities.* Minneapolis: University of Minnesota Press. 324-327.

Cohen, D. (2010). Initial Thoughts on the Google Books Ngram Viewer and Datasets. Dan Cohen's *Digital Humanities Blog.* http://www.dancohen.org/2010/12/19/initial-thoughts-on-the-google-books-ngram-viewer-and-datasets/ (accessed 11 November 2012).

eBook as Ecosystem of Digital Scholarship

Long, Christopher P.
longc@psu.edu
The Pennsylvania State University, United States of America

Socratic and Platonic Political Philosophy: Practicing the Politics of Reading (forthcoming Cambridge University Press) is an enhanced digital book that attempts to use digital media technology to cultivate the political practice of collaborative reading for which it argues.

The book's central argument is that there is an analogy between the ways Socrates practices politics with those he encounters in the dialogues and the ways Platonic writing turns us as readers toward ideals of speaking and acting capable of transforming our lives and the community in which we live. For Socrates, politics involves speaking words to individuals that will require them to turn their attention to questions of justice, beauty and the good, ideals that are at once alluring and yet always also elusive. For

Plato, politics involves writing words to readers that will require us to do the same. The book traces the practices of Socratic political speaking and Platonic political writing through five dialogues: *Protagoras, Gorgias, Phaedo, Apology* and *Phaedrus.*

This short paper will begin with a very brief overview of the basic philosophical argument of the book and the prior Digital Humanities scholarship from which it emerges. We will then turn to the technologies the enhanced digital book will deploy to further cultivate a community of readers capable of performing the politics of reading and writing for which the book itself argues. In the final third of the paper, deeper philosophical questions will be raised about the nature and limitations of authorial authority and about the connection between the book's design and the ideals of reading and writing for which it advocates.

The book begins with what I call an "Overture" designed to open a space into which the reader is invited to enter. This space is intended to extend beyond the physical book into an online digital community of dialogue both with me as the author of a reading of Plato, and also with others interested in the possibilities raised by the book. This online community already exists in the ecosystem of online digital spaces I have created over the past five years (via Twitter, Facebook, Google Plus, Flickr, YouTube, and my system of blogs called the Long Road)[1] . During this time I have sought not simply to push information to others about my scholarly work, but rather to engage in substantive digital dialogue with a wider community about issues in which we share an interest.

So a community of scholarly and educated readers has already been cultivated in the course of the writing of the book itself, because the research for it was facilitated by an ongoing public dialogue with colleagues who joined me in discussion on my podcast, the Digital Dialogue[2] . There are currently 57 episodes of the podcast[3] , eleven of which are explicitly referenced in the book itself. In producing these podcasts, I invited scholars to join me to talk about their work and in the course of our discussion, my work was enriched and I came away with new or deepened perspectives.

The publication of the enhanced digital book is designed to further cultivate and enlarge the influence of this community of readers. A book that argues for reading as a collaborative endeavor should be published in a way that performs and enables collaborative reading.

To do this, the Cambridge University Press and I are developing a dynamic enhanced digital book that will embed the audio of the eleven podcasts into the digital book itself, enabling readers to listen to the podcasts directly as they encounter them in the text. In order to cultivate a community of collaborative reading, the enhanced digital book will also enable the reader to make all highlighting and annotations public if desired. Those annotations and markings will then themselves generate a feed that interfaces with a blog plug-in like Comment Press or some other form of integration by which the annotations and highlights can appear in public in ways that are open to further response. Although readers might decide to publish the annotations to a preferred social media site, the annotations should also be accessible to a blog I manage and moderate so that I can respond to and engage with readers as they engage with the book itself.

The publication of this enhanced digital book is designed to facilitate an ongoing dialogue about the book, its ideas and the larger questions of what I call in the book the "politics of reading." As the conversation develops, I would envision recording new episodes of the Digital Dialogue podcast with readers who have had particularly insightful annotations or comments. Those podcasts too, if desired, could be made available in the enhanced digital book.

I envision the printed version of the book as another way to give readers access to this enhanced digital version and the community of dialogue to which it is intimately connected. We hope to include QR codes or some other method of moving the reader from the physical book to the online conversation. The hope is that the enhanced digital book will interface with existing systems of curated annotations as those found on the Kindle via Amazon.com and sites like Findings.com and Apple's iBookstore.

Finally, the question of authorial authority emerges as central to this project. On one hand, by inviting readers to become active participants in an ongoing conversation about the ideas articulated in the book, the enhanced digital book is designed to recognize and cultivate the hermeneutical imagination of its readers and open new perspectives on the text for the author. In this sense, the publication of the book is designed to open a site of ongoing scholarly dialogue in which the author is but one of multiple interlocutors. On the other hand, by affording the author ongoing opportunities to moderate and shape the discussion associated with the book, its mode of publication could be taken to further reinforce the centrality and authority of the author. A central question thus emerges: how can digital technologies be deployed that will ensure civil, deliberative dialogue without re-inscribing the hegemony of the author into the published text?

The ideals of reading for which the book argues — the importance of attentively caring for individuals as such, of cultivating ethical imagination, and of orienting our lives toward ideals of justice, beauty and the good — cannot simply be argued for, they must be put into practice. This short paper at DH2013 will be part of the larger attempt to engage an audience of interested scholars in order to further augment the community of active readers on which the success of the project ultimately depends.

Notes

1. See: http://www.la.psu.edu/chrislong
2. See: http://www.personal.psu.edu/cpl2/blogs/digitaldialogue/blog/
3. See: http://ets.tlt.psu.edu/wiki/Digital_Dialogue

Ontology and collaborative knowledge environment in Digital Humanities: the Cardano Case

Luzzi, Damiana
luzzi@rinascimento-digitale.it
Fondazione Rinascimento Digitale, Italy

Baldi, Marialuisa
marialuisa.baldi@unimi.it
Università degli Studi di Milano, Italy

1. Introduction

The *Girolamo Cardano Project: the Knowledge and the Arts of the Renaissance* [1] is devoted to a major author in Renaissance philosophy and science, although not yet fully studied. Girolamo Cardano(1501-1576) was a polymath, philosopher, mathematician, physician, astrologer, encyclopedist and autobiographer[2] . In this paper we focus on our experiment on one of his most important works in medicine, the *Contradicentia medica* [3] , according to the methodological approach of the semantic web ontology(Kotis, et al. 2010; Domingue, et al. 2011). The multidisciplinary nature of the text, its encyclopedic references and citations, its many sources, ancient and modern, explicit and implicit, make the *Contradicentia* a significant case study for the methodology adopted. Our aims are:

- making available online the digital edition of the *Contradicentia* and its transcription, together with other texts related to it
- providing a collaborative environment for editing, reading, studying, researching, and posting annotations and comments.

A semantic web ontology has been designed to tackle the complexity of the *Contradicentia,* reconstructing Cardano's ideas in medicine and in the philosophy of nature, and capturing knowledge about significant contextual information. Its application to the text of Cardano is an absolute novelty in Renaissance studies. The semantic web ontology is an excellent choice for representing Cardano's encyclopedic knowledge. The main characteristics of an ontology are:

- flexibility
- extensibility
- portability

it is shared and persistent over time, too.

The ontology and the semantic approach allow us to express the concepts drawn from the text and create links between them favoring the development, sharing, reuse, and updating of knowledge. The semantic web ontology can be published in the form of Linked Data (Eero, 2012; Heath et al., 2011) to facilitate sharing, interoperability and reuse of information.

Cardano himself, were he alive today, would be very interested in ICT and semantic web technologies, as he thought knowledge was a network[4] and that philosophers had to find hidden relationships between things and concepts.

2. Cardano's ontology

According to the OWL 2 (Motik, et al. 2009):

- the core classes represent Cardano's thought
- the "general" classes represent the knowledge base designed to express the information about the temporal and spatial aspects, the persons and groups, the cataloging of the texts and bibliography references. To manage this information we have implemented classes and predicates, taking into account the standards[5] , in order to make an open and flexible system available in which information, with a different granularity, can be integrated and continuously updated by the users themselves, once enabled.

As an example (Fig. 1), we represent a significant portion of the ontological model which describes the

evolution of Cardano's thought over time, as it is expressed in some portions of the text.

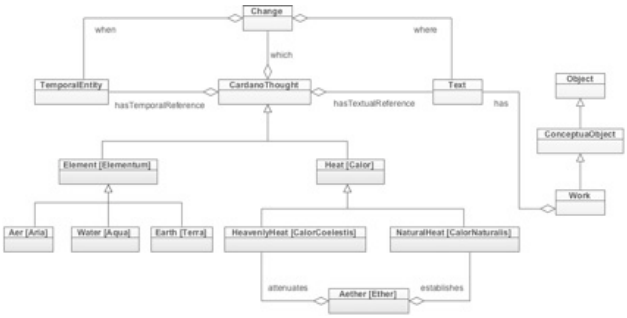


Fig. 1:
Basic description in UML (Unified Modeling Language); in brackets the Latin term used by Cardano.

Cardano's concepts (e.g. "Elementum") are contained in the superclass *CardanoThought.* The "lifespan" of any concept is defined by the class *TemporalEntity* linked by the property *hasTemporalReference.* The class *TemporalEntity* can be aligned by equivalence relation to the class with the same name within Time ontology [6] . To express the evolution of Cardano's concepts in time, related to a specific part of the text, the class *CardanoThought* is linked by the property *changes* with the class *Change,* and the property *hasTextualReference* with the class *Text.* In addition, to define when, where and which concept has changed, the class *Change* has been defined and linked by property:

- *when* to the class *TemporalEntity*
- *where* to the class *Text*
- *which* to the class *CardanoThought.*

The class *Work* can be aligned by equivalence relation to the class with the same name within FRBR-oo standard[7] .

3. Web Application

Reperio[8] , a collaborative knowledge environment for Digital Humanities and Science, indicates a technological solution to meet the needs of specific scholarly projects. It takes advantage of experience gained during the experimental project Pinakes [9] , of which it is an evolution; Reperio is used in some national and international projects[10] . Its vision and mission is to help eliminate isolation within different research communities by facilitating collaborative ways of working and sharing content and resources, while respecting the intellectual property of the individual scholar. Reperio is a multi-user, modular, collaborative, flexible and customizable web work environment consisting of two modules:

- Ontology Editor to edit classes and properties, and to insert instance in the ontology
- Text to edit and manage texts and images, which provides specialized tools:
  - text editor
  - digital image manager
  - automatic importer of texts/images
  - metadata, full-text and semantic search
  - annotation editor
  - comparison and collation manager
  - textual and linguistic analysis

The Ontology Editor and Text modules communicate in a dynamic way: any changes made to the ontology schema (such as the addition, modification, cancellation of a class or a predicate) is "instantly" visualized in Text. The text is thus connected to the ontological schema that it uses for the different types of annotations.

The markup is stored and can be displayed as XHTML, XML, RDF. The data is stored in a Sesame Triple Store. A user-friendly interface allows the user to perform SPARQL queries. The URI [11] identifies digital resources, too. Reperio's source code will be open-source.

4. Annotation

The Annotation Tool, developed on the basis of the standard Open Annotation Collaboration[12] , enables making various types of annotations on the text (and/or image): comments, links or cross-references to other resources, ontological annotations, etc. Users are allowed to select portions of the text and associate the tags proposed by the system or that they themselves enter.

In this context we present the collaborative Ontological Annotation Tool of Reperio, because it is the most widely used tool for the studies of *Contradicentia* and annotating is a core practice to scholars, too (Corcho, 2006). The expressive power hidden in the texts can be further maximized by combining ontology and annotation: annotation expresses, in a formal manner, the meaning of a text using the "terminology" provided by the ontology. Thanks to this type of annotation, users can study and analyze the text in different ways, philological, syntactic, morphological, grammatical. They can even comment on physical materials, inks, colors, etc.

Particularly, the semantic annotation (Agosti et al., 2007) helps to bridge the ambiguity of natural language in expressing notions and their computational representation

in a formal language. The annotation operation can be performed:

1. through concepts associating classes to the selected term/s. If the class relating to the concept on the text is not in the ontology, it can be easily inserted (by user with appropriate permission) opening the ontology editor.
2. through instances:
   a. the instances are already in the ontology: selecting (Fig. 2) *"Philosophus"* and connecting it to the instance *"Aristotle"* of the class Person and its subclass Philosopher(i.e. search for *"Aristotle"* shows the results where Aristotle is appointed *"Philosophus"* too, and a search for person or philosopher shows Aristotle in the result). Another example: Cardano writes "calor animalium secundum Philosophum est calor non igneus, sed coelestis" [13] : by semantically linking Cardano's citation with the instance *De longitudine et brevitate vitae* of the Title class of the ontology, you get the reference to Aristotle's text and its bibliographic information. Therefore a portion of the ontological schema is based on FRBR-oo standard. In addition, if Aristotle's book is present in a digital edition in Reperio or in another digital library, through a URI connection, you could read the page to which Cardano refers.
   b. directly populating the ontology: selecting *"aqua"e* and inserting it as instance of the class Elementum.

Annotation operations and/or the ontology population may be performed manually or in a semi-automatic manner by text parsing performed on the basis of the classes and/or instances.

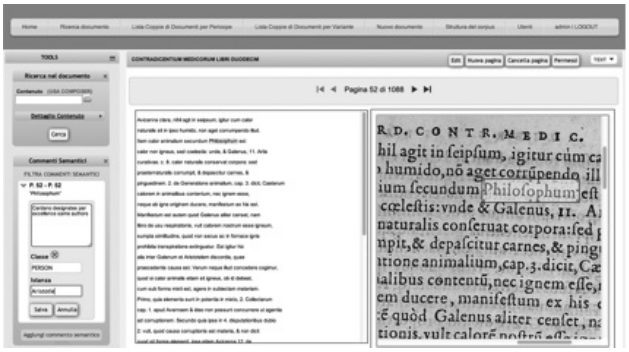


Fig. 2:
Reperio: Semantic Annotation

The semantization of knowledge significantly improves accuracy and relevance of search results. The annotation ontological tool also allows you to write additional information on the annotation, and then to perform a search on them, too.

5. Conclusion

Ontology and an environment system like Reperio can be considered an evolving open ecosystem, that offers scholars the freedom to search and explore. On the experience of the project, the process of designing the ontology has been very useful, because it offers different views and perspectives on texts and the concepts they have, and will open new ways for further study and analysis. Such an "enhanced" search allows you to infer and deduce new knowledge based on what is available.

References

Agosti, M., G. Bonfiglio-Dosio, and N. Ferro (2007). A historical and contemporary study on annotations to derive key features for systems design. *International Journal on Digital Libraries,* 8(1): 1-19.

Corcho, O. (2006). Ontology based document annotation: trends and open research problems. *International of Journal Metadata, Semantics and Ontologies.* 1: 47-57.

Domingue, J., D. Fensel, and J.A. Hendler (eds.) (2011). *Handbook of Semantic Web Technologies.* Berlin Heidelberg: Springer-Verlag.

Eero, H. (2012). *Publishing and Using Cultural Heritage Linked Data on the Semantic Web,* Palo Alto, CA: Morgan & Claypool.

Heath, H., and C. Bizer. (2011). *Linked Data: Evolving the Web into a Global Data Space.* Palo Alto, CA: Morgan & Claypool.

Kotis, K., and G. Vouros (2010). Ontological Tools: Requirements, Design Issues and Perspectives. In Poli, R., M. Healy, and A. Kameas (eds.) *Theory and Applications of Ontology,* Netherlands: Springer, 155 -173.

Motik, B., P.F. Patel-Schneider, and B. Parsia (2009). OWL 2 Web Ontology Language Structural Specification and Functional-Style Syntax. In *W3C Recommendation,* http://www.w3.org/TR/owl2-syntax/ (accessed 5 March, 2013).

Notes

1. The project is coordinated by the University of Milan-Department of Philosophy http://dipartimento.filosofia.unimi.it/ (accessed 5 March, 2013) in collaboration with the Digital Renaissance Foundation

(FRD, http://www.rinascimento-digitale.it (accessed 5 March, 2013).
2. Cardano G. (1663). *Opera Omnia: tam hactenus excusa,* cura Caroli Sponi, 10 vols, Huguetan J. A., Ravaud M. A., Lyon, henceforth OO. The volume is available at: *Girolamo Cardano. Strumenti per la storia del Rinascimento in Italia settentrionale,* Baldi M., Canziani G. (edited by): *Opera Omnia,* http://www.cardano.unimi.it/testi/opera.html.
3. Cardano G. (1565). *Contradicentium Medicorum. Libri duo, quorum primus centum & octo, alte vero totidem disputationes continet,* Macaeus I., Parisiis.
4. Cardano G.: *De vita propria liber,* XLI, OO, I.
5. *The conceptual models used are: Time ontology in OWL to express the information about the temporal aspects* http://www.w3.org/TR/owl-time/ (accessed 5 March, 2013); Friend of a Friend (FOAF) to express information about persons and groups), http://www.foaf-project.org/ (accessed 5 March, 2013); Functional Requirements for Bibliographic Records-object oriented (FRBR-oo) to catalogue texts and bibliography references, http://www.ifla.org/publications/functional-requirements-for-bibliographic-records (accessed 5 March, 2013).
6. Cfr. note 5.
7. Cfr. note 5.
8. Developed by the FRD in collaboration with the Institute of Computational Linguistics "Antonio Zampolli" (ILC-CNR), http://www.reperio.it (accessed 5 March 2013).
9. It was developed from 2004 to 2011 by the FRD and the ILC-CNR, in collaboration with the MiBAC and the Galileo Museum.
10. See, for example: *Cataloging and Management of Digital Documents (Transcript of Texts and Images, Manuscripts, Books, Lectures, etc.)* of the Pontifical Gregorian University Archives in Rome; *DiTMAO: Information System for Old Occitan Medical Terminology; Bulletin (1893 - 1923) of the Società Dantesca Italiana.*
11. Uniform Resource Identifier, the same identifier at the basis of Linked Data.
12. OAC, http://www.openannotation.org (accessed 5 March, 2013) and Open Annotation Data Model (W3C draft, http://www.openannotation.org/spec/core/ (accessed 5 March, 2013).
13. G. Cardano, *Contradicentium medicorum libri,* cit., I, I, XI.

Should the Digital Humanities be taking a lead in Open Access and Online Teaching Materials?

Mahony, Simon
s.mahony@ucl.ac.uk
UCL Centre for Digital Humanities, United Kingdom

Tiedau, Ulrich
u.tiedau@ucl.ac.uk
UCL Centre for Digital Humanities, United Kingdom

The digital age has introduced new possibilities but also new problems for higher education. Academics within the same departments have always shared teaching materials but a cultural change is taking place in universities, with academics using the internet to share their research (Open Access) and teaching and learning resources (OER: Open Educational Resources) more widely. This spirit of collaborative working is increasing, and potentially opens up higher education to a wider and global market, giving students and teachers greater access and flexibility. Education for all has taken on a new meaning in the digital age and the true rationale of Openness is one of reclaiming original academic practice and collaboration; consequently the move towards openness extends beyond resources and also increasingly includes Open Educational Practices, or just Open Education. How do new initiatives such as OERu, Open Text Books, and iBooks Author contribute to this? To change academic culture and to encourage open educational practices requires much more than technological changes. It will require an understanding of the challenges facing the educational community today and how OERs can help them achieve their goals in teaching and learning. This is particularly pertinent in the Digital Humanities where there is great emphasis on the teaching of the collaborative and communication skills that are requirements for our researchers and research projects as well as important job skills for our students that do not stay in the academy.

This paper draws as case studies on the experiences and publishes the results of two completed and one currently inprogress JISC funded projects for the creation, use, and importantly reuse of OERs. Firstly, *VirtualDutch* (www.dutch.ac.uk) where a lesser taught language (despite being the one most closely related to English) subject

community have collaborated in joint teaching projects. A wide range of Open Educational Resources has been developed since the start of the programme, including self-access reading skills courses, learner's grammars, online reference works and some 30 multimedia study packs for autonomous learning. These resources cater for various levels of linguistic competence, ranging from topics such as individual Dutch or Flemish authors, like Multatuli or Louis Couperus, to the sociolinguistic situation of Brussels and the multicultural society in the Netherlands today. These are also being used in a series of distancelearning programmes.

The second is *Digital Humanities Open Educational Resources* (DHOER: www.ucl.ac.uk/dhoer) which was set up to create and release a comprehensive range of introductory materials on approaches, topics, and methods in the digital humanities; these are based on modules currently taught as part of the Master's programme at the UCL Centre for Digital Humanities. Importantly, these resources go beyond the digital humanities sphere and support many cognate disciplines, including the whole spectrum of the arts and humanities, cultural heritage, information studies, library studies, computer science and engineering. Indeed by pushing the disciplinary boundaries DHOER has contributed considerably to the advancement of the OER idea and helped to start a movement to bring about the cultural change that the UKOER programme envisages.

The third project, OER CPD4HE/Sustexts (*OER Continued Professional Development for Higher Education, Sustainable Texts and Disciplinary Conversations*), a collaboration with HEDERA.org (HE Development, Evaluation and Research), recognises that subject discipline is a key part of academic identity and that narratives are important in both learning and professional development. It has set out to increase the reach of open practice and engage with institutional policy makers.

Each of these projects has been involved in the awarenessraising of OER, by presenting at workshops, conferences and organizing several UKOER programme- and institutionwide events. The focus for each is on building communities of users and contributors to ensure sustainability and to develop standards of best practice. We can gather download statistics simply enough, just as we can for journal articles and other academic resources, but that is no indication of whether or not they have actually been used or indeed reused as a teaching resource. How this might be achieved is one of the challenges addressed here. This paper develops themes introduced in a recent publication: 'Open access and online teaching materials for digital humanities', Warwick et al eds. (2012) *Digital Humanities in Practice*, charting the progress and results of new initiatives since that chapter was authored.

Digital humanities should be taking a lead in the development of open access online teaching materials. In our community we take openness and collaboration for granted and commitment to these principles should be central to the development of any of our teaching programmes. By making our teaching resources openly and freely accessable to all we make our ideas and methodologies available to the wider academic and research communities. It is in this way, by creating synergies rather than silos and making our educational practice, particularly our critical and methodological approaches to teaching and learning, available to others, that we will overcome the sometimes sceptical reaction to the value of the work that we do (see Bradley 2010, which is taken from a paper he gave at DH2009). Experience has shown that digital humanities teaching must be relevant to students' studies and research interests and we must be clear that what we offer are not 'skills training' courses (although they too play a part) but new methodologies and new ways to think about our material. Institutional barriers need to be overcome. Colleagues across the arts and humanities and other faculties need to be able to see how students benefit from the digital humanities approach, and then they will better understand our work and support the training of future digital humanities students and researchers (see Mahony and Pierazzo, 2013).

References

Bradley J. (2010). No job for Techies: Technical Contribution to Research in the Digital Humanities. In Deegan and McCarty (eds.) *Collaborative Research in the Digital Humanities,* Ashgate http://www.ashgate.com/isbn/9781409410683

Centre for Excellence in Teaching and Learning (CETL) Open Educational Resources (2010) Universities' Collaboration in eLearning (UCEL), http://www.ucel.ac.uk/oer10. DHOER: Digital Humanities Open Educational Resources http://www.ucl.ac.uk/dhoer

Hirsch (ed.) Digital Humanities Pedagogy: Practices, Principles and Politics, Open Book Publishers. http://www.openbookpublishers.com/product/161/digital-humanitiespedagogypractices principlesandpolitics

JISC Phase 1 (2010). Open Educational Resources Programme Phase one: JISC, https://cms.jisc.ac.uk/whatwedo/programmes/elearning/oer.aspx

JISC Phase 2. (2011). Open Educational Resources Programme Phase 2 http://www.jisc.ac.uk/whatwedo/programmes/elearning/oer2.aspx

JISC Phase 3. (2012). Academy/JISC Open Educational Resources Programme Phase 3 http://www.jisc.ac.uk/oer

Mahony, Tiedau, and Simons (2012). Open access and online teaching materials for digital humanities. In Warwick et al., (eds). *Digital Humanities in Practice*. Facet.

Mahony, and Pierazzo (2013). Teaching Skills or Teaching Methodology?, in Hirsch (ed). *Digital Humanities Pedagogy: Practices, Principles and Politics*, Open Book Publishers.

SustextsOER, Sustainable Texts & Disciplinary Conversations, http://www.ucl.ac.uk/calt/cpd4he/

Warwick, Terras, and Nyhan (eds). (2012). *Digital Humanities in Practice*. Facet. http://www.facetpublishing.co.uk/title.php?id=7661

VirtualDutch Open Educational Resource http://www.ucl.ac.uk/alternativelanguages/OER/

This is Not a Novel: Experimental Literature as Prototype

Mauro, Aaron
mauro@uvic.ca
University of Victoria, Electronic Textual Cultures Lab, Canada

Introduction

This short paper describes how digitizing experimental print texts can be used to innovate the aesthetics of the web, produce new models for human-computer interaction, and further theorize touch interfaces. Matthew Kirschenbaum has long identified the "tactile fallacy" often associated with digital media (43), but, as Sebastian Heath pointed out more recently at the 2011 Digital Humanities meeting, "'digital materiality' does not yet have a fixed meaning." By building upon work by Katherine Hayles, Alan Liu, Marlene Manoff, and others, this paper will help account for the material metaphors of interface design and how touch interfaces have become a central concern for a supposedly immaterial medium. Because of the advanced design techniques used by commercial web developers and the universalized standards and backward compatibility of HTML markup, popular manifestations of the web have a great deal to teach academic discourses about the aesthetics of reading and visualizing the materiality of texts. I turn, therefore, to contemporary experimental literature as a means of grappling with the materiality of print and testing the limits of presentation semantics in web markup. As a first step, this paper will present various web based prototypes of CSS styling to visualize Jonathan Safran Foer's decidedly print text *Tree of Codes* (2010).

At the core of my approach are two simple observations. First, since experimental authors have been pushing the limits of print and remaking traditional conceptions of the book for decades, any digital humanities discourse that presumes to revolutionize literature through a radical reshaping of its formal possibilities must also account for the long tradition of experimental literature. Second, contemporary fiction is published in an increasingly digital media environment, and contemporary authors are responding to this new context through a critique of digital publishing. As Elizabeth Eisenstein argues, the printing press has long been an agent of social and political change, but it is also becoming a tool for technological critique. In short, the purpose of my paper will be to show how this critique can be leveraged to produced innovative methods of web development.

Digitizing Experimental Literature as a Prototyping Framework

Specifically, this paper will present several style sheets in CSS 3 that attempt to digitize Foer's *Tree of Codes*. Since the failure of any perfect mimesis is a foregone conclusion, this failure demonstrates a method for producing new text forms but also offers a vantage point from which to make interpretations about the importance of materiality in Foer's text. Through a combination of 3D transforms and animations made possible with CSS3 and canvas elements, I will work to show the complexity of our still new and evolving web standards. Technological experimentation makes manifest the purpose of all cultural and literary experimentation. Because literary traditions perform and inform the history of the book along with a broader technological context, prototyping experimental literature in the populist medium of the web is most capable of formally bridging reading technologies. It is, I argue, the points of rupture between mediums that signals new directions for development and experimentation. Furthermore, this methodology exists at the threshold between electronic literature community and the digital humanities, while offering a means of reconciling these often disparate discourses.

Web Development and Literary Authoring

In a commercial context, developers like Mike Kuniavsky, Indi Young, Luke Wroblewski, Peter Merholz, Jeffrey Kalmikoff, and John Zeratsky have become the theorists of the ever emergent web. The tools that have thus far been created by engineers and designers must, however, incorporate a broader understanding of the

function and purpose of literature and literacy. It is the role of academically informed web designers to interpret contemporary design culture through the history of the book. It is now well understood that digital literacy must not relegate pedagogical practice to simply teaching a new user interface. Because of the great potential for "computers as modelling machines" (McCarthy 27), prototyping and making digital books is a fundamental methodology of the digital humanities and, as Alan Galey has described it, "the design ethos of thinking through making" (111). In the context of Jerome McGann's earlier claim that "The next generation of literary and aesthetic theorists who will most matter are people who will be at least as involved with making things as with writing texts" (19), I argue for literary authoring and web development to become a unified creative act. As Manoff has observed, "The content or text of a book cannot be separated from the physical object that houses it" (319). In a print context, the storage and delivery of content is linked absolutely, while authoring is a privileged external act. In a digital context, creative web development has the potential to link literary authoring and delivery into a single act, while storage and preservation becomes an external function of literary production taken up by digital humanists.

Platform Development and Repository Building

This paper represents only the earliest stages of a multi-year project to build an experimental literature web platform that hosts a repository of open source CSS modules and the tools to develop online literary projects. As a member of the multidisciplinary research team in The Electronic Textual Cultures Lab, I am currently seeking funds to mobilize the additional expertise required to complete a project of this size and complexity. Set within the context of the early prototyping and feasibility stages of a broader digital humanities project, this paper represents a template for the future implementation of this longer term development strategy with the technical support at the Humanities Computing and Media Centre at the University of Victoria.

Acknowledgements

I would also like to thank the generous input and mentorship of Raymond Siemens and Jentery Sayers at the University of Victoria during the development of this project. This work is supported by the Social Science and Humanities Research Council of Canada.

References

Foer, J. S. (2010). *Tree of Codes*. Belgium and The Netherlands: Visual Editions.

Galey, A. and Ruecker, S. (2010). How a Prototype Argues. *Literary and Linguistic Computing* 25.4 :405-424.

Gramazio, F. K. (2008). *Digital Materiality in Architecture*. Baden: Lars Mäller. .

Hayles, N. K. (2003). Translating Media: Why We should Rethink Textuality. *The Yale Journal of Criticism* 16.2: 263-290.

Heath, S. (2012). The Digital Materiality of Early Christian Visual Culture: Building on John 20:24-29. http://dh2011abstracts.stanford.edu/xtf/view?docId=tei/ab-315.xml . (accessed October 25, 2012).

Kirschenbaum, M. G. (2002). Editing the Interface: Textual Studies and First Generation Electronic Objects. *Text: An Interdisciplinary Annual of Textual Studies* 14: 15-51.

Kuniavsky, M. (2010). *Smart Things: Ubiquitous Computing User Experience Design.* Burlington: Morgan Kaufmann.

Leonardi, P. M. (2010). Digital Materiality? How Artifacts Without Matter, Matter. *First Monday* 15.6(June 2010): http://www.uic.edu/htbin/cgiwrap/bin/ojs/index.php/fm/ article/viewArticle/3036/2567 . Accessed October 25, 2012. Web.

Liu, A. (2004). Transcendental Data: Toward a Cultural History and Aesthetics of the New Encoded Discourse. *Critical Inquiry* 31.1:49-84.

Manoff, M. (2006). The materiality of digital collections: theoretical and historical perspectives. *Portal: Libraries and the Academy* 6.3 311-235. http://dspace.mit.edu/handle/1721.1/35689 . (accessed October 25).

McCarthy, W. (2005). *Humanities Computing*. London: Palgrave.

McGann, J. (2001). *Radiant Textuality: Literature After the World Wide Web.* New York: Palgrave.

Merholz, P. (2008). *Subject To Change: Creating Great Products & Services for an Uncertain World: Adaptive Path on Design.* Sebastobol: O'Reilly.

Wroblewski, L. (2002). *Site-Seeing: A Visual Approach to Web Usability.* New York: Hungry Minds.

Young, I. (2008). *Mental Models: Aligning Digital Strategy with Human Behaviour.* New York: Rosenfeld.

Becoming interdisciplinary

McCarty, Willard
willard.mccarty@kcl.ac.uk
King's College London, United Kingdom

Research in the digital humanities is by definition *interdisciplinary* and is frequently, though not always, *collaborative*. Both terms are often used in discussions of the field. The qualities they name are vigorously promoted and so close to becoming unquestioned virtues that they may seem merely descriptive of how research should or even must now be done. Yet we have a very poor grasp of what these qualities mean, how they shape practice, what they entail and how they change the disciplines and problems they involve.

Help is at hand for *collaborative* from the more than 30 years' work on the laboratory sciences by historians, sociologists and others — work that remains largely untapped by the digital humanities. For *interdisciplinary*, however, there is much less help, despite the fact that interdisciplinary research has been discussed on and off since the early 20th Century. In the proposed paper I will continue this discussion but in quite practical terms: not asking, as is so often done, what interdisciplinarity is, rather how becoming interdisciplinary can be intelligently attempted. I will briefly describe existing work on the topic and say why I think *interdisciplinary* is poorly understood and why it is important that we get it right. I will then exemplify a way forward by describing a doctoral-level course which I have taught for the last four years to students from the humanities and the social, health, and physical sciences. I will argue for disciplines as starting-points rather than "islands of knowledge" (Hacking 2012) to which we are necessarily marooned. I won't say much about how institutions work against our becoming as interdisciplinary as we can be, though they certainly do that.

Almost all discussion of the topic takes one of two forms. Most of it is framed by the ontological question: What is interdisciplinarity? How is it different from multidisciplinarity, transdisciplinarity and so on? (Klein 1990). The remainder assumes that interdisciplinary research is what happens when individuals from different disciplines collaborate. Both have their merits. Asking the ontological question leads to useful work in the history and sociology of knowledge (Frodeman et al. 2010); examining the work of teams illumines the transmission of knowledge across cultural boundaries (Gorman 2010) and their complex social and institutional dynamics (Strober 2010). But neither the ontological nor the sociological approach is of much help to individual scholars attracted by ideas from elsewhere or forced by the logic of their situation to take on a foreign discipline. Their problem is how to proceed immediately, by themselves, from initial curiosity to an understanding sufficient to make responsible use of another perspective. And neither approach addresses the problem illustrated by Myra Strober: how individuals within a collaborative team come to understand what the others are talking about and why, or fail to do so.

The ontological question tends also to mislead by using an abstract noun ("interdisciplinarity" &c) to name what is a way of acting rather than a aspirational state or class which scholarship can achieve. To use the abstraction implies the fixed form and properties of some*thing*, or a history of successive forms and properties. The individual scholar, with his or her immediate, practical concerns, needs help with acquiring knowledge *how*, not knowledge *that*.

Reframing interdisciplinary research as a way of acting clarifies the problem but does not make it easy. Stanley Fish has famously argued that "being interdisciplinary is so very hard to do" (1989), by which he meant impossible. Many attest that making the attempt is severely challenging. Gillian Beer has perhaps most eloquently of all spelled out the difficulties (2006) but nevertheless continues undaunted into "open fields" of knowledge (1996). Fish's argument turns on the impossibility of achieving a neutral and therefore perfectly interdisciplinary standpoint. Granted: there can be no perspective on disciplines unaffected by one's discipline of origin. But Fish goes badly wrong in asserting that any *attempt* is therefore not only impossible but also a moral error, Alan Liu points out (2008). I argue that trying is all for those of us who would extend our knowledge beyond what we have been conditioned to know in the ways we have been conditioned to know it.

Being interdisciplinary is difficult because from the get-go academic training situates the researcher within a specific field of discourse conducted, as Richard Rorty has said, "within an agreed-upon set of conventions about what counts as a relevant contribution, what counts as answering a question, what counts as having a good argument for that answer or a good criticism of it" (1979: 320). For this reason, in proportion to differences in its conventions, research in a discipline to which one is alien is difficult to see as good research, or even to see as research at all. The outsider presenting to insiders is apt to be greeted by incomprehension, misapprehension, indifference, hostility — or, what is worst of all, he or she may not be heard as saying much of anything.

Practitioners in the digital humanities cannot avoid putting themselves in the path of such danger if they are to be more than experts in a range of techniques. As a matter of course we practitioners are thrust into cross-disciplinary encounters in which operating intelligently within a foreign set of conventions is essential. The question is not whether to engage with strangers at a more than superficial level, rather how to do this well.

We face the problem in three areas.

First, as builders of things for others we cannot simply impose received methods onto problems in the form in which colleagues bring them: both problems and methods must be rethought in the light of each other. Each time the situation demands we converse in a suitable pidgin to negotiate the difficult "trading zone" between another discipline and computing as we know it (Gorman 2010). A conversational language which avoids the difficulties inherent to such an interchange by reducing it to the machine's known capabilities may lead to useful resources but stretches neither the client discipline nor our own. In fact it reduces the digital humanities to a mere service.

Second, like all those others who have brought new programmes of research into the academy, we need help from the older disciplines. Hence we need to look into them. Their help is not without its dangers, however, since ways of working and thinking, like physical instruments, implicitly carry baggage across that trading zone that can significantly affect how the recipient field subsequently does its work. We interdisciplinary researchers need to know insofar as possible not simply what the insider of another discipline knows but how he or she knows it and what its tendencies and entailments are.

Third, as cultural critics and educators we must be alert to the refiguration of disciplinary thought. This refiguration is powered by two covert forces: the incursion of methodology into the formerly non-methodological humanities (Gadamer 2000/1960), and the unavoidable temptation to stray far and wide in response to the riches presented to us and our students by JSTOR and the like. How possibly can scholars fulfil their responsibilities as educated commentators and teachers without the skills of interdisciplinary navigation?

In the proposed paper I argue that the most promising approach to the problem in all three of its forms is ethnographic, taking disciplines to be "epistemic cultures" (Knorr Cetina 1991) with the objective of discovering "the native's point of view" (Geertz 1983). Again this does not make the project of taking on a foreign discipline any less of a challenge, but it does provide a starting point. The usefulness of anthropological tools in computer science to improve the fit of system to users suggests their immense practical value (Crabtree, Rouncefield, and Tolmie 2012; cf. Nardi 2010). But as far as I can determine, little to no attempt has been made to train researchers to apply them to interdisciplinary encounters.

I will illustrate how I have done this by presenting the syllabus of a course developed from my own research practices and summarizing what I have learned from teaching it. After a brief theoretical introduction, this course, *Exploring Disciplines*, takes up a series of case-studies, each with readings in the discipline under consideration: philosophy, biology, history, literary studies, computer science, cultural studies and archaeology and epigraphy. (The Syllabus for 2013 may be found at http://tinyurl.com/bzl8755 .) Its ethnographic orientation is reflected in the organizing metaphor of an archipelago — again Hacking's "islands of knowledge". But sensitivity to the operative metaphor is cultivated by returning again and again to explicit consideration of its tendencies and to brief consideration of other possibilities (McCarty 2006).

References

Beer, G. (1996). *Open Fields: Science in Cultural Encounter*. Oxford: Oxford University Press.

Beer, G. (2006). *The Challenges of Interdisciplinarity*. Speech for the Annual Research Dinner. held 26 April 2006 at Durham University. www.dur.ac.uk/ias/news/annual_research_dinner/ (7 October 2012).

Crabtree, A., M. Rouncefield, and P. Tolmie (2012). *Doing Design Ethnography*. Berlin: Springer Verlag.

Fish, S. (1989). "Being Interdisciplinary Is So Very Hard To Do". *Profession 89*. 15-22. New York: Modern Language Association.

Frodeman, R., J. T. Klein, and C. Mitcham (eds). (2010). *The Oxford Handbook of Interdisciplinarity*. Oxford: Oxford University Press.

Gadamer, H.-G. (2000/1960) *Truth and Method*. 2nd rev. edn. trans. by Weinsheimer, J. and D. G. Marshall. New York: Continuum.

Geertz, C. (1983). 'From the Native's Point of View': On the Nature of Anthropological Understanding". *Local Knowledge: Further Essays in Interpretative Anthropology*. 3rd edn. New York: Basic Books.

Gorman, M. E. (2010). *Trading Zones and Interactional Expertise: Creating New Kinds of Collaboration*. Cambridge, MA: MIT Press.

Hacking, I. (2012). *The Anthropology (and Archaeology) of Numbers.* The Henry Meyers Lecture 2012. held at Royal Anthropological Institute. London. www.therai.org.uk/ (accessed 7 October 2012).

Klein, J. T. (1990). *Interdisciplinarity: History, Theory, & Practice*. Detroit: Wayne State University Press.

Knorr Cetina, K. (1991). "Epistemic Cultures: Forms of Reason in Science". *History of Political Economy* 23(1). 105-22.

Liu, A. (2008). "The Interdisciplinary War Machine". in *Local Transcendence: Essays on Postmodern Historicism and Database*. 169-85. Chicago: University of Chicago Press.

McCarty, W. (2006). "Tree, Turf, Centre, Archipelago — or Wild Acre? Metaphors and Stories for Humanities Computing". *Literary and Linguistic Computing* 21(1). 1-13.

Nardi, B. A. (2010). *My Life as a Night Elf Priest*. Ann Arbor: University of Michigan Press.

Rorty, R. (1979). *Philosophy and the Mirror of Nature*. Princeton: Princeton University Press.

Strober, M. (2010). *Interdisciplinary Conversations: Challenging Habits of Thought*. Stanford: Stanford University Press.

Exquisite Haiku: Experiments with Real-Time, Collaborative Poetry Composition

McClure, David William
david.mcclure@virginia.edu
Scholars' Lab, University of Virginia Library, United States of America

Introduction

There is a long tradition of artistic practices that have tried to reduce or remove the role of the author. Text generators produce poems according to algorithmic recipes; Oulipo procedures shuffle existing texts into new, unanticipated combinations; and collaborative processes produce collages, mash-ups, and other hybrid formats that combine a collection of individual voices into a single work. These author-reducing practices often take the form of rebellions against the "false prophets of genius and inspiration" (Motte, 203; qtd. Ramsay, 28), efforts to transcend the unity and rationality of single-author works and create new forms that "could not be created by one brain alone" (Breton, qtd. Waldberg, 95).

I ask the opposite question: Is it possible to create an "authorless" text that nonetheless looks and acts exactly like a single-author work? Drawing inspiration from the "Exquisite Corpse," a Surrealist parlor game, I present Exquisite Haiku, a real-time, interactive web application that makes it possible for groups of people to compose short poems. The software enables an extremely granular form of literary collaboration in which the intentions of the individual players are scrambled beyond recognition "into" or "below" the individual word selections. By pushing the register of collaboration down to an extended process of deliberation that takes place for each word, I argue that the resulting texts lack authors in the sense that they are completely *unattributable* — no individual person is responsible for any addressable part of the text. And yet they emerge from a completely deterministic, human-willed process and exhibit an "aura of sensefulness" (Ramsay, 30) and "closure in meaning" (Laxon, 30) that makes them indistinguishable from single-author texts.

The Exquisite Corpse and the "Fold"

First played by the friends of André Breton in Paris in the 1920's, the Exquisite Corpse[1] was one of the most influential of the Surrealist parlor games. A group sits at a table and passes around a sheet of paper — each player writes or sketches a new bit of material onto the page, and then folds the paper to conceal all but a small portion of the composite image. The end result is a phantasmagorical mash-up of the individual contributions, all highly individuated but bound together by a web of associations emerging from each player's partial knowledge of the whole. For Breton and the Surrealists, this disjointedness was a feature, not a bug — these "chain games" were efforts to invent new modes of artistic praxis that abandoned the constraints of single-author rationalism and moved towards a collaborative engagement with the *sur*-reality of Freudian free-association, a hop-scotching movement among images and motifs (Kern, 3-28). To achieve this fragmentation, the "links" of the chain needed to be blocky, rough, and obviously collaborative.

Interestingly, though, these theoretical goals start to unravel as the Corpse becomes more "competent" from a technical standpoint. Susan Laxon describes a variation of the game introduced in the mid-1930's by Valentine Hugo that tried to combine the energies of the individual players into a more unbroken, continuous form of collaboration. The physical "folds" were replaced with a system of discreet markings that protected the physical purity of the page, and the contour of the final image was traced over in an effort to conceal discontinuities and standardize the shading technique (Laxon, 41). For Laxon, this essentially an artistic error, at least when held up against the intellectual goals of the movement — the effort to "smooth" or "improve" the collaboration has the paradoxical effect of pushing the final product back in the direction of single-author unity. The Corpse stakes its claim to originality on the abruptness and defamiliarization brought on by the juxtaposition of the separate sec-tions. As they bleed together and start to reassemble towards cohesion, the Corpse cedes its *raison d'être*.

The altered rule set never caught on; before long, the fold was reinstated and the error reversed. But what if the "mistake" had been allowed to continue? Hugo's modified game points to an intriguing proposition: What if the collaboration actually became so granular — the individual contributions so completely pulverized into

the text — that it would be impossible to draw a line between any discrete unit of meaning in the work and an intending consciousness? Is it really true that a *complete* fragmentation of single authorship would, paradoxically, look almost identical to single authorship? Could this be achieved in a real artistic praxis? What kind of system or a game could actually reach this hypothetical endpoint?

Exquisite Haiku and Literary Consensus

In an effort to explore this question, I built Exquisite Haiku, a real-time web application written in JavaScript using Node.js that makes it possible for groups of people — as few as two, and theoretically as many as about 1,000 — to work together in an interactive, synchronous environment to compose three-line English haiku in a 5-7-5 syllable pattern. The words in the poem are selected one-by-one and in-order by a series of "word rounds," structurally identical cycles of play that last for a fixed amount of time (1-10 minutes). At the end of each round, the word that emerges with the most points from a game-like process in which the group collectively evaluates the proposals of individual players is locked into the next position in the poem, and the process restarts for the next word.

At the start of each round, players are presented with the words that have already been selected, a countdown timer that shows the amount of time left for the current word selection, and a "points" counter (all players get the same number of points, replenished at the start of each round). Players submit words by typing into a box at the top of the screen, and the words are pushed onto a continuously updated stack of unique possibilities. Starting immediately, and continuing as more proposals are submitted, players can influence the ordering of the stack by spending out little bursts of positive or negative points, of variable size, onto individual words by clicking and dragging up or down.

As soon as a vote is cast, the software immediately propagates the player's activity to all of the other players in the poem, all of whom see a millisecond-current interface designed to communicate a constant stream of information about the individual and aggregate sentiments of the players. The interface becomes a field of shifting, dilating, recoloring, reordering words that assembles the activity of the group into a single visual representation and sets the stage for an ongoing process of conversation, experimentation, and brinksmanship. Words are floated to the top and floated back down, tried out and then abandoned in favor of other ideas. A new submission or a large vote can trigger a sudden flurry of agreement or dissent. Competing factions form around words, the players tactically pacing their point expenditures to protect or bolster the position of a word in the stack.

Unlike other computer-assisted forms of writing, though, the software doesn't actually *do* anything at all — it doesn't pick words, inject any element of randomness or chance, or *do* anything else to materially impact the outcome of the process. It is essentially an elaborate scorekeeper or referee, a set of rules and processes according to which the sentiments of the players are collected, tallied, and redisplayed, with the final effect of distilling a single course of action from the divergent energies of the group. In this sense, the software becomes an exercise in applied political philosophy. It is a microcosmic legal or economic system, a legislative process for word selection, an artistic social contract that transplants the lowest-level mechanics of literary meaning-making — the raw, semiotic process of selecting signifiers and linking them together into syntagms — into a system of democratic governance.

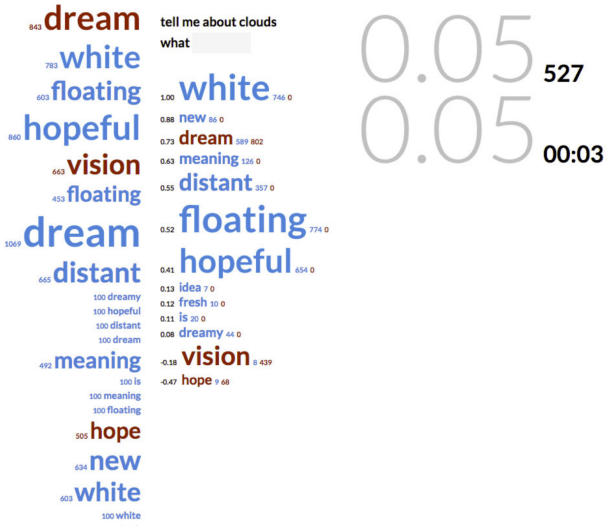


Figure 1:
A screenshot of Exquisite Haiku in the middle of the second line of a poem, near the end of a word round. In three seconds, "white," the top word in the main ranking stack below the poem, will get locked and the process will restart for the next word: "tell me about clouds / what white _____."

Authorless authorship

In the way that e^{-x} approaches 0 as x increases, I argue that the texts that emerge from this process approach complete *unattributability* as the number of players increases and the contribution of any individual becomes increasingly marginal. Unlike other "authorless" texts, though, the democratic nature of the game causes the

poems to gyroscopically evolve in the direction of highly specific, concrete meanings. At any given point in the composition process, there is a strong social tendency for the players to form a majority consensus around words that are semantically coherent in the context of the words that have already been selected, and that set the stage for future word choices that will result in poems that hold together overall[2] :

winter's color is
a clanging echo of some
distant summer song

electrical days
arrive without incident
pass without pleasure

in all shades of love
some element of hazard
glows forever pale

despite quiet thought
quote roaring ideas that
compel agreement

Although far from unequivocal (especially the fourth example, with the idiomatic strangeness of "quote" and "roaring"), there is an unmistakable structural firmness at play. At the level of pragmatic, ordinary-language interpretation, specific "theses" or "statements" are advanced.[3] Compare to some of the classic outputs of the original Corpse, which, quite deliberately on the part of their creators, almost provide negative definitions of this notion of sensefulness:

> The exquisite corpse will drink the young wine.
> The dormitory of friable little girls puts the odious box right.
> The Senegal oyster will eat the tricolor bread.
>
> (Rubin, 278)

I will argue that this tension between how the Exquisite Haiku are *produced* and how they *behave* as literary objects make them highly peculiar from a theoretical standpoint. They possess characteristics that seem to satisfy all of our intuitive requirements for "authored-ness" in the strong sense of the concept: They are formed exclusively by the volitional actions of human agents; there is no element of randomness or chance involved; they bear well-formed "sentence meanings" that say specific things. And yet they emerge from a radically collaborative process that in fact destroys exactly this concept of authorship.

References

Dowling, W. C. (1983). Intentionless Meaning. *Critical Inquiry*. 9.4 (June, 1983): 784-789.

Fish, S. E. (1980). *Is There a Text In This Class? : The Authority of Interpretive Communities*. Cambridge, MA: Harvard University Press.

Gadamer, H.-G. (1975). *Truth and Method*, trans. Garrett Burden and John Cumming. New York: Crossroad.

Hayles, N. K. (2008). *Electronic Literature: New Horizons for the Literary*. Notre Dame: University of Notre Dame Press.

Iser, W. (1972). The Reading Process: A Phenomenological Approach. *New Literary History*. 3.2 (Winter, 1972): 279-299.

Kern, A. M. (2009). From One Exquisite Corpse (in)to Another: Influences and Transformations from Early to Late Surrealist Games. *The Exquisite Corpse: Chance and Collaboration in Surrealism's Parlor Game*. Katana Kockhar-Lindgren, Davis Schneiderman and Tom Denlinger. Lincoln: University of Nebraska Press. 3-28.

Knapp, S., and W. B. Michaels (1982). Against Theory. *Critical Inquiry*. 8.4 (Summer, 1982): 723-742.

Laxon, S. (2009). 'This is not a drawing.' *The Exquisite Corpse: Chance and Collaboration in Surrealism's Parlor Game*. Katana Kockhar-Lindgren, Davis Schneiderman and Tom Denlinger. Lincoln: University of Nebraska Press. 29-48.

Mathews, H., and A. Brotchie. (1998). *Oulipo Compendium*. London: Atlas Press.

Motte, W. F. (1998). *Oulipo : a Primer of Potential Literature*. 1st Dalkey Archive ed. Normal, IL: Dalkey Archive Press.

Ramsay, S. (2011). *Reading Machines: Toward an Algorithmic Criticism*. Urbana: University of Illinois Press.

Rubin, W. S. (1968). *Dada and Surrealist Art*. New York: H. N. Abrams.

Searle, J. R. (1994). Literary Theory and Its Discontents. *New Literary History*. 25.3, 25th Anniversary Issue (Part 1) (Summer, 1994): 637-667.

Waldberg, P. (1997).*Surrealism*. New York: Thames and Hudson.

Notes

1. The name comes from the sentence that was produced the first time the game was played: "*Le ca-davre exquis boira le vin nouveau*" ("The exquisite corpse will drink the new wine").
2. No doubt these group dynamics reflect culturally informed values about what poetry should look like. It would be interesting to see if groups of players with different cultural backgrounds approach the game in different ways – both in terms of the types of poetry that

the group attempts to produce, and in terms of how players engage with the political (and even ethical) process of choosing words.

3. This is not always the case; sometimes the process "fails" in this regard. For example: "*who told god that we / are capable of seeing / shapes in dead leaves that*" – in this case, the players (I was one of them) miscounted the syllables in the last line, bringing the poem to a logical conclusion with "leaves," even though there was still one syllable remaining on the line. The group basically just gave up, the final "that" perhaps a gesture towards an idea that the phrase could continue beyond the boundaries of the poem. This is a technical error – remove the last word, and the poem is actually quite well-formed. In other cases, though, the poem just fails to close towards any immediately obvious, top-level meaning: "*we intend just what / you would fear nothing at the / hour of meaning*."

Approaching Algorithmic Media Analysis in the Humanities: An Experimental Testbed

McDonald, Jarom Lyle
jarom_mcdonald@byu.edu
Brigham Young University, United States of America

Hunter, Ian
beanland2@gmail.com
Brigham Young University, United States of America

With all of the growth and advancement in text analysis and visualization over the past several years[1] , more attention is being paid both to analyzing Humanities data as well as to exploring the relationship the interpretive mind of the Humanist has with the quantitative evidence gathered by machines. Recently, Steve Ramsay has called for more attentive development of an "algorithmic criticism" for literary analysis, where a reader uses computers to reveal patterns not easily evident or retrievable without the computational power a machine can offer (34). The key for Ramsay, however, is that the computer itself struggles with *interpretation*, and it behooves the literary critic to read the patterns as one might read a text to find meaning rather than rely on the data as a sort of factual proof of any Humanistic argument. And while, as Ramsay has pointed out, this sort of computational reading is not without conceptual or methodological obstacles (not to mention the technical ones that new tools, projects, and ideas constantly encounter), there is enormous value in "allow[ing] computer-assisted criticism to be situated within the broader context of literary study" (13). In fields ranging from literary studies (consider Franco Moretti's concept of distant reading) to computer science (such as Ben Shneiderman's work on information visualization) to predictive cultural analysis (exemplified in today's popular culture by Nate Silver and his arguments for more work in computer modeling of complex systems), the act of harnessing the power of algorithms to help read "text" demonstrates how vital it is for the digital humanities to avoid the pitfalls that come from crunching words without an ability to also understand them.

It's our contention that something is often missing from the conversation, however—algorithmic media analysis. If Digital Humanists want to more fully make sense of the relationship between meaning and computation, investigating quantitative reading of media is a far too under developed endeavour. The immediate initial reaction is often one of skepticism—after all, even when focusing solely on textual data, computer-assisted analysis is making strides but is still in its infancy in terms of being able to offer patterns to read that move beyond linguistic models or sentiment analysis. Yet outside of the Digital Humanities an ever-growing number of projects and computational tools from big names in "big data" are exploring the application of traditional text analysis approaches within a wider variety of media, ultimately allowing researchers to discover how multimedia might aid in the evolution of tools and paradigms for making more of quantitative analysis,[2] and we see an algorithmic approach from the Humanistic perspective a significant complement to what might be happening in the commercial enterprise R&D labs.

Of course, we're not in any way claiming that no work on machine-based analysis of multimedia data has ever taken place in the Humanities; quite the opposite, actually. There has always been a small but steady attention paid to how digital humanists might utilize new models for computationally approaching complex textual narratives, semantic relationships, image corpora, audio, and other such media [3] ; we might look to the work ongoing at the University of California, San Diego's Software Studies Initiative (such as their "FilmHistory.viz" project generated with the CineMetrics software tools); Lev Manovich's recent writings on visualization of visual media; the ShotLogger project, and Jason Mittell's media studies theories of what he calls complex television. But nothing has ever taken off in the way that text analysis has recently, so a question arises; how might the digital humanist more fully embrace multimedia modes of expression as a valid, if not even more informative, subject of algorithmic criticism?

What would, for example, a text analysis model for video look like, and might a quantitative approach to something like television or film yield a better understanding of some core Humanistic concepts such as story, character, or emotion? What sorts of quantitative analysis models might we apply to a medium such as television as a gateway for a deeper understanding of humanistic inquiry? And how useful might those models be, given some of the fundamental differences between text or language and newer communicative media of the past century?

In this short presentation, I will lay out this conceptual conversation and demonstrate some of this historical work that exists as examples of attempts to develop analytical tools for quantitative media reading. We are excited about how we might design our experiments to expand the realm that quantitative analysis of video might explore. I will then offer some of the initial trials we are undertaking to explore more robust approaches to an algorithmic criticism of multimedia within the digital humanities. Our preliminary trials are designed to explore what sorts of tools and visualizations might be truly useful for something as complex as a television episode or feature film. For example, whereas a digitized text might have only words to offer, a digital video object will have a text (the transcript), an audio stream (which may overlap with the text but which also includes points of analysis such as intensity, tone, pitch, speed, background music, ambient noise, etc.), and a linear sequence of images (which may be analyzed separately or in relation to other frames). Our various proof of concept models place all of these modes of information in conversation with each other, and we will begin to theorize the types of things that humanists might glean from further development of our initial experiments. For example, we look at what sorts of patterns might arise, from a visualization of a script placed in juxtaposition with a spectrogram of the audio track; how might the spatial and temporal difference between pixel color from one frame to the next as it might relate to a motion analysis.

As I report on our successes and failures with various quantitative models, seeing them truly as preliminary experiments that we hope will lead to tools, we will close with a case study that we hope to use as a demonstration of the real humanistic value of an algorithmic approach to multimedia. Given the underlying principle that algorithmic criticism is only useful inasmuch as a reader can discover patterns and apply semantics to those patterns, we might theorize that an effective way to evaluate a quantitative analytical tool would be to compare the patterns it reveals to patterns discovered manually by critics external to our experiments. We have taken a collection of pilot episodes of 30 American television series from the past 20 years, and and have given them to several different content experts at our university — media studies scholars, literary scholars specializing in narrative principles, contemporary American Studies scholars, etc. We have asked them to generate simple narrative maps of the episodes (i.e. periods of exposition, rising action, location of the climax, scenes depicting various emotions, etc.), as well as to generate some discussion as to what sorts of narrative features the various episodes have in common. Our purpose is to apply quantitative analysis approaches to the digital videos as well, with the ultimate goal being a theory as to what sorts of data-driven calculations, based on observation of technical details of digital video, might correlate best with the manual patterns that our scholars discover with traditional techniques. Can we use changes in audio intensity to help us better understand moments of conflict in a multimedia narrative? Can we find relationships between motion analysis and changing plot action?

Computers are still quite far from being able to perform tasks such as "understanding" nuance, tone, humor, irony, and other components of full semantic comprehension of literature. And we're not in any way claiming that our analytical approaches are a silver bullet in machine learning of humanities text. But we're arguing that they are the next step; that the digital humanities can advance the utility of algorithmic analysis by driving full-force in making multimedia the object of our study.

References

Cinemetrics Tools. http://cinemetrics.lv/

Heras, D. C. (2012). The Malleable Computer: Software and the Study of the Moving Image. *Frames Cinema Journal*. http://framescinemajournal.com/article/the-malleable-computer/

Manovitch, L. Visualization Methods for Media Studies. http://softwarestudies.com/cultural_analytics/Manovich.Visualization_Methods_Media_Studies.pdf

Mani, I. Computational Narratology. *The Living Handbook of Narratology*. Hamburg University Press. http://hup.sub.uni-hamburg.de/lhn/index.php/Computational_Narratology

Mittell, J. (2006). Narrative Complexity in Contemporary American Television. *The Velvet Light Trap* **58**. 29-40.

Moretti, F. (2000). Conjectures on World Literature. *New Left Review* **1**. http://newleftreview.org/II/1/franco-moretti-conjectures-on-world-literature

Porter, M. J., D. L. Larson, A. Harthcock, and K. B. Nellis. Examining television narrative structure. *Journal of Popular Film and Television.* **30**. 23-30.

Ramsay, S. (2011) *Reading Machines: Toward an Algorithmic Criticism*. Champaign, IL: University of Illinois Press.

Shotlogger 2.0. http://shotlogger.org/aboutusV2.php

Silver, N. (2012) *The Signal and the Noise: Why So Many Predictions Fail, but Some Don't*. New York: Penguin Press.

Notes

1. I most certainly recognize that text analysis is far from a new field, and has been an activity at the intersection of technology and the Humanities since the days of Father Busa. What I'm referring to, however, is the central primacy that text analysis and similar methodologies have recently acquired in the Digital Humanities discourse.
2. For example, Microsoft ("Multimedia Search and Mining"), IBM ("Semantic Learning and Analysis of Multimedia"), and Google ("Google X Laboratory") are all heavily invested in automated processing of multimedia for the purpose of pattern recognition.
3. In addition to more comprehensive work, a cursory scan reveals that last year's Digital Humanities conference offered presentations that dealt with computational analysis (or related strategies such as data mining, visualization, etc.) of acoustic ecology, literary genres, spatial readings, user generated content, aural and prosodic patterns in text, and computational narratology.

A Community Fab Lab: Introductions to Making

McGrath, Robert E.
robert.mcgrath216@gmail.com
Champaign Urbana Community Fab Lab, United States of America

Introduction

Digital technology is revolutionizing both the study of Humanities and the practices of "humanistic" disciplines (Schreibman, Siemens et al. 2004), enabling new forms of creative and scholarly communities (e.g. HASTAC (HASTAC 2013), The Pool (New Media Collective 2013), (Gauntlett 2011)) and new forms of expression (e.g. (Craig 2013; Latulipe, Wilson et al. 2010; Schiphorst 2009; Sherman and Craig 2003; Smith 2011)). Furthermore, as more digital technologies become ubiquitous, humanistic practices are reintegrating the living culture of contemporary life in many ways, including syntheses such as "ArtScience" (Edwards 2008), "Alternative Reality" (McGonigal 2011), and "Learn to Teach, Teach to Learn" (PBworks 2013). This paper considers one example of such reintegration in progress, a local community-based Fab Lab, in which we are developing an approach to fostering collaboration and creativity.

Digital Fabrication Technology

In recent years, digital fabrication technology has become widely available at moderate prices, opening the era of personal fabrication (Anderson 2012; Gershenfeld 2005; Mota 2011; Rischau 2011). Personal fabrication technology is seen as revolutionary; potentially reordering the manufacturing economy and changing the relationship between designer, producer and consumer (Anderson 2012; Gauntlett 2011; Johnson 2010; Rischau 2011). From the perspective of a humanist, this technology can be viewed as transforming creativity and — literally — putting tools in the hands of workers; by empowering every individual to fabricate whatever he or she desires, provided he or she learns how to do so.

Digital Fabrication technology crosses the boundaries between physical and virtual worlds, and this crossing is decisive. Once something (in this case, the design for making an artifact) is digitized, it becomes possible not only to manipulate it, but also to mix, remix, sell, and share it. Just as digital music and video have become universally available, nearly for free, uploaded as well as consumed; the ability to design and make things will be universally shareable.

The availability of low-cost digital fabrication technology has led to the emergence of a variety of community-based social spaces, including Fab Labs affiliated with MIT (FabWiki 2013), independent Maker Spaces (HackerspaceWiki 2013), and similar groups (e.g. (BioCurious 2013)). These spaces deploy low-cost digital design and fabrication techniques in small, community-based workshops featuring an empowering and creative "do it yourself" ethos within a supportive, diverse, and multidisciplinary setting. Digital technology can enhance and magnify long existing creative drives because knowledge can be shared in a very direct way: knowledge (i.e., designs) can be uploaded and then downloaded and "executed" (e.g. from Shapeways (Shapeways 2013), Thingiverse (Makerbot Industries 2013) or Instructables (Instructables 2013)), thereby directly connecting local labs, businesses, and individuals world wide (Anderson 2012; Gauntlett 2011; Gershenfeld 2005).

A Local Community Fab Lab

The Champaign Urbana Community Fab Lab (CUCFL), located on the University of Illinois Urbana campus, is a volunteer-operated, open community of people who like to

design and make things (CUCFL 2013; Ginger, McGrath et al. 2012; Watson 2011). The CUCFL makes available to the local community resources, including skilled volunteers, computers, computer-controlled machines, and electronics assembly tools. These high tech tools, and an open, informal environment, make it possible for people of all ages and skill levels to learn to imagine, design, and build.

The CUCFL is affiliated with the International Fab Lab Network that originated at MIT (FabFolk 2013). The more than 140 member labs around the world are operated independently, sharing and cooperating through standards and a common vision (FabWiki 2013). In addition to the MIT affiliated Fab Labs, hundreds of independent "Maker Spaces" and "Hacker Spaces", which provide similar creative environments (HackerspaceWiki 2013). These spaces form a large, informal, global community of makers, who share a plethora of information and enthusiasm via the Internet.

Who Does What in the CUCFL

The CUCFL has built a community of people with a range of expertise and a desire to share; volunteers who love to learn how to make things, and who love to share their knowledge with others. On any given day you may find kids and parents, University students and staff, artisans, entrepreneurs, school teachers, and retirees; working side by side on projects of their own design; learning and teaching.

An important goal of the CUCFL is to provide an environment in which everyone, including young women and kids from underserved communities, can imagine, make, and share; learn and, in turn, become a teacher. The CUCFL serves kids from youth groups, schools, and home school groups, as well as hobbyists, handicrafters, inventors developing prototypes of new products, and artists exploring and fabricating new creations.

In keeping with a humanistic spirit, many projects develop collaboratively in the lab, through free-wheeling discussion, experimentation, and iteration, which can produce eclectic explorations of academic, technological, and artistic concepts (e.g. (McGrath, Rischau et al. 2012)). The CUCFL hosts workshops and design sessions, such as a one-day "Fab Off", attended by participants from other Fab Labs in the region (CUCFL 2011). Several founders of the CUCFL have graduated University and founded a local design and prototyping company (The Product Manufactory 2013).

How We Introduce People to Making

We work hard to make the CUCFL something you "do", not something you "watch," therefore, one of the most important activities is inviting new users to begin making. Many first time visitors find a Fab Lab to be an alien environment; with an exciting but daunting array of high tech tools. And, at first, many people have little notion of what can be done, and might imagine that they are incapable of creating their own designs.

Introductory sessions consist of an introduction to the lab, followed by a hands-on project tutored by experienced volunteers. Based on the principle of "show, don't tell", the novice has the opportunity to, and is strongly encouraged to, design and make a simple object, such as a personalized key ring or sticker — to actually go through the process of making something *on the first day*. These objects are insignificant in themselves, but are highly meaningful because of who made them and how: the experience of making is simultaneously empowering and an encouraging introduction into our (humanistic) community of makers.

The "Fab Lab to Lab Fa"b Initiative

In 2011, the global Fab Lab initiative proposed a challenge to see if labs could be created according to the power of ten; if and how labs might be created at a series of scales, roughly designated by levels of cash outlay: $100, $1000, $10,000, *etc*. In effect, this extends the Global Fab Lab Network with another layer of "capillaries". Inspired by this challenge, the CUCFL, in collaboration with the Center for Digital Inclusion of the University of Illinois at Urbana Champaign, is creating a network of mini labs in a variety of local venues targeted at students from 8 to 18 years of age, including school classrooms, public libraries and dedicated clubhouse/community centers (Ginger, McGrath et al., 2012).

The CUCFL and the collaborative sites have a mutual set of responsibilities. The mini labs commit to use their resources in the promotion and development of skills and capabilities that align with the Fab Lab mission: personal growth, economic development and cross-cultural understanding. In return, each local mini lab receives a starter set of equipment and supplies that will enable the people at that site to create a variety of objects of their own design. The main CUCFL provides training, assistance and materials, including tutorials and starter project kits, and, as users develop skills, they will be able to use the larger capabilities of the main lab.

Conclusion

Fab Labs provide technology within a local, informal community of knowledge, to encourage and enable innovation and creativity by the people who work in or in cooperation with the labs, and, we believe, enabling

personal growth, economic development and cross-cultural understanding. The Champaign Urbana Community Fab Lab is one such lab, where we are exploring how to create and sustain a community of learners and makers, with deep local roots, and wide global connections.

These technologically enabled networks of local workshops represent an important trend in the contemporary practices of humanism, (re-)integrating art, design, engineering, and entrepreneurship, and crossing the boundaries between physical and virtual creativity, through social interaction, and knowledge sharing. Beyond the technical and economic implications, these technologies can have profound and exciting psychological and cultural effects. It is thrilling when a kid's face lights up and he or she holds up an object and says, possibly for the first time in their life, "*I* made this". And that is only the beginning — the kids inevitably teach other kids (and adults).

Notes

Thanks to Betty J. Barrett, Jeff Ginger, and Peter Organisciak for reading and discussing the drafts of this paper. Thanks to all the volunteers and community partners who make the CUCFL possible.

Funding

This work was partly supported by the Office of the Provost, Office of Public Engagement, and the Center for Digital Inclusion of the University of Illinois, Urbana-Champaign.

References

Anderson, C. (2012). *Makers*. New York: Random House.

BioCurious. (2013). *BioCurious — your Bay Area hackerspace for biotech*. http://biocurious.org/ (accessed February 12, 2013).

Craig, A. B. (2013). *Understanding Augmented Reality: Concepts and Applications (forthcoming)*. San Francisco: Morgan Kaufman.

CUCFL. (2011). *Fab Off Slideshow!* http://cucfablab.org/blog/fab-slideshow (accessed February 12, 2013).

CUCFL. (2013). *Champaign Urbana Community Fab Lab* http://cucfablab.org/ (accessed February 12, 2013).

Edwards, D. (2008). *Artscience: Creativity in the Post-Google Generation*. Cambridge: Harvard University Press.

FabFolk. (2013). *The International Fab Lab Association*. http://fablabinternational.org/ (accessed February 12, 2013).

FabWiki. (2013). *Labs*. http://wiki.fablab.is/wiki/Portal:Labs (accessed February 12, 2013).

Gauntlett, D. (2011). *Making is Connecting: The social meaning of creativity from DIY and knitting to YouTube and Web 2.0*. Cambridge: Polity.

Gershenfeld, N. (2005). *Fab: The Coming Revolution On Your Desktop — From Personal Computing to Personal Fabrication*. New York: Basic Books.

Ginger, J., et al. (2012). *Mini Labs: Building Capacity for Innovation Through A Local Community Fab Lab Network.* 'World Fab Conference (Fab8)'. held 22-28 August 2012 in Wellington, NZ.

HackerspaceWiki. (2013). *hackerspaces*. http://hackerspaces.org/wiki/Hackerspaces (accessed February 12, 2013).

HASTAC. (2013). *Humanities, Arts, Science Advanced Collaboratory*. http://hastac.org (accessed February 12, 2013).

Instructables. (2013). *Instructables — Make, How To, and DIY*. http://www.instructables.com/ (accessed February 12, 2013).

Johnson, J. (2010). *Atoms Are Not Bits; Wired Is Not A Business Magazine*. http://gizmodo.com/5457461/atoms-are-not-bits-wired-is-not-a-business-magazine (accessed February 12, 2013).

Latulipe, C., et al. (2010). *Exploring the design space in technology-augmented dance.* 'CHI '10 Extended Abstracts on Human Factors in Computing Systems'. held in 2010 in Atlanta, Georgia.

Makerbot Industries. (2013). *Thingiverse*. http://www.thingiverse.com/ (accessed February 12, 2013).

McGonigal, J. (2011). *Reality is broken: why games make us better and how they can change the world*. New York: Penguin Press.

McGrath, R. E., J. Rischau, and A. B. Craig. (2012). Transforming Creativity: Personalized Manufacturing Meets Embodied Computing. *Knowledge Management and E-Learning* 4.2: 157-173.

Mota, C. (2011). *The Rise of Personal Fabrication.* Creativity & Cognition. held November 3-6 in Atlanta.

New Media Collective. (2013). *the pool*. http://pool.newmedia.umaine.edu/ (accessed February 12, 2013).

PBworks. (2013). *Learn 2 Teach, Teach 2 Learn*. http://learn2teach.pbworks.com/w/page/15779288/Learn%202%20Teach%2C%20Teach%202%20Learn (accessed February 12, 2013).

Rischau, J. (2011). *Custom Digital Fabrication in Industrial Design*. MFA thesis. Urbana-Champaign: University of Illinois.

Schiphorst, T. (2009). Body Matters: The Palpability of Invisible Computing. *Leonardo* 42(3): 225-230.

Schreibman, S., R. G. Siemens, and J. Unsworth. (eds.) (2004). *A companion to digital humanities*. Malden, MA: Blackwell Publishing.

Shapeways, I. (2013). *shapeways*. http://www.shapeways.com/about/ (accessed February 12, 2013).

Sherman, W. R. and A. B. Craig. (2003). *Understanding Virtual Reality: Interface Application and Design*. San Francisco: Morgan Kaufmann.

Smith, B. D. (2011). *Telematic Composition*. Ph.D. thesis. Urbana-Champaign: University of Illinois.

The Product Manufactory. (2013). *The Product Manufactory*. http://www.theproductmanufactory.com/ (accessed February 12, 2013).

Watson, G. (2011). The Champaign-Urbana Community Fab Lab. *ACM interactions* 18.5: 86-87.

The Digital Scholarship Training Programme at British Library

McGregor, Nora
nora.mcgregor@bl.uk
British Library, United Kingdom

Farquhar, Adam
adam.farquhar@bl.uk
British Library, United Kingdom

Introduction

Research libraries and cultural heritage institutions must ensure that staff skills and core competencies keep pace with a rapidly changing research environment if they are to continue to effectively support and engage with scholars [1] . The American Council of Learned Societies Commission on Cyberinfrastructure for the Humanities & Social Sciences (ACLS 2006) observes that: In recent practice, “digital scholarship” has meant several related things:

a) Building a digital collection of information for further study and analysis
b) Creating appropriate tools for collection-building
c) Creating appropriate tools for the analysis and study of collections
d) Using digital collections and analytical tools to generate new intellectual products
e) Creating authoring tools for these new intellectual products, either in traditional forms or in digital form

The British Library is realigning its services and structure and in 2010 the Digital Scholarship department was established with a remit to ensure the Library develops its strategy and service propositions to enable and support these digital scholarship activities. The Digital Curator team was created within it to build the staff capacity to deliver on this strategy and ensure that the entire collections workforce is fully versed in the opportunities that digital content and new technologies can offer. This paper discusses the design and implementation of our in-house Digital Scholarship Training Programme at British Library.

Training Objectives

An important first step in shaping our training initiative was the establishment of four clear objectives for what we hoped to achieve for colleagues, and by extension Library users.

1. Staff across all collection areas are familiar and conversant with the foundational concepts, methods and tools of digital scholarship.

Outside the purview of digital scholarship were courses in basic computer literacy: this training was already available to staff through Human Resources.

2. Staff are empowered to innovate.

Seb Chan(Cooper-Hewitt Museum) and Rob Stein(Dallas Museum of Art) stress that innovation can come from anywhere within an organisation and institutions should be careful to avoid erecting unintentional barriers by allotting space and resources too selectively[2] . Our programme would underscore that staff across the Library have the power to innovate, and would provide the support to do so.

3. Our internal capacity for training and skill-sharing in digital scholarship are a shared responsibility across the Library.

Within the organisation, there are areas of world-class expertise in digital content, research, and scholarship. The programme must leverage and amplify this by working with these staff to develop and deliver course modules.

4. Collaborative digital initiatives flourish across subject areas within the Library as well as externally.

The training programme would open up direct communication between colleagues across subject areas as well as digital scholars, ensuring opportunities for collaboration and improvements on service arise.

Design & Development

In April 2012 the Digital Curator team embarked on an intensive three-month survey of the current digital scholarship landscape.

Having conducted a literature survey, the team sought out scholars working at the intersection of computing and scholarship and joined them for informal chats about their research[3] . Perhaps inevitably, we were frequently drawn to activities within the field of Digital Humanities, its very existence the embodiment of trends towards more digital scholarly practice in academia[4] . We consulted the proceedings of major conferences across Europe such as Digital Humanities 2012 in Hamburg and the Digital Humanities Congress 2012 at University of Sheffield and surveyed the skills which academics were acquiring by attending pertinent training courses [5] and reviewing open syllabi[6] and course materials.[7]

By August 2012 the team had outlined the specific concepts, methods and tools which were of direct relevance to library staff. We initially considered taking an advisory approach whereby we would point staff to externally available training opportunities in the areas we had outlined, but found this would not suffice in meeting our objectives; existing courses were by-and-large written for academics or the private sector and the cost of sending a preponderance of staff on them was prohibitive.

This informed our decision to design and deliver our own curriculum in-house and we subsequently drafted individual briefs and learning outcomes for what would become our core offering of 15 one-day courses. Each of the three Digital Curators took responsibility for managing five of the courses and worked with our internal advisory board and instructors from within the Library and institutions on the leading edge of digital scholarship such as King's College London, Open University, University College London and University of Oxford to finalise the courses.

Instructors were asked to consider the following when preparing course materials:

- Content should be aimed at "intelligent novices", that is, staff who have heard about the concepts but haven't had the time, space or opportunity to explore them in any depth.
- Focus on the wider concepts, methods and processes which tools enable rather than teaching to the tools.
- Include a hands-on practical element wherever possible, preferably using British Library digital content.
- Deliver from the library practitioner perspective and highlight the Library's current work, or potential for such work. It is crucial that staff clearly connect the relevancy of this new knowledge to their role at the Library.
- Deliver a one-day workshop onsite rather than online. Courses would not be held online as that could unnecessarily alienate an audience with varied technical skills. A full-day commitment would also provide necessary time and mental space away from business-as-usual activities while underscoring this development is a priority.

The Curriculum

We launched the two-year programme officially in November 2012 and the first of four planned semesters ran through the end of March 2013 with these fifteen courses:

101 What is Digital Scholarship?
102 Digital Collections at British Library
103 Digitisation at British Library
104 Communicating our collections online: Copyright considerations and Opportunities
105 Crowdsourcing in Libraries, Museums and Cultural Heritage Institutions
106 Text Encoding Initiative
107 Data Visualisation for Analysis in Scholarly Research
108 Geo-referencing and Digital Mapping
109 Information Integration: Mash-ups, API's and Linked Data
110 Social Media: Introduction to the Library's Social Media Policy, Twitter and Blogging
111 Working collaboratively: Using the British Library Wiki, Yammer and Google Drive
112 Presentation skills: From Powerpoint to Prezi
113 Foundations in working with Digital Objects: From Images to A/V
114 Behind the Screen: Basics of the Web HTML, CSS, XML
115 Metadata for Electronic Resources: Dublin Core, METS, MODS, XML

The content of each was carefully designed to specifically suit the Library's point-of-view. For example, the course 'Information Integration: Mash-ups, API's and Linked Data' provided a broad overview of the terms[8] , but also stressed to staff the immediate potential for these technologies in

connecting our digital content with external sources. A practical hands-on exercise[9] walked them through accessing our own British National Bibliography API as a digital scholar might and highlighted its potential as a rich resource for answering complex research questions. The exercise also showcased how our data formats helped or hindered such queries, providing a useful perspective for staff who may create API's in future.

Early Progress & Lessons Learned

A total of 86 staff members took part in the first semester, attending an average of 2.7 courses. Nine courses were led by external instructors, while the remaining six were taught by British Library staff. Feedback was captured via free text evaluation forms collected at the end of each course which have given us some good indications about what has worked and what needs addressing in the next semester.

- When asked what they enjoyed most on any given course, staff consistently noted they valued time to freely brainstorm ideas with colleagues.
- The inclusion of practical hands-on activities alongside lectures was also highly valued. Highly structured exercises with clear step-by-step directions were favoured over unstructured time devoted to free exploration of tools.
- We set course capacities too ambitiously which made them a challenge to deliver. We reduced capacities from 30 to 15 for courses with hands-on exercises.
- Courses which are tool-based, for example 110 Social Media, will be broken into discreet modules so participants need only attend the sections they require.
- Curators were given first priority on all courses initially as we had considered them our target audience. We changed this policy shortly after launch and opened courses to all interested staff as there was little justification for maintaining a waiting list in light of such positive demand.
- Initial take-up has benefited from a core of early adopters and new hires. As this demographic complete the courses, we will need to be more strategic and creative in marketing the programme to those less inclined.
- How we capture information now will be crucial for gauging impact over the long-term. Several attendees have alerted us to project ideas they intend to take forward and we must ensure a mechanism is in place to log and monitor such activities[10] .

This short paper reports on this model and our experience with the hope that it may be useful for similar institutions.

References

The American Council on Learned Societies (ACLS) (2006). *Our Cultural Commonwealth: The report of the American Council of Learned Societies Commission on Cyberinfrastructure for the Humanities and Social Sciences.* New York: ACLS. http://www.acls.org/cyberinfrastructure/ (accessed 07 March 2013).

Digital.Humanities@Oxford Summer School (2012) *Full Programme.* http://digital.humanities.ox.ac.uk/dhoxss/2012/fullprogramme.html (accessed 07 March 2013)

Digital Humanities 2012 Hamburg (2012). *Programme.* http://www.dh2012.uni-hamburg.de/conference/programme/ (accessed 07 March 2013)

Digital Humanities Congress (2012). *Programme and Presentations.* http://www.sheffield.ac.uk/hri/dhc2012 (accessed 07 March 2013)

Edwards, C. (2013). DH Syllabi. *The CUNY Digital Humanities Resource Guide blog,* http://commons.gc.cuny.edu/wiki/index.php/DH_Syllabi (accessed 07 March 2013).

Research Information Network (2011). *Social media: A guide for researchers.* http://www.rin.ac.uk/system/files/attachments/social_media_guide_for_screen_0.pdf (accessed 07 March 2013)

Ridge, M. (2012). War, Plague and Fire' and 'Bootstrapping Innovation in Museums' at 'Museum Ideas 2012-Museums in the Era of Participatory Culture. *Open Objects blog,* November 03, 2012. http://openobjects.blogspot.co.uk/2012/11/war-plague-and-fire-and-bootstrapping.html (accessed 07 March 2013).

Robichaud, A., and C. Blevins (2011). *Tooling up for Digital Humanities.* http://toolingup.stanford.edu/ (accessed 07 March 2013)

Schreibman, S., R. Siemens, and J. Unsworth (eds.) (2004). *A Companion to Digital Humanities.* Oxford: Blackwell. http://www.digitalhumanities.org/companion/ (accessed 07 March 2013).

Schreibman, S., and R. Siemens (eds.) (2007). *A Companion to Digital Studies.* Oxford: Blackwell. http://www.digitalhumanities.org/companionDLS/ (accessed 07 March 2013).

Weller, M. (2009). *The Digital Scholar: How Technology is Changing Academic Practice.* London: Bloomsbury Academic .doi: http://dx.doi.org/10.5040/9781849666275 (accessed 07 March 2013).

Notes

1. See also Martin Weller's *The Digital Scholar: How Technology is Changing Academic Practice* http://dx.doi.org/10.5040/9781849666275
2. From their talk *Bootstrapping Innovation in Museums* at Museum Ideas 2012 http://openobjects.blogspot.co.uk/2012/11/war-plague-and-fire-and-bootstrapping.html
3. Several groups meet regularly in the immediate vicinity of British Library such as Decoding Digital Humanities London (DDHL) https://sites.google.com/site/ddhlondon/ and the Bloomsbury Digital Humanities Group.
4. Texts such as *A Companion to Digital Humanities* (2004) http://www.digitalhumanities.org/companion/ and *A Companion to Digital Studies* (2007) http://www.digitalhumanities.org/companionDLS/ were highly influential in the early formation of the department as well as the initial framing of the training offering.
5. Digital.Humanities@Oxford Summer School http://digital.humanities.ox.ac.uk/dhoxss/
6. A brief collection of DH-related syllabi has been helpfully collate here: http://commons.gc.cuny.edu/wiki/index.php/DH_Syllabi
7. Tooling Up for Digital Humanities at Stanford http://toolingup.stanford.edu/ , David Birnbaum's http://dh.obdurodon.org/ , and Research Information Network's Social media: A guide for researchers http://www.rin.ac.uk/our-work/communicating-and-disseminating-research/social-media-guide-researchers are notable examples.
8. As one attendee remarked, "Great to have something often referred to demystified!"
9. See the full exercise here: http://www.meanboyfriend.com/overdue_ideas/2013/02/introduction-to-apis/
10. Phil Hatfield, Curator of Canadian & Caribbean Studies, attended several of the courses and is now formally taking forward a project to visualise a portion of his collection: "We have a large collection of Canadian photographs and associated data at the Library and I'd been considering for some time now ways in which to work with them beyond simply hosting them in a typical image gallery. The course on Data Visualisation gave me the space to play around with some of my ideas for visualisations and pointed me in the direction of free tools out there such as Google Fusion Tables. I hadn't realised it was so easy to get started and was able to see the shape of the collection almost immediately."

Ambiances: A Framework to Write and Visualize Poetry

Meneses, Luis

ldmm@cs.tamu.edu
Center for the Study of Digital Libraries — Texas A&M University, United States of America

Furuta, Richard

furuta@cs.tamu.edu
Center for the Study of Digital Libraries — Texas A&M University, United States of America

Mandell, Laura

mandell@tamu.edu
Initiative for Digital Humanities, Media, and Culture — Texas A&M University, United States of America

Poetry visualization is used mostly to highlight patterns detected using computational tools. These tools aim to help scholars carry out the critical analysis of poetry and are based on data mining algorithms and visual tools (Bradford Paley). The purpose of these tools is to synthesize and bring forward certain key elements such as the structure of the narrative, the organization of the poem, the language elements, and the metaphors employed (Chaturvedi, 2011). This purpose is achieved by intentionally bringing forward the graphical elements of poetry by carefully layering annotations (Tufte, 1990) based on literary criticism, thus creating new methods to analyze poetry. Additionally, these visualization tools can place special emphasis in viewing the poems from different perspectives (Meneses et al., 2011) and visualizing the textual representation of the poetic texts in their formal structure (Audenaert et al., 2007).

However, poetry visualization can also be used to create new forms of expression. These new forms of expression apply deformation and transformation techniques to the original poem to create a new art form that in some cases retains the original characteristics of the poem. Examples of this transformation include *Visualizing Text* by Diana Lange (Lange), *Poetry on the Road* (Schaffors et al.), *Text Universe* (Rapati) and Ira Greenberg's *Syntactic Arthropod* (Greenberg).

We believe that the tools used nowadays to visualize poetry have three important characteristics. First, the visualizations are created after the poem has been published. Second, the scholars who create and use the visualization usually do not have any relationship to the author who created the original literary work. Third, the visualizations are a direct consequence of the transformations applied to the original text. When we put together all three characteristics we are left with a visualization tool that serves its purpose when highlighting certain passages in a poem, but does not have any effect or significant influence on the author as the poem was written beforehand.

In our research, we are attempting to challenge the three notions that in our opinion have defined how poetry is visualized. For this purpose, we are building an interactive framework to write and visualize poetry. We have named this framework “Ambiances”. The main goal of our framework is to create a tool that affords a symbiotic relationship between writing and visualizing a poem. In our framework, the process of writing a new poem influences its resulting visualization and the visualization also affects the process of writing. At first, we envisioned mechanisms that allowed the author to carry out the writing and visualizing processes interactively. However, our tool also brings forth new possibilities of interaction when a different author designs the visualization. In this case, this interaction between the author of the poem and the designer of the visualization can engage in collaborative authorship. Additionally, our prototype borrows some of the concepts behind Storyspace (Bernstein, 2002, Bernstein, 2009). One of our goals is to design a tool that allows the author to create visualizations (Bernstein, 2009) that give aesthetic clues about the contents of the narrative.

We developed the first prototype of the system using Processing , an open-source programming language designed specifically for visual artists and designers with the purpose of teaching the fundamentals of programming. Casey Reas and Ben Fry started Processing as a project in 2001. However, Processing has evolved through the years and it is used beyond its original pedagogical scope. Nowadays, Processing is also widely used by visual artists and designers to create new ways of displaying of data, animations and digital artworks.

Using Processing has some obvious advantages that are concerned how with how the source code is executed and its portability. Processing is an extension of the Java Imaging Library, so Processing files can be executed as Java programs. Additionally, the Processing language is also available as a Javascript port , making it possible to run this system entirely on a modern web browser. However, we believe that the main advantage derived from using Processing is the simplification of the development process by encapsulating the complex data structures into simpler objects and methods.

The Ambiances framework is composed of different areas or “environments”. The first environment is a text editor where the author writes and composes the poem. The second area consists of a minimalist-programming environment optimized for writing Processing code. The third is a visualization environment that uses the poem and the source code as input materials to create a resulting visualization interactively. We specifically chose to emphasize the visualization environment and make it an integral part of the system. Figure 1 shows a screenshot of a prototype of our framework using the poem “Bright Star” by John Keats.

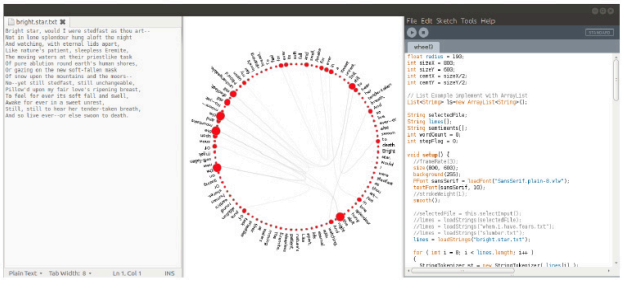


Figure 1.
Ambiances and a visualization of the poem “Bright Star” by John Keats.

The layout of the environments in the user interface allows the authors to engage in interesting ways. People have a natural tendency to create visual structures to organize resources when they share a workspace, which helps them to communicate interpretations and coordinate activities (Shipman et al., 2001a). This tendency towards the creation of visual structures facilitates the exchange and interpretation of visual information. In our prototype, the environments are linked together in a single widow frame.

The layout of the environments allows users to collaborate synchronously: authoring the poem and the visualization at the same time, or asynchronously: working at different times. Additionally, the interface encourages the authors to receive instant feedback through the visualization. Given that the authors cannot critique each other directly and can communicate through the visualization, Ambiances provides an unobtrusive way of writing poetry collaboratively that encourages unexpected interactions. For example: in the specific case where the visual elements are developing in syncopated opposition, we believe that the visualization and the interactions will provide hints that will allow the author of the poem to modify certain figures of speech accordingly. Our vision also allows authors to collaborate remotely from different locations. We will place special emphasis on evaluating the interactions in this case, as authors will communicate mostly through the visualization environment. However, at this point we do not discard allowing the authors to communicate using other means.

We plan on evaluating the affordances and the implications of our system by recording and analyzing the user interactions with the system prototype. This method of evaluation has been used in the past successfully in hypertext systems. In the specific case of the Visual Knowledge Builder (VKB), Shipman et al. recorded, logged and analyzed user interactions and found that their user interface created a strong relationship among documents, environments and their evolution through time (Shipman et al., 2001b). Similarly, our system allows authors to save sets of documents for a later time. This allows authors to expand

their creative and explorative processes through multiple sessions, thus making their use of the system into a more complete experience.

In the end, developing the prototype for Ambiances left us with questions that we still need to answer. One of them is if Ambiances will work better with authors in specific genres, writing styles, languages and cultural backgrounds. It is not unconventional to assume that it will, but we still need proof. For this purpose, we are gathering a diverse group of authors to participate in this study. The feedback that we will collect from the authors and their interactions will be included in the new iterations of the system.

The contribution of our research is to explore the advantages, issues and challenges that a framework for writing and visualizing poetry interactively and can provide. More specifically, we are interested in the feedback that authors obtain in real time from the visualizations and how it can affect their writing. Additionally, we will also address the multiple authorship issues and the implications that will surface from the use of our framework. In the end, our research aims to implement, analyze and study new interactive methods for creating and visualizing poetry.

References

Flare — Data Visualization for the Web http://flare.prefuse.org (accessed 1 November 2012.)

Juxta — Collation Software for Scholars http://www.juxtasoftware.org (accessed 1 November 2012.)

Many Eyes http://www-958.ibm.com/software/data/cognos/manyeyes/ (accessed 1 November 2012.)

Processing.js http://www.processingjs.org/ (accessed 22 August 2012.)

Processing.org http://www.processing.org/ (accessed 2 April 2012).

Audenaert, N., U. Karadkar, E. Mallen, R. Furuta, and S. Tonner *Viewing Texts: An Art-Centered Representation of Picasso's Writings. Digital Humanities 2007,* University of Illinois, Urbana-Champaign. 14- 16.

Bernstein, M. (2002). *Storyspace 1. Proceedings of the thirteenth ACM conference on Hypertext and hypermedia — Hypertext '02.* College Park, Maryland, USA.

Bernstein, M. (2009). *On hypertext narrative. Proceedings of the 20th ACM conference on Hypertext and hypermedia — Ht '09.* Torino, Italy.

Bradford Paley, W. *TextArc.org Home* http://textarc.org (accessed 1 November 2012.)

Chaturvedi, M. (2011). *Visualization of TEI Encoded Texts in Support of Close Reading.* M.A. thesis, Miami University.

Greenberg, I. syntactic_arthropod: Built with Processing http://iragreenberg.com/poetess/viz02/ (accessed 22 August 2012).

Lange, D. Visualizing text — OpenProcessing http://openprocessing.org/sketch/44133. (no date).

Meneses, L., C. Monroy, R. Furuta, and E. Mallen (2011). Computational Approaches to a Catalogue Raisonné of Pablo Picasso's Works. *Interdisciplinary Journal for Germanic Linguistics and Semiotic Analysis.*

Rapati, T. Text Universe http://openprocessing.org/sketch/10383 (accessed August 22,2012.)

Schaffors, A., B. Muller, and F. Pfeffer *esono.com — Poetry on the Road 2006* http://www.esono.com/boris/projects/poetry06/ (accessed 22 August 2012.)

Shipman, F., R. Airhart, H. Hsieh, P. Maloor, P. Moore, J. M. Moore, and D. Shah *Visual and spatial communication and task organization using the visual knowledge builder. Proceedings of the 2001 International ACM SIGGROUP Conference on Supporting Group Work, 2001* Boulder, Colorado, USA. 500325: ACM, 260-269.

Shipman, F., H. Hsieh, P. Maloor, and J. M. Moore *The visual knowledge builder: a second generation spatial hypertext. Proceedings of the twelfth ACM conference on Hypertext and Hypermedia — Hypertext '01, 2001b* Arhus, Denmark. 113-122.

Tufte, E. (1990). *Envisioning Information,* Cheshire, CT, Graphics Press.

Digging into Human Rights Violations: phrase mining and trigram visualization

Miller, Ben
miller@gsu.edu
Georgia State University, United States of America

Li, Fuxin
fli45@mail.gatech.edu
Georgia Institute of Technology

Shrestha, Ayush
ashrestha2@cs.gsu.edu
Georgia State University, United States of America

Umapathy, Karthikeyan
k.umapathy@unf.edu

University of North Florida

I. Introduction

Digging into Human Rights Violations (DHRV) is developing a computational reader for large text archives of human rights abuses so as to discover the stories of hidden victims and unidentified perpetrators only apparent when reading across large numbers of related documents. In part, this project began with an observation drawn from Benetech's Human Rights Data Analysis Group's (HRDAG) report on the Bosnian Book of Dead (Ball 2007). In their report on the tabulation of fatalities resulting from ethnic cleansing undertaken by the Milosevic regime, HRDAG was highly concerned with the de-duplication of entries. This over-reporting of individual victims within human rights corpora is endemic, and represents an opportunity for a system that can read across a corpus. For example, in the 511 interviews comprising our test corpus, one named individual appears more than 60 times. How many times might an unnamed individual reoccur? Automated readers exist that classify documents, produce summaries (Nenkova 2011), extract significant information (Strassel 2008) and highlight sentiment on a perentity, sentence, paragraph, or document basis (Pang 2008). This type of analysis works best with well-defined figures, such as occur in newspaper articles or government documents. That partially describes reports of human rights violations, as each report generally describes a victim's perspective of one limited event. Currently, these systems have difficulty parsing peripheral entities, indeterminate language, or references that go beyond the boundaries of one document and are only significant when traced across documents; many reports peripherally describe the fates of other victims. These implicit, buried links and duplicate reports amongst records enable horizontally reading across records collections, rather than vertical reading through one record.

The technical goal of DHRV is to create an NLP system that facilitates cross-document coreference of entities in collections of witness statements and interviews within the domain of rights. This project's approach to resolving the task (Kibble and Deemeter 2000) described as "whether or not two mentions of entities refer to the same person," begins by considering the subtype of anaphora (indicative language within a document) known as exophora (indicative language across more than one document), and relies on placing pronominal entities within a high-order Event Trigraph of location, time, and name. Because temporal information is so often referential and ambiguous, and therefore difficult to extract and correlate (Northwood 2000) our approach uses a phrase-based establishment of semantic context to support identifying the temporal context. This noun- and verb-phrase extraction, collocation detection, and semi-automated matching, feeds a 2D planar visualization similar to network graph models. Uncertainties in the document and information retrieval processes are visualized to allow researchers to confirm whether entity occurrences should be conflated. Because much human rights documentation contains sensitive information that cannot be made public, this project is prototyping with another historically significant corpus that shares many structural features to our primary data: the World Trade Center Task Force Interviews conducted with first responders to the attacks of September 11, 2001.

II. Event Summarization Based on Matching Phrases

Our main stratagem is to situate entities in the series of events that define their appearances. Phrases useful for this process accord to a "journalist template," of Who, What, When, Where, and Why, and are situated in the events reporting schema developed by Patrick Ball for human rights violations reporting (Chang 2012, Ball 1996). The goal is a system that can automatically extract these important entities as phrases, and based on these extracts, allow for the recognition of duplicate entities across documents. A perceptual diagram of the system is shown in Fig. 1.

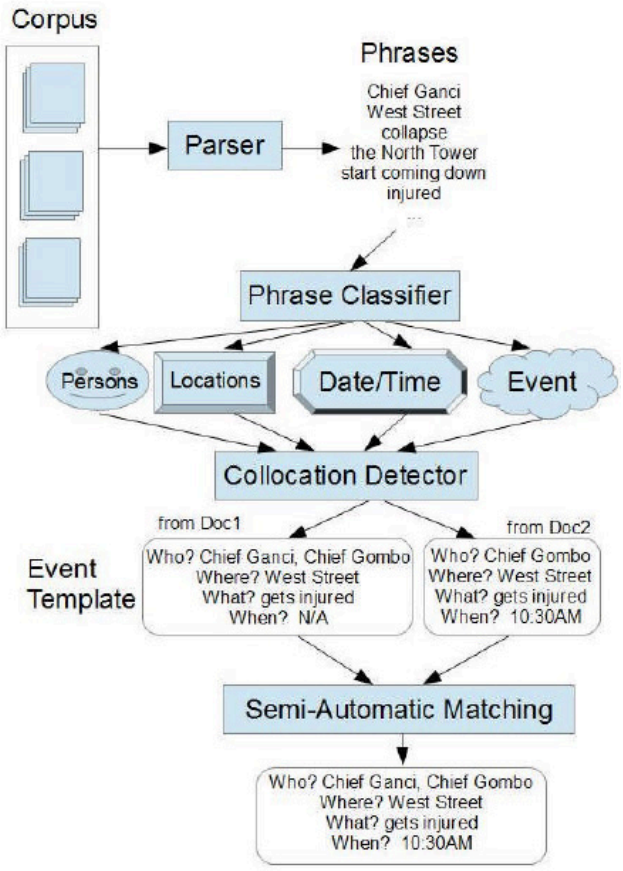


Figure 1:
Phrase mining perceptual diagram

In the system, the noun phrases and verb phrases are first extracted from a parser. Then a phrase classifier is used to determine which phrases fall into important entity categories such as Person names/Geographic Locations/ Date/Time or depiction of an event. After classifying these phrases into categories, a module called Collocation detector is ran to detect which of the detected entities are described in the same context within a passage in the corpus. This collocation of phrases is different from the collocation of words usually used in NLP, in that it captures instances of the collocations, instead of a global probability. A collocation is only true when multiple elements correlate. After a set of collocated phrases have been detected, they are placed into the event template and fed to a visualization engine. Human observers then decide which cross-document entities are identical. The engine computes automatic scores to make suggestions to the observers on which events and entities should be merged.

III. Phrase Extraction and Classification

The first step of phrase extraction is done by running a full parser on each document and then extracting all the retrieved noun phrases and verb phrases from the parse tree. We decided against using a shallow parser (chunker) because it has lower recall (may not capture all the desired phrases) than a full parser. The parser we are using is the Stanford parser (Klein 2003(1), Klein 2003(2)). From the extracted phrases, we formulate a classification task for labeling important phrases for event extraction. The important phrases in our research is different from the traditional named entity recognition (NER) problem in NLP, in that we are seeking to connect names to unnamed entities. We have 8 categories for important phrases: *Organization, Person, Title, Location, Date, Time, Event, Miscellaneous* and the background category of *Unimportant.* Of these categories, some are traditional NER or TimeML categories. *Event* and *Miscellaneous* labels are new, and determine some important phrases that might not be readily interpreted as named entities. Phrases such as “the pedestrian bridge,” “the ferry,” or “the second tower” which are not identifiable as a particular named entity, but might be crucial in depicting the event are classified as *Miscellaneous*.

To maximally utilize human knowledge in the phrase labeling phase, an unsupervised selection mechanism selects the phrases to be labeled. In this mechanism, phrases are ranked by a score that is similar to a frequency or N-gram model, but discounts the probability of a phrase if it is very common in a background corpus

$$Sc(phrase) = \log P(phrase) - \max(\log P_{bg}(phrase) - \log P(phrase), 0) \quad (1)$$

where log P(*phrase*) is computed by an N-gram language model trained on the current corpus, and log P_{bg}(*phrase*) is based on a Ngram language model trained on a background corpus that is supposed to contain documents of all kinds. Under this model, the probability of a phrase is only discounted if P_{bg}(*phrase*) > P(*phrase*). This application of Term Frequency–Inverse Document Frequency (Cohen 2002) helps us to find frequent phrases in the corpus which are not popular in the background corpus. The phrases with top scores are manually labeled. By this approach we can obtain the labels for the most frequent and unique phrases in the corpus, which are likely to be more important in isolating an event. Our N-gram training uses the modified Kneser-Ney smoothing (Chen 1999) from the MitLMpackage (Hsu 2008). The background language model is obtained from Microsoft Web Ngram Services. Given a set of human-labeled phrases, we then train two levels of classifiers on these phrases. At the first level, a binary Important versus *Unimportant* phrase classifier is trained. At the second level, a one-against-all multi-class classifier is trained for each of the phrase category described above, except *Miscellaneous*, which serves

as the background category for Important phrases. The features used for the classifiers are common NER features (Zhang 2003, Ratinov 2009), plus standard bagof- words features. For the *Date* and *Time* phrases, we make use of the SUTime library (Chang 2012) which matches date and time expressions using an extensive set of rules defined by regular expressions. The classification of these *Time* and *Date* phrases do not depend on our own human annotation.

IV. Collocation Detection and Event Templates

For the collocation we use a simple metric: a Gaussian kernel on the distance between mentions of different phrases. Formally, the collocation probability of one occurrence of a phrase, given a set of other phrases is defined as:

$$P(p_1|p_2, p_3 \ldots, p_k) = \exp(-\beta \sum_i (S(p_1) - S(p_i))^2) \quad (2)$$

where $S(p_i)$ is the sentence number where pi occurred. Given the defined conditionals, one can compute the joint probability $P(p_1, p_2, p_3, \ldots, p_k)$ and use a threshold to determine which phrase set goes to an event template.

V. Visualizing Unvertainty in Event Trigraphs

Depending upon the context, uncertainty refers to statistical uncertainty, ranged values or missing data (Pang 1997). Uncertainty can be introduced in the data during acquisition, processing or even visualization. In this paper, since the underlying data has been extracted from narratives, which in turn lacks precision, uncertainty is introduced right from the data acquisition phase. These uncertainties include the temporal, “By this time, it had to be 11:00 o'clock at night” and “at that time I noticed,” locative, “I guess that would be North End Avenue,” and entity, “At this point I had my five guys" [WTCTF 9110250]. One of the major goals of this project is to visualize the triadic relationship amongst character, time, and location while incorporating these uncertainties.

In (Skeels 2010), the authors present a survey of the existing works on uncertainty visualization. The survey shows that most research in this area is in the field of geospatial (MacEachren 2005) or scientific visualizations (Lodha 1996, Grigoryan 2004, Wittenbrink1996). It also states, "[t]he main techniques developed include adding glyphs, (Wittenbrink 1996, Lodha 1996) adding geometry, modifying geometry, (Grigoryan 2004) uncertainty in a model to decision makers (Walker 2003) modifying attributes, animation (Lodha 1996, Gershon 1992) and sonification.(LodhaSonic 1996).” To convey the time-location-character data and the associated uncertainty visually, in this paper, we introduce a trigram based visualization called an Event Trigraph.

An Event Trigraph is a 2D planar diagram consisting of events as its building blocks. An event in this case is a 3-tuple consisting of the basic elements derived from the phrase mining method detailed above to yield three elements: time, location and entity. Events are represented visually as triangles with the aforementioned basic elements as vertices connected with weighted edges. The weights represent the confidence value in the relation as obtained via our text mining methods. These confidence values show how likely is the connection both between two elements and amongst the trigram. Figure 2 shows a trigram with confidence values and elements.

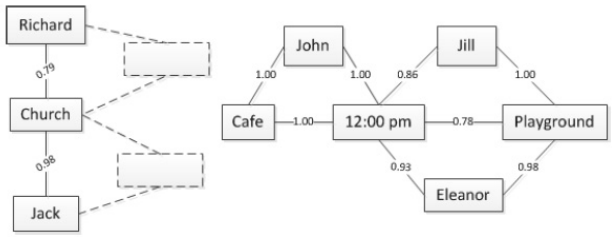


Figure 2.
Event trigraph with uncertainty values and voids

Confidence values peak at 1, or a certain relationship between two elements. To reduce visual clutter due to the excess number of edges and vertices, we employ details on demand (Shneiderman 1996) and filtering capabilities on the dataset. The users can also drag and drop the events over other events and manually enter the confidence values to present the associativity between two or more events. Since the user loads multiple documents at a time, this feature enables users to visually associate events in different documents — the main forte and a unique aspect of our visualization. Furthermore, since the resulting data structure is a weighted network graph, many graph theory algorithms can be readily applied, thus making it scalable.

VI. Conclusion

As reports are processed by this method, the diagram builds in complexity and supplements each individual event trigraph with ones that potentially correlate. In the figure above, John-Café-12:00 pm is correlated to Jill-Playground-12:00pm across the time element,

and Jill-Playground-12:00 pm correlates to Eleanor-Playground-12:00 pm across both time and location elements. This aggregated event trigraph builds to represent a corpus, and offers a method for correlating the collocation of entities across documents, and potentially identifying unnamed entities within said corpus.

References

Ball, P. (1996). *Who Did What to Whom?: Planning and implementing a large scale human rights data project*. Washington, D.C.: American Assoc. for the Advancement of Science, Science and Human Rights Program.

Ball, P., E. Tabeau and P. Verwimp (2007). *The Bosnian Book of Dead: Assessment of the Database*. Sussex: Households in Conflict Network, Institute of Development Studies.

Chang, A. X., and C. D. Manning (2012). Sutime: A library for recognizing and normalizing time expressions. In *International Conference on Language Resources and Evaluation* (LREC 2012).

Chen, S. F., and J. T. Goodman (1999). An empirical study of smoothing techniques for language modeling. *Computer Speech and Language*, 13:359–393.

Cohen, W. W., and J. Richman (2002). "Learning to Match and Cluster Large High-Dimensional Data Sets for Data Integration." *Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (KDD).

Gershon, N. D. (1992). "Visualization of fuzzy data using generalized animation." *IEEE Symposium on Visualization 1992*. Chicago: IEEE Computer Society Press, 268–273.

Grigoryan, G., and P. Rheingans (2004). "Point-based probabilistic surfaces to show surface uncertainty." *IEEE Transactions on Visualization and Computer Graphics* 10(5): 564–573.

Hsu, B. J. and J. Glass (2008). Iterative language model estimation: Efficient data structure and algorithms. In *Proceedings of Interspeech*.

Klein, D. and C. D. Manning (2003). Accurate unlexicalized parsing. In *Proceedings of the 41st Meeting of the Association for Computational Linguistics*, 423–430.

Klein, D. and C. D. Manning (2003). Fast exact inference with a factored model for natural language parsing. In *Advances in Neural Information Processing Systems* 15, 3–10.

Lodha, S. K., A. Pang, R. E. Sheehan, and C. M. Wittenbrink (1996). "UFLOW: Visualizing Uncertainty in Fluid Flow." *IEEE Symposium on Visualization* Chicago, IL: IEEE Computer Society Press, (1996): 249–255.

Lodha, S. K., C. M. Wilson, and R. E. Sheehan (1996). "LISTEN: Sounding uncertainty visualization." *Proceedings of the Visualization 1996*. Los Alamitos, CA: IEEE Computer Society, 189–195.

MacEachren, A. M., et al. (2005). "Visualizing geospatial information uncertainty: What we know and what we need to know." *Cartography and Geographic Information Science* 32.3: 139-160.

Nenkova, A., and K. McKeown (2011). "Automatic Summarization." In Functions and Trends in Information Retrieval. 5.2-3: 103-233.

Northwood, C. (2010). *TERNIP: Temporal Expression Recognition and Normalisation in Python*. Sheffield: Department of Computer Science.

Pang, A. T., C. M. Wittenbrink, and S. K. Lodha (1997). "Approaches to uncertainty visualization." *The Visual Computer* 13(8): 370-390.

Pang, B., and L. Lee (2008). "Opinion Mining and Sentiment Analysis." In Foundations and Trends in Information Retrieval. 2(1-2): 1-135.

Ratinov, L. and D. Roth (2009). Design challenges and misconceptions in named entity recognition. In *Proceedings of the Thirteenth Conference on Natural Language Learning* (CONLL 2009).

Shneiderman, B. (1996). "The eyes have it: A task by data type taxonomy for information visualizations." *Visual Languages. Proceedings*, IEEE Symposium on IEEE, 1996.

Skeels, M., et al. (2010). "Revealing uncertainty for information visualization." *Information Visualization* 9(1): 70-81.

Strassel, S, M. Przybocki, K. Peterson, Z. Song, and K. Maeda (2008). "Linguistic Resources and Evaluation Techniques for Evaluation of Cross-Document Automatic Content Extraction." In *Proceedings of the Sixth International Language Resources and Evaluation* (LREC'08).

Walker, W. E., et al. (2003). "Defining uncertainty: A conceptual basis for uncertainty management in model-based decision support." *Integrated Assessment* 4(1): 5–17

Wittenbrink, C.M., A. T. Pang, and S. K. Lodha. (1996). "Glyphs for visualizing uncertainty in vector fields." *IEEE Transactions on Visualization and Computer Graphics* 2(3): 266–279.

Zhang, T., and D. Johnson (2003). A robust risk minimization based named entity recognition system. In *Proceedings of the Seventh Conference on Natural Language Learning* (CONLL 2003).

Introducing Anvil Academic: Developing Publishing Models for the Digital Humanities

Moody, Fred
fmoody@anvilacademic.org
Anvvil Academic, United States of America

Spiro, Lisa
lisamspiro@gmail.com
Anvvil Academic, United States of America

Jackson, Korey
kjackson@anvilacademic.org
Anvvil Academic, United States of America

Background

Academic publishing provides vital services, including acquisition, peer review, editing, and distribution of scholarship. Yet there are too few venues for publishing digital humanities works, particularly on native platforms that represent the richness of this work and allow for interaction. This paper will introduce Anvil Academic [1], a new joint initiative of the Council for Library and Information Resources (CLIR) and the National Institute for Technology in Liberal Education (NITLE). Other sponsors include Stanford University, Washington University, Amherst College, Middlebury College, Bryn Mawr College, Southwestern University, NINES at the University of Virginia, and the Brown Foundation.

Anvil aims to address the current crisis in academic publishing by applying the time-honored editorial and peer review practices of publishing to the emerging world of digitally mediated humanities scholarship. In particular, Anvil will focus on four key (and sometimes overlapping) genres: data-driven projects that explore patterns in rich collections of humanities information; multi-modal titles using various modes of display and representation; networked authorship projects that facilitate conversation and connections; and flexible, interactive, media-rich educational resources. As an entrepreneurial, open-access publisher of peer-reviewed digital humanities scholarship, Anvil intends to help bring the digital humanities into the academic mainstream by forging new, sustainable models for publishing as the world of scholarly communication turns steadily digital.

Starting up a new open access, all-digital academic press for the digital humanities presents several challenges. What peer review standards and processes should be used to give credibility to digital scholarship, offer useful feedback for improving it, and promote community and conversation? What are the best technical and logistical approaches to publishing works that exist on a range of platforms and in a diversity of formats? How will these published works be preserved? What business models will enable an open access humanities press to provide core services and sustain itself? While we do not have definitive answers to these questions, we will discuss the contexts and strategies that inform Anvil's approach.

The Rationale for Anvil

Over the last decade, a rich body of humanities work featuring digital curation, data visualization, network analysis, text mining, multimodal argumentation, and other approaches has demonstrated the vitality and creativity of the digital humanities[2] . Despite its growth, this scholarship has yet to enter the academic mainstream in the way that analog research (words printed on paper or screen and distributed in the form of monographs and journal articles) has. Publishers still print and distribute academic monographs through sales channels, albeit to ever-fewer purchasers, most of which are libraries that find circulation of such books plummeting[3] . Yet traditional book/journal publication still enjoys unequalled status in the evaluation of applications for tenure and promotion while just as — or even more — deserving work in digital scholarship often goes unrewarded[4] .

Barriers facing digital scholarship include the conservatism of academic culture and the lack of respected entities for evaluating and disseminating this work. Some academics regard digital scholarship as second-rate work that has not undergone the scrutiny of credible peer review[5] . Others are reluctant to jeopardize their tenure and promotion prospects by disseminating their work through non-traditional means. In a prestige marketplace where a publisher's reputation is a primary determining factor in measuring the worth of an academic's published research, digital humanities need an organization to fulfill a role similar to that of the publisher in the analog world.

Anvil: The Way Forward

In its start-up phase, Anvil is confronting three core challenges:

Validating and Improving Scholarship through Peer Review and Editing

In essence, Anvil Academic aims to determine how a publisher of digital humanities scholarship can bring useful evaluation rubrics to the digital world, and to work with scholars to optimize and improve their work in order to make it publishable through rigorous editing and peer review. By defining and stepping into this role, Anvil seeks to bring order and coherence to a digital scholarship space that currently lacks the kind of guidance toward refined argument traditionally provided to the academic world by the scholarly publisher.

To demonstrate the credibility of digital scholarship, Anvil will involve both traditional and digital scholars in our peer review and acquisitions procedures. Drawing upon prior work on evaluating digital scholarship, we also will develop and implement rigorous standards for acceptance by Anvil. That acceptance will be contingent upon the rationale for the project's being digital and for what the project contributes to the state of the (digital) art in the humanities, as well as the project's contribution to the overall body of scholarship in its subject area. By carefully editing (in the senses of acquisitions editing, developmental editing, line editing, copy editing) the work, Anvil aims to demonstrate both the value of the publisher's role in the digital humanities and the value of digital work to humanities disciplines.

Platforms and Technical Skills

Anvil believes that a digital humanities publisher needs to be platform-independent: that is, the published works must be displayed in their native environments, so as to avoid the substantial costs and headaches of re-authoring complex works for the sake of making them suitable for a one-size-fits all authoring/viewing environment. Anvil thus focuses on providing editorial, marketing, distribution, cataloging, and preservation services rather than platforms in order to make electronic publishing of complex works economically feasible and to allow authors the fullest possible creative freedom. To facilitate the sustainability and reusability of digital scholarship, Anvil will encourage the use of open platforms, open licenses and open web standards. Anvil also uses the Internet Archive's Archive-It service to catalog and store titles in the Internet Archive and in LOCKSS, and is exploring a similar arrangement with HathiTrust. For some works that can be more constrained in their technical approaches, Anvil may provide hosting, but many Anvil publications will be hosted in their native environments, marked as Anvil publications and made available in Anvil's catalog. Anvil thus explores what it means to publish in a disaggregated environment where some services are provided through partnerships.

Business Models

Core to Anvil's mission is making publications available as open access, an approach consistent with the values and practices of the digital humanities community. Unfortunately, no single model for sustaining open monograph publishing has yet emerged. As Anvil explores how to provide the funding necessary to sustain digital publishing, it will experiment with several approaches, such as membership, fee-for service, sales of distilled app versions of Anvil titles, and grant funding.

It should be stressed that Anvil is in a proof-of-concept startup phase; as we grow, and as we acquire more funding, we will be adding to our arsenal of editorial, curatorial, and preservation resources and partnerships. In discussions with us, preservation of titles is the leading item of concern to authors of digital work.

Conclusion

The Anvil experiment hopes to bring digital humanities scholarship to a larger readership and to evangelize to the higher-education establishment for new-form scholarly work. Part of our mission is to work at persuading department chairs and administrators of the worth (and the measurability of that worth) of digitally mediated scholarly research and argument. Also critical to the Anvil experiment and mission is our intent to be as public as possible with our processes, so as to demonstrate to skeptics the worth and rigor of our editorial effort and the related worth of the resulting published work. Finally, Anvil aims to advance the state of the art in digitally mediated humanities scholarship, helping to set and demonstrate standards for assessing not only the scholarly content of such work but its methodology. We hope as well to demonstrate both the worth and the feasibility of publishing in the post-monograph space, so that more publishers will join in this effort, our combined presence growing into a cooperative/competitive effort that in the long run benefits—and advances—digital humanities as a whole.

Notes

1. http://anvilacademic.org/

2. Some leading examples of such work can be found in Christa Williford and Charles Henry, One Culture. Computationally Intensive Research in the Humanities and Social Sciences. CLIR, June 2012, http://www.clir.org/pubs/reports/pub151.
3. See "Report of the Executive Collection Development Executive Committee Task Force on Print Collection Usage Cornell University Library," http://staffweb.library.cornell.edu/system/files/CollectionUsageTF_ReportFinal11-22-10.pdf (Oct. 22, 2010, revised Nov. 22, 2010), p. 2.
4. Notwithstanding efforts like the Modern Language Association's guidelines for evaluating digital humanities scholarship, first approved by the MLA Executive Council in May 2000 and last reviewed by the Committee on Information Technology in January 2012. See "Guidelines for Evaluating Work in Digital Humanities and Digital Media," http://www.mla.org/guidelines_evaluation_digital.
5. See, for example, Gary A. Olson, "How Not to Reform Humanities Scholarship." The Chronicle of Higher Education, February 9, 2012, http://chronicle.com/article/How-Not-to-Reform-Humanities/130675/.
6. See Diane Harley, Sophia Krysz Acord, Sarah Earl-Novel, Shannon, Lawrence and C. Judson. King, *Assessing the Future Landscape of Scholarly Communication: An Exploration of Faculty Values and Needs in Seven Disciplines*. UC Berkeley: Center for Studies in Higher Education, 2010, http://cshe.berkeley.edu/publications/publications.php?id=351.
7. Among our exemplars in evaluating scholarship is NINES (http://www.nines.org/about/scholarship/peer-review/).
8. See Raym Crow, *Income models for Open Access: An overview of current practice*. SPARC, September 2009, http://www.arl.org/sparc/publisher/incomemodels/imguide.shtml

Semantic Augmentation and Externalization in the Humanities: a Demonstrative Use Case

Morbidoni, Christian
christian.morbidoni@gmail.com
Università Politecnica delle Marche, Italy

Grassi, Marco
margra75@gmail.com
Università Politecnica delle Marche, Italy

Nucci, Michele
m.nucci@univpm.it
Università Politecnica delle Marche, Italy

Fonda, Simone
fonda@netseven.it
NET7, Italy

1. Introduction

The Web is rapidly becoming an important source of documents and information for scholars in diverse disciplines and it is opening new scenarios for communication and collaboration. Scholars do not only need to easily find and access open content online, but also to be able to work with it, producing new knowledge and exchanging it with others. By means of content annotation and augmentation, two of the so-called "scholarly primitives" (Unsworth, 2000; Palmer et al., 2009), scholars have been doing this for decades. Their work consists, among other things, in enriching texts (or other kind of intellectual works) with new information, to advance the knowledge of a certain domain. Finally, externalization is an equally important scholarly primitive as it allows presenting results to the community; this usually corresponds to write a paper in the "traditional" academic world.

Effectively translating these primitives in the digital world is the great challenge and opportunity of Digital Humanities. With such purpose, it is commonly accepted that structuring data and metadata about digital objects using standard formats and schemas is a needed step to make content effectively accessible on a global scale. The Semantic Web technologies and Linked Data paradigm have become growingly accepted a way of representing, contextualizing data and making it interoperable(Gradmann, 2010).

The basic idea behind this paper is that the Semantic Web technologies can be used not only to properly represent "static" metadata but also to effectively structure annotations and make their semantics "explicit". On one side, this allows scholars to create new data and to contribute to the Linked Data Web. On the other, it enhances the externalization of such created knowledge easing the creation of rich and innovative data visualization applications. This enables a "virtuous circle" in which the knowledge generated by scholars can be merged with other data and become the learning and researching object for other scholars.

This paper presents an experimental scenario showing how annotation, augmentation and externalization of knowledge can be performed with (Semantic) Web tools

that are currently under development and evaluation in the SEMLIB[1] and the DM2E[2] EU projects.

After having shortly introduced Pundit (Morbidoni et al., 2011), we present a demonstrative prototype, based on Edgemaps (Dörk, Carpendale and Williamson 2011), where structured annotations are reused in a Web application to visualize a graph of influences among philosophers.

2. Pundit and Semantic Augmentation

Pundit is a semantic annotation tool that allows building structured data about digital objects, annotating entire Web pages down to single paragraphs, sentences or words. Web contents can be semantically augmented, establishing typed relations among different kinds of "entities", and contextualized, linking them to the Web of Data. For example, a scholar can state that a certain text excerpt "cites" another one, that it "describes" the subject of a picture, or that it "refers to" a place or a person.

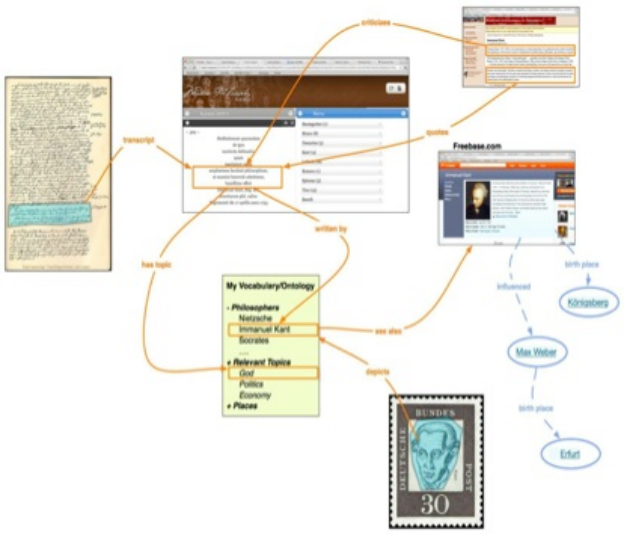


Figure 1:
Augmenting original content with semantically structured annotations

Augmentations always preserve authorship and are represented in RDF using Named Graphs. Augmentations, even when made by different users, can be merged to form a semantic graph, as the one shown in Fig. 1, where text fragments are connected with images, vocabularies entries and Linked Data sources as Freebase. Note that, as Freebase is an RDF data source itself, links from an annotation to a Freebase entity can be used by machines as possible gateways towards the Web of Data, where they can collect additional information (e.g. Kant date and place of birth)and further augment the original knowledge.

Nowadays, while some of the basic activities scholars do, as reading and writing papers, are already well supported in the digital world, some essential scholarly primitives, such as annotation, augmentation and externalization, do not yet have a clear support in terms of appropriate software tools.

Figure 2:
Pundit in action

- *Augmentation of online content:*Pundit provides different GUIs allowing the annotation of several media contents at different level of granularity and complexity, ranging from simple comments and semantic tags to triples(statements), where different kinds of "items" as text excerpt, images and fragments on images are connected by semantically defined relations.
- *Contextualization*, by linking contents parts to the Web of Data (e.g. DBPedia, Freebase)or to controlled custom vocabularies.
- *Simple aggregation.* It allows collecting (and reusing in annotations) items of interest.
- *Collaboration.* In Pundit annotations are collected in "notebooks" (each user can have multiple notebooks)that can be kept private or shared with others.

In Fig. 2, a screenshot shows the Pundit GUI. Annotations con be composed as triples of the form "subject-predicate-object" (as shown in Fig. 3). More details about Pundit, as well as a live demo, can be found in (Grassi et al., 2012; Nucci et al., 2012) and on the project Web site[3] .

Figure 3:
The Pundit triple composer.

3. Consuming collective knowledge: externalization

In Pundit, semantic augmentations are stored as RDF named graphs and made accessible to software by means of SPARQL endpoints and RESTful HTTP APIs. This allows external applications to easily fetch data and mix it with other Linked Data sources. For example, in the SEMLIB project the collective knowledge created by annotators is reused by a Semantic Recommender System (Policarpio, et al. 2012; Fossati, et al. 2012), which creates similarity links among contents in a digital library, offering an additional navigation layer to users.

3.1 Edgemaps Visualization: A Demonstrative Use Case

Data visualization is not a new topic in the Digital Humanities, as witnessed by projects such as Edgemaps(Dörk, Carpendale and Williamson 2011), where an interactive graph visualization represents philosophers in a timeline or in a similarity graph, showing their influences. The demo shows influence relations coming from Freebase, a well know general-purpose Linked Data repository[4] .

While the visualization is intuitive and has been highly appreciated in the Digital Humanities community, scholars are also concerned with questions like: “Why exactly the graph says that Marx influences Gramsci?”, ”What is the evidence of that in the primary sources?”, “Who said that?”.

The simple idea behind the proposed demonstrative application is to feed the influences graph with precise statements made by scholars, so that each edge in the graph can be linked to an annotation that “justifies” its existence, linking back to primary sources.

In our example, we annotated open contents on Wikisource.org, which publishes in a wiki-form a big amount of primary literature. To do this, Pundit has been customized to accommodate semantic relations extracted from the CiTO ontology[5] . The relation set includes predicates like “cites” and “quotes”, as well as other more specific ones like “discusses”, “cites as sources”, “agrees with”, etc. The Pundit bookmarklet allows loading the Pundit annotation environment on any Web page, and annotating the text while browsing it in its original location (Wikisoruce.org).

Figure 1 shows how Pundit can be used to produce an annotation that connects two texts from different philosophers. The dropdown menu allows specifying a precise relation among the ones proposed (e.g. cites, agrees with, etc.). In terms of RDF triples the annotations would look like the following:

- :text_1 cito:agreesWith :text_2.
- :text_1 dc:creator freebase:John_Locke.
- :text_2 dc:creator freebase:George_Berkeley.

Figure 4:
Creating a semantic annotation, composing a triple(or statement)with Pundit.

Finally, we created a simple Web application based on Edgemaps to load influences relations from users augmentations: each time an annotation exists that connects texts from two different authors a corresponding edge is created in the graph connecting the two authors. When browsing the graph, a scholar can see all the annotations that “establish” a specific influence link with another philosopher selecting an author, as shown in Figure 2. The demonstrative application[6] implements a simple HTTP API with the following parameters:

- *nbs,* comma separated IDs of notebooks (users personal collections of annotations) to get data from
- *source:* [freebase|pundit], tells the application if load data from freebase only, from pundit only or from both.

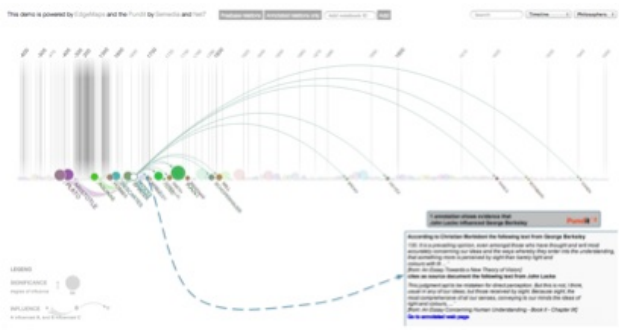


Figure 5:
Showing evidences of philosopher influence with a Timeline Visualization

For example, at the following URL: http://metasound.dibet.univpm.it/thepund.it/edgemaps_demo/demo.html?nbs={c4a2729c}&source={freebase,pundit}#phils;map;;/en/john_locke ; the Edgemap shows relations among John Locke and other philosophers retrieved both from freebase data and from a Pundit notebook (whose id is c4a2729c). By mouse over on Berkeley, a box appears showing an annotation that "justifies" the relation. A scholar can easily see who is the author of the annotation, read the annotated text and go to the annotated Web page to see the information in its original context.

4. Conclusions

In this paper, we introduced the concept of semantic augmentation on which Pundit annotation system is founded and presented an example of how semantic data created by scholars or professional con drive live externalizations of a research activity. More experiments are currently undergoing to implement the proposed externalization paradigm also in other scenarios. Another example application has been developed for data journalism, whose detailed description is out of scope here, which allows to put in relations public declarations of politicians (annotated from online newspapers) with the trend of financial indicators[7] .

Acknowledgements

The research leading to these results has received funding from the European Union's Seventh Framework Programme managed by REA-Research Executive Agency [SEMLIB — 262301 — FP7/2007-2013 — FP7/2007-2011 — SME-2010-1]. The research is also supported by the DM2E project, funded by the European Commission's "ICT Policy Support Programme" (ICT PSP), agreement No. 297274.

References

Fossati, M., C. Giuliano, and G.Tummarello. (2012). Semantic Network-driven News Recommender Systems: a Celebrity Gossip Use Case. In *International Workshop on Semantic Technologies meet Recommender Systems & Big Data* held at ISWC 2012.

Gradmann, S. (2010). Knowledge = Information in Context: on the Importance of Semantic Contextualisation in Europeana. http://www.scribd.com/doc/32110457/Europeana-White-Paper-1

Grassi, M., C. Morbidoni, M. Nucci, S. Fonda, and G. Ledda (2012). Pundit: Semantically Structured Annotations for Web Contents and Digital Libraries in Mitschick, A., F. Loizides, P. Predoiu, A. Nürnberger, and S. Ross (eds.) *Semantic Digital Archives 2012. Proceedings of the Second International Workshop on Semantic Digital Archives (SDA 2012)*, held 27 September in Paphos, Cyprus. CEUR-WS.org/Vol-912, urn:nbn:de:0074-912-6.

Morbidoni, C. and M. Grassi, and M. Nucci (2011). Introducing SemLib Project: Semantic Web Tools for Digital Libraries, *Proceedings of the International Workshop on Semantic Digital Archives — sustainable long-term curation perspectives of Cultural Heritage held as part of the 15th International Conference on Theory and Practice of Digital Libraries(TPDL)*. Berlin, 29th September 2011.

Nucci, M., M. Grassi, C. Morbidoni, and F. Piazza (2012). Enriching Digital Libraries Contents with SemLib Semantic Annotation System. *Proceedings of the Digital Humanities 2012 Conference* held 16-20 July 2012 in Hamburg, Germany.

Palmer, C., L. Teffeau, and C. Pirmann (2009). Scholarly Information Practices in the Online Environment: Themes from the Literature and Implications for Library Service Development. Report. Development. http://www.oclc.org/resources/research/publications/library/2009/2009-02.pdf

Policarpio, S., S. Brunk, and S. Tummarello (2012). Implementation of a SPARQL Integrated Recommendation Engine for Linked Data with Hybrid Capabilities. *Artificial Intelligence meets the Web of Data Workshop at ECAI'12*. held 27-28 August in Montpellier, France.

Unsworth, J. (2000). Scholarly Primitives: what methods do humanities researchers have in common, and how might our tools reflect this? Symposium on Humanities Computing formal methods experimental practice. http://www3.isrl.illinois.edu/~unsworth/Kings.5-00/primitives.html

Notes

1. SEMLIB Project: http://www.semlibproject.eu/
2. DM2E Project: http://dm2e.eu/
3. Pundit: http://thepund.it
4. Marian Dörk EdgeMaps: http://mariandoerk.de/edgemaps/
5. CiTO Ontology: http://purl.org/spar/cito/
6. http://metasound.dibet.univpm.it/thepund.it/edgemaps_demo/demo.html
7. Political rhetoric vs. STOXX 600: http://thepund.it/journalism-how-to.php

The FAST-CAT: Empowering Cultural Heritage Annotations

Munnelly, Gary
munnelg@scss.tcd.ie
Trinity College Dublin, Ireland

Hampson, Cormac
cormac.hampson@scss.tcd.ie
Trinity College Dublin, Ireland

Ferro, Nicola
ferro@dei.unipd.it
University of Padua, Italy

Conlan, Owen
owen.conlan@scss.tcd.ie
Trinity College Dublin, Ireland

Introduction

The role of annotations in digital humanities is well known and documented (Agosti et al., 2004, 2007) (Bélanger, 2010). Subsequently, many different tools which allow for the annotation of digital humanities content have been developed. Unfortunately, tools designed specifically for an individual portal are typically only compatible with that system. More general solutions, which can be easily distributed across various sites, have been produced, but these systems often have limited functionality (only annotating a single content type, no sharing features etc.) (Okfn) (TILE, 2011).

FAST-CAT (Flexible Annotation Semantic Tool — Content Annotation Tool) is a generic annotation system that directly addresses this challenge by implementing a convenient and powerful means of annotating digital content. It provides a reliable, portable manner of annotating both textual and image content in documents. The annotations are stored remotely by the FAST service which means that they may be shared across different sites maintaining the same data without the need for modification.

FAST-CAT has been developed as a module for the Drupal 7 content management system making it extremely easy to add to an existing Drupal site.

This paper introduces FAST, the backend service providing powerful annotation functionalities, and CAT, the frontend Web annotation tool, and discusses how its features are tackling important challenges within the Digital Humanities field.

Features of FAST

The FAST annotation service adopts and implements the formal model for annotations proposed by (Agosti and Ferro, 2008). Since then, FAST has been completely reengineered with added functionality such as provenance, logging and extended searching. According to this model, an annotation is a compound multimedia object which is constituted by different signs of annotation. Each sign materializes part of the annotation itself; for example, a textual sign would contain the textual content of the annotation, an image sign would contain images, etc. Each sign is characterized by one or more meanings of annotation, which specify the semantics of the sign, e.g. a sign whose meaning corresponds to the “title” field in the Dublin Core (DC) metadata schema or a sign carrying a question from the author whose meaning may be “question” or similar.

The flexibility inherent in the annotation model allows us to create a connective structure, which is superimposed to the underlying documents managed by digital libraries. This can span and cross the boundaries of different digital libraries and the Web, allowing the users to create new paths and connections among resources at a global scale.

FAST defines three different scopes which determine the visibility of an annotation — private, public and group. With FAST-CAT, the scope of an annotation is set to private

by default, meaning that only the person who created the annotation can see it. These annotations may serve as reminders or notes within the document for the user, akin to writing in the margin of a page.

The user can choose to make an annotation public, allowing other users to read their comments on a document. This has numerous applications for users of all levels of expertise. For instance, experienced users may choose to create public notes and annotations which can expand on the text of the document, helping less experienced users to comprehend the content. Less experienced users may indicate parts of the document which they would like to be explained further.

Group annotations allow users to provide viewing and editing permissions on an annotation to specific users. This means that a team of people working towards a similar goal could communicate directly through the medium of annotation. In this way, it can be seen that FAST-CAT can play a crucial role in collaboration.

Features of CAT

CAT is a web annotation tool whose development began in July 2012. It has been developed with the goal of being able to annotate multiple forms of document content and assist in collaboration in the field of digital humanities. At present, CAT allows for the annotation of both text and images. The current granularity for annotation of text is at the level of the letter. For image annotations, the granularity is at the level of the pixel. This allows for extremely precise document annotation, which is very relevant to the Digital Humanities domain due to the variety of different assets that prevail.

There are two types of annotation which may be created using CAT; a targeted annotation and a note. A targeted annotation is a comment which is associated with a specific part of the document. This may be a paragraph, a picture or an individual word, but the defining feature is that the text is directly associated with a specific entity. Conversely, a note is simply attached to the document. It is not associated with a specific item therein. Typically, this serves as a general comment or remark about the document as a whole. Further to allowing a user to comment on document text, the annotations created using CAT allow a user to link their annotations to other, external sources. Hence CAT can be used to construct a narrative through a number of documents. This is hugely beneficial for teachers using digital cultural collections and for students from primary to university level. For example, using these links a teacher can construct a predetermined path for their students to follow through a series of sites relevant to their chosen topic. Importantly, each link has comment text associated with it, allowing an educator to explain why this specific link is important or what the student should seek to gain from following this particular path.

While CAT is beneficial for researchers and educators, it is also being used as an important source of user data for the content provider. Websites such as Amazon and YouTube are able to provide increasingly accurate recommendations for their individual users. These recommendations are facilitated by a user model which is driven by a combination of ratings, recently viewed items and numerous other factors. For a digital humanities site, annotations provide an insight into which entities are of interest to a user. If a user is frequently annotating a document, it is likely that this document is of interest to them. Furthermore, if the text being annotated is analysed, it may be possible to discern specific items of interest within the document. A digital humanities site which can determine what a user is attempting to study, then anticipate and recommend sources that may be of use to them in the future is profoundly useful. If well implemented, curators of digital humanities portals will see a dramatic improvement in the effectiveness with which researchers interact with their domain.

FAST-CAT and CULTURA

At present, FAST-CAT is being developed as part of the CULTURA project (Hampson et al., 2012a, 2012b). A key aspect of CULTURA is the production of an online environment that empowers users, of various levels of expertise, to investigate, comprehend and contribute to digital cultural collections. FAST-CAT is a key component of this environment and is currently being trialled with the help of three different user groups.

A team of MPhil students and professional researchers will use the tool as part of their teaching, collaboration and research into the 1641 depositions. These users will be testing FAST-CAT in a free form manner. How they choose to annotate and what content they label is entirely determined by their own needs. The 1641 depositions are text only content, so these students will serve only to evaluate the text annotation aspect of the tool.

Providing an alternative insight is a group of secondary school students from Lancaster whose teacher will use the annotations to guide them through a lesson. These students will also be working with the 1641 depositions.

Masters students in Padua will test the image annotation functionality of FAST-CAT as part of their research into the IPSA collections of illuminated manuscripts. Similarly to the MPhil students, their approach to annotating documents will be determined by their own research methodology.

The various features offered by FAST-CAT and its user interface will be evaluated in detail and comparisons will be drawn between the manner in which different user groups availed of annotations depending on their level of expertise

and document content. Furthermore, FAST-CAT will also help to drive CULTURA's comprehensive user model by providing the site with updates on the user's behaviour regarding document annotation.

Future Work

Much of the further enhancement of FAST-CAT will be based on the feedback given by the user groups mentioned in the previous section. However there are already plans to expand and improve the system for future versions.

While FAST-CAT is supported by modern browsers, to improve portability, implementations for older web browsers will be developed.

FAST-CAT is a Drupal 7 module which means that, at present, it is only available for the annotation of websites which are built using the Drupal content management system. However, it only utilises a small amount of Drupal functionality to relay messages from the client computer to FAST. This dependency is easily removed as the majority of functionality is either client side or independent of Drupal. Designing and implementing a more server agnostic php script will allow FAST-CAT to be deployed on any website. This is one of the main items for future development of the system and will help to ensure that FAST-CAT can be utilised by as wide a range of content based websites as possible.

References

Agosti, M., G. Bonfiglio-Dosio, and N. Ferro (2007). A Historical and Contemporary Study on Annotations to Derive Key Features for Systems Design. *International Journal on Digital Libraries (IJoDL)*. 8.1. 1-19.

Agosti, M., & N. Ferro (2008). A formal model of annotations of digital content. *ACM Transactions on Information Systems (TOIS)*. 26.1. 3:1-3:57.

Agosti and Ferro, I. Frommholz, & U. Thiel (2004). Annotations in Digital Libraries and Collaboratories — Facets, Models and Usage. In *Proc. 8th European Conference on Research and Advanced Technology for Digital Libraries (ECDL 2004)*, 244-255. LNCS 3232, Springer: Heidelberg.

Bélanger, M.-E. (2010). *Ideals*. https://www.ideals.illinois.edu/bitstream/handle/2142/15035/belanger.pdf?sequence=2 (accessed October 25, 2012)

Hampson, C., M. Agosti, N. Orio, E. Bailey, S. Lawless, and O. Conlan (2012). *The CULTURA Project: Supporting Next Generation Interaction with Digital Cultural Heritage Collections.* Limassol, Cyprus.

Hampson, C., S. Lawless, E. Bailey, S. Yogev, N. Zwerdling, and D. Carmel (2012). *CULTURA: A Metadata-Rich Environment to Support the Enhanced Interrogation of Cultural Collections*. Cádiz, Spain.

Okfn. *Okfn Annotator*. http://okfnlabs.org/annotator/ (accessed June 2012).

TILE. (2011). *TILE: text-image linking environment*. http://mith.umd.edu/tile/ (accessed July 2012).

Possibilities of narrative visualization: Case studies of lesson-learned-oriented archiving for natural disaster

Nameda, Akinobu
a.nameda7@googlemail.com
Ritsumeikan University, Japan

Wakabayashi, Kosuke
kwakaba@fc.ritsumei.ac.jp
Ritsumeikan University, Japan

Nakatsuma, Takuya
t.nakatsuma@gmail.com
Ritsumeikan University, Japan

Hatano, Tomomi
y0300528@pl.ritsumei.ac.jp
Ritsumeikan University, Japan

Saito, Shinya
ken@arc.ritsumei.ac.jp
Ritsumeikan University, Japan

Inaba, Mitsuyuki
inabam@sps.ritsumei.ac.jp
Ritsumeikan University, Japan

Sato, Tatsuya
satot@lt.ritsumei.ac.jp
Ritsumeikan University, Japan

Abstract

After the Great East Japan earthquake struck, the need for learning lessons from the vast amount of information has been increased in order for preparing the next Great earthquake. Therefore, the present study explores a method of displaying textual information of experiences and detecting narratives to learn lessons. We conducted two visualizations of database on the Great earthquakes and considered the outcome of each. We found that the outcome of one of the visualization was gaining factual information on phenomena, whereas the other that was designed to have future perspectives allowed to find the demands or views for the future as well as the facts in the past. If taking the perspective for extracting narratives and learning lessons, the outcome having future views is more contributing to gaining meaningful story. The future-oriented perspective would be recommended to add into the process of designing archive to learn lessons on natural disaster.

1. Background and purpose

As a means of succeeding collective memory, the social significance of digital archives has been increased. Since the Great East Japan earthquake in particular, organizations both in public and private sectors have created archives on the Great earthquake. As can be seen in numbers of archiving projects, large amount of descriptions, pictures, and movie images has been collected (examples seen in references section).

In creating archives, on the one hand, it is very important to preserve the whole and various kinds of information as much as possible. On the other hand, if considering the natural disaster will occur again, there is an emergent need to learn lessons from the experiences on the natural disaster and convey the lessons to the next generation in order to prepare for the next strike of the Great earthquake. As such, the issue here is to develop an effective way to learn lessons from large amount of information in archive.

One of the ways to learn lessons is focusing on narratives on the disaster and visualizing them. Narrative here is defined as the one having story line which is associated with time and space. According to the Bamberg (2012), when people do storytelling, the narrators "position characters in space and time and, in a broad sense, give order to and make sense of what happened" (85). In applying this into the context of visualizing archive, it would be beneficial to extract the time order and things making sense in order to learn lessons that are supposed to be something making sensible and meaningful. The present study, therefore, explore a method of displaying textual information of experiences and detecting narratives to learn lessons. More concretely, we conduct two visualizations of database on the Great earthquakes and discuss the outcome of each from the perspective of leveraging information in archives.

2. Case studies

2-1. Study 1: Visualizing information database of the Great Hanshin-Awaji earthquake

We visualized the information database of the Great Hanshin-Awaji earthquake of 1995, which was created and published by the government of Japan (Cabinet Office, Government of Japan, 2006). The database included textual information in relevant articles to what happened in local places within the Hanshin-Awaji area in Western Japan from 1995 to 2005, 10 years after the earthquake. These articles were classified into four periods. Details on the data were presented in Table 1.

Period 1 (from the actual earthquake to 72 hours after)	Period 2 (from Day4 to 3 weeks after the earth quake)	Period 3 (4 weeks to 6 months after the earth quake)	Period 4 (6 months after the earth quake)
1-01.Actual earthquake	2-01.Operating evacuation centers	3-01.Building emergency dwelling	4-01.Reconstruction of life
1-02. Initial reactions	2-02.Supporting life in disaster area	3-02.Rebulding residences and life	4-02.Revival of the industry and cities
1-03.People's behavior	2-03.Determining the situation	3-03.Planning reconstruction	
1-04.Rescue and emergency medical treatment	2-04.Volunteers	3-04.Demolishing damaged buildings	
1-05.Fire handling	2-05.Re-establishing urban infrastructure	3-05.Industry recovery	
1-06.Emergency transportation			
1-07.Food and supplies			
1-08.Health and hygiene			
1-09.Lifeline			
1-10.Reaction among companies			
1-11.Preventing a second disaster			

Table 1.
Content structure of articles in the information database of the Great Hansin-Awaji earthquake (from Cabinet Office, Government of Japan, 2006)

For the visualization, we utilized KACHINA CUBE (KC) system that is a web-based platform allowing to store, plot, and display information in three-dimensional space (Saito, Ohno, & Inaba, 2009; Ohno, Saito, & Inaba, 2010). Manually inputting data on the Great Hanshin-Awaji earthquake into the KC system, we designed the KC output which contained a 2D geographical map of the Hanshin-Awaji area, and a time line on the vertical axis. Segmented textual information from the database was plotted as information fragments into three-dimensional space. More concretely, the textual information including names of municipality was plotted in the space where a municipality was given a particular coordinate. Information fragments in each time period had each color. A total of 1,000 fragments were created in the KC system.

The KC system operated on three main functions (Figure 2). First was to access the original articles in the database by clicking information fragments plotted as segmented information. Second was to allow the cube in the system to rotate so that viewers could locate each information fragment on the map according to time period, at different angles. Finally, the system was equipped with a search function to retrieve the information that was specifically interested in by the users.

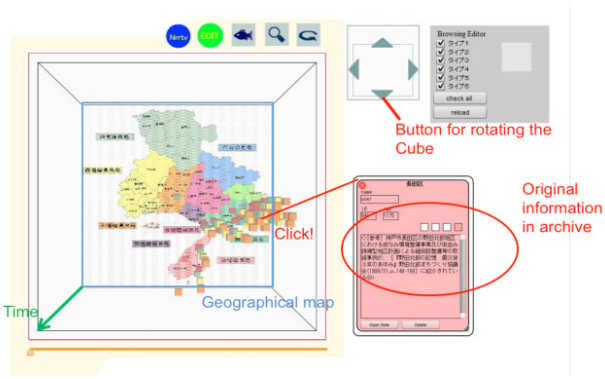


Figure 1.
Visual image of KACHINA CUBE output (as seen from the top) and its functions of rotating and accessing original information of a fragment

One of the outcomes of the visualization of the study 1 was inferred by paying attention to this quantitative distribution of articles in the geographical map. Comparing the number of articles across particular places and time periods reveals how textual information on the Hanshin-Awaji earthquake was accumulated. Few information fragments were related to the Nagata area in periods 1, 2, and 3 (Figure 3). During these periods, the Nagata area experienced the most damage, with the large number of houses destroyed in Kobe city (City of Kobe, 2010). However, the Nagata area did not have as much information as Nishinomiya city or Ashiya city which is close to the Nagata area. Thus, we could hypothesize that gaining information from severely damaged places was more difficult.

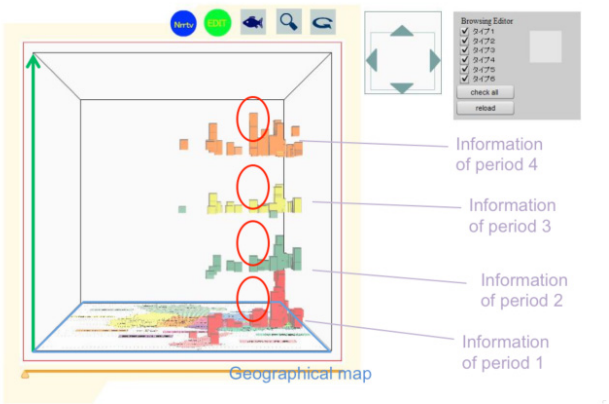


Figure 2.
The quantitative change in information fragments related to the Nagata area from periods 1 to 4. The fragments in red circles are the articles on Nagata area in each period (the KACHINA CUBE system as seen from the front)

2.2 Study 2: Visualizing the survey data on the Great East Japan earthquake

We also created a 3D visualization of the free textual responses in a questionnaire survey data on the Great East Japan earthquake by using the KC system. The survey was conducted by Kashima city, which is located in Ibaraki prefecture in the east coast Japan. The aim of the survey was to preserve and hand down the experiences on the Great East Japan earthquake to next generation and to gain useful information for making a disaster prevention plan (Kashima city, 2012).

As the target and database of the visualization, the textual information of the responses in the survey was sorted and organized into 14 areas in the city. The textual information of the responses was also separated and organized by the time periods of past, present and future. Showing the definition of the time periods, textual information in the past includes the experiences when the Great earthquake happened, the one in the present includes the experiences when people are in evacuation center, and the one in the future includes the demands or wishes for the future in relation to the earthquake. The contents of the textual information included the descriptions of the situation immediately after the earthquake, the behavior and actions taken in evacuation, recovery of infrastructure and procurement of foods and goods. An example of the table in the original survey report was in Figure 3.

Figure 3.
Part of table including textual information of the participants' responses (In the figure above, the textual information was sorted and organized by time periods and areas of the city, and by colors for showing each of contents)

For the visualization, we utilized the KC having a 2D geographical map of the Kashima city, and a time line on the vertical axis. One area had a particular coordinate in the space of the cube, and the textual information as information fragments was manually put into the space of each area. The color of the information fragments was classified according to the contents of the information as in the Figure 3.

Similar to the outcome of the study 1 above, the lopsided gathering of information fragments was seen in the study 2. Taking an overview to the distribution of information fragments in KC, the quantity of information fragments in future time period was larger than those of the other time periods (Figure 4 and 5). In order to understand why the quantity of information fragments was larger in the future time period, we focused on the Hirai area where the tendency was remarkable and explored what was happening in each time period by accessing the original texts in KC. As the results, in the past time period which was immediately after the earthquake hit, we could find a voice in an information fragment that presented the experience of feeling difficulty in catching wireless broadcast for evacuation in Hirai area. There was also a voice showing that it was unclear of the primary school as an evacuation center. Even in the present time period in the life in evacuation center, a voice showed that it is not easy to catch wireless broadcast clearly. In the future time period, there were voices showing the participants' demands of giving clear pictures of evacuation routes to evacuation center, keeping evacuation route in good condition, setting evacuation center to the familiar or close place and conducting a disaster drill. From these results, we were able to recognize the possibility of the rise in the citizens' awareness of disaster evacuation in Hirai area.

Figure 4.
Image of KACHINA CUBE output in the study 2, as seen from the top

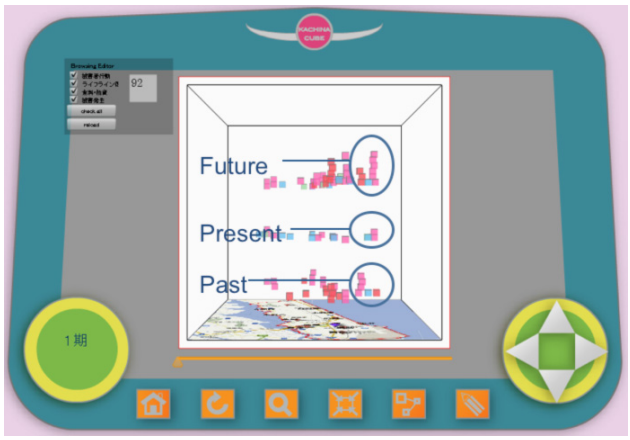


Figure 5.
The quantitative distribution of information fragments in Hirai area in each time periods.

3. Discussions

The outcome of the visualization of the database on the Great Hanshin-Awaji earthquake was gaining the factual information on phenomena, whereas the visualization of the survey data in Kashima city on the Great East Japan earthquake allowed us to find the demands or views for the future as well as the facts in the past. If taking the perspective for extracting narratives and learning lessons, the outcome having future views is more contributing to gaining meaningful story. As presented in the visualization with the survey data in the study 2, therefore, the future-oriented perspective would be recommended to add into the process of designing archive, in order for leaning lessons for the next natural disaster.

In addition to the suggestion of future-oriented archive for gaining lessons, the uniqueness of the present study is

having both processes of extracting and displaying stories or narratives in visualization. As equal to conveying stories with visualizations (e.g. Segel & Heer, 2010), extracting the stories or narratives from the vast amount of information would also be important to effectively leverage textual information in archive. The visualization with the KC system will be also useful as a web-based visualization tool for supporting the research collaborations or interactions such as Heer and Agrawala (2008)have pointed out.

References

Bamberg, M. (2012). Narrative analysis. In Cooper, H., P. M. Camic, D. L. Long, A. T. Panter, D. Rindskopf, and K. J. Sher (eds.) *APA handbook of research methods in psychology* Washington, DC: American Psychological Association. 85-102.

Cabinet Office, Government of Japan (2006). *Information database of the Great Hanshin-Awaji earthquake.* http://www.bousai.go.jp/1info/kyoukun/hanshin_awaji/index.html

City of Kobe (2012). *The Great Hanshin-Awaji earthquake statistics and restoration progress.* http://www.city.kobe.lg.jp/safety/hanshinawaji/revival/promote/index-e.html

Heer, J. and M. Agrawala(2008). Design consideration for collaborative visual analytics. *Information Visualization,* 7: 49-62.

Planning Division of Kashima City (2012). *Survey report collection on the Great East Japan Earthquake.* http://city.kashima.ibaraki.jp/info/detail.php?no=5671 (accessed 17 October 2012).

National Diet Library, Japan *Archiving Project of The Great East Japan Earthquake.* http://www.ndl.go.jp/jp/311earthquake/disaster_archives/index.html (accessed 17 October 2012).

Reischauer Institute of Japanese Studies. *Digital archive of Japan's 2011 disasters*. http://rijs.fas.harvard.edu/earthquake/index.php (accessed 17 October).

Tohoku University, The Research Group on Disaster Prevention and Management. *"Michinoku-Shin-Roku-Den" Digital Archive Project of The 2011 Great East Japan Earthquake Disaster.* http://shinrokuden.irides.tohoku.ac.jp/shinrokuden/summary (accessed 17 October 2012).

Ohno, S., S. Saito, and M. Inaba(2010). A platform for mining and visualizing regional collective culture. In Ishida, T. (ed), *Culture and Computing, LNCS (Lecture Notes in Computer Science)*, 6259:189-199. Berlin: Springer.

Saito, S., S. Ohno, and M. Inaba(2009). A platform for visualizing and sharing collective cultural information. *International Conference Digital Archives and Digital Humanities* held in Taipei, Taiwan.

Segel, E., and J. Heer (2010). Narrative visualization: Telling stories with data. *IEEE Transactions on Visualization and Computer Graphics*. 16: 1139-1148.

Uncovering the "hidden histories" of computing in the Humanities 1949–1980: findings and reflections on the pilot project

Nyhan, Julianne
julianne.nyhan@gmail.com
University College London, United Kingdom

Welsh, Anne
anne.welsh@gmail.com
University College London, United Kingdom

Introduction

Despite the relative longevity of Digital Humanities (its origins are usually pinpointed to at least 1949, when Fr Roberto Busa, an Italian, Jesuit priest, began work on an *index variorum* of some 11 million words of medieval Latin in the works of St Thomas Aquinas and related authors) very little is known of history. In this paper we will present the key findings of the pilot stage of the 'Hidden Histories' project, which is building an oral prosopography that explores the social, cultural and intellectual factors that helped to shape the work of Digital Humanities scholars since c.1949. We will also demonstrate that Oral History is an important and productive methodology in such research and that it has resulted in the creation of primary sources that not only offer new information and interpretations of the field but that can also be used for further historical research.

Research context and key aims

Contributions and notes towards a history of the field have been appearing since at least 1996 (for example, Fraser 1996; McCarty 2003; Hockey, 2004). The most substantial contribution published to date, that of Hockey

(2004), is a chronological outline that emphasises key developments such as TEI. As welcome and important as such contributions are they neither are, nor aim to be, comprehensive histories of the field. As McCarty has argued "For computing to be *of* the humanities as well as *in* them, we must get beyond catalogues, chronologies, and heroic firsts to a genuine history. There are none yet." (McCarty 2008: 255).

Nevertheless, more recent and ongoing research is enabling us to fill in aspects of the broad outlines of such a history. Barnet has researched, *inter alia*, the evolution of the Memex (2009) and the hypertext editing systems HES and FRESS (2010); Rockwell et al (2011) have presented on the incunabular history of computing in Canada and both Willard McCarty (forthcoming, see http://www.mccarty.org.uk/) and Edward Vanhoutte (forthcoming, see http://www.edwardvanhoutte.org/onderzoek/index.htm) are at work on book-length studies of DH literary history and editions.

Looking to the history of computing we can identify a number of the technological developments that gave rise to the possibility of viewing computers as more than mere number crunchers that were unsuitable for use in Humanities research. Given the nature of the Humanities, and its long standing oppositions — between, *inter alia*, techné and episteme; lone scholars and collaborative teams; discipline and interdiscipline— we hold that the very fact that the computer was used in Humanities research at all is as significant as the results it has yielded. However, the social, cultural and educational factors that helped to shape the early uptake of computing in the Humanities are little understood; it is this gap that the project sought to address. A key aim of the pilot was to investigate the appropriateness of Oral History as a methodology for capturing memories, observations and insights that are rarely recorded in the scholarly literature of the field. Accordingly, we carried out a number of pilot interviews in order to test and refine our methodology, aims and research questions. This paper will describe the key findings of the pilot project 'Uncovering the "hidden histories" of computing in the Humanities 1949–1980,' which has resulted in the creation of primary sources that will further research in this area (for example, McCarty, Nyhan et al forthcoming; Rockwell, Nyhan et al forthcoming; Short, Nyhan et al forthcoming; Siemens, Welsh et al forthcoming; Unsworth, Welsh et al forthcoming).

Methodology

During the pilot phase oral history interviews were carried out with ten prominent scholars in the field including, among others, Willard McCarty, Geoffrey Rockwell, Harold Short, Ray Siemens, John Unsworth. Now into a further iteration of the project the range of people who are being interviewed has been extended. This is to ensure a better balance between the contributions of emeritus, established, mid- and early career scholars as well as gender. Also being interviewing are those who have not necessarily followed a purely academic career path but have worked in research management, funding, foresight and strategy and service or support positions.

The chief methodology of the project was an oral history one. Oral history has been defined as "the investigation of the past by means of personal recollections, memories, evocations or life stories, where the individual talks about their experiences, attitudes and values to a researcher" (Hitchcock and Hughes 1995, 220). Semi-structured interviews were based around the following questions:

1) Please tell me about your earliest memory of encountering computing technology
2) Did you receive formal training in programming or computing?
3) How did you first get involved in what we now refer to as Digital Humanities?
4) Which people particularly influenced you and how?
5) What about scholars who were not using computers in their research do you have some sense of what their views about humanities computing were?
6) What was your first engagement with the 'conference community' and how did that come about?

Content analysis was initially investigated in order to analyse the interviews; however, this was ultimately rejected in favour of a close reading approach. As will be explored in greater detail in the full paper one of the chief strengths of an oral history methodology is its propensity to expose, rather than gloss over the heterogeneity, dissent and difference that is an integral part of the history of computing in the humanities. In seeking to move away from the hitherto evolutionary and catalogue-based approach to the history of computing in the humanities that has often been pursued it seemed especially important to give due regard to this.

Key themes that emerged from the interviews

The forthcoming publications referred to above are to transcripts and audio recordings of five interviews carried out during the pilot phase. The paper will present the key findings of the pilot project based on an analysis of all of the interviews carried out. Specifically we will:

(i) Identify and reflect on the advantages and disadvantages of using an Oral History methodology for such research

We will also explore in detail two of the key themes that emerged from the interviews and their implications:

(ii) The interviews revealed that historically, both the entry routes through which people entered the field and the levels of formal training in computing that they had access to differ substantially. For example, Willard McCarty did not have access to formal training in computing when he was at Reed college in the 1960s, instead he learned programming 'by doing' (McCarty, Nyhan et al forthcoming). In contrast, Ray Siemens, had access to formal training in computing when he was in the English department of the University of Waterloo in the mid-1980s (Siemens, Welsh et al forthcoming). All data collected on this theme will be presented and possible implications discussed. For example, looking to the present time, as more and more dedicated Digital Humanities courses are being founded in countries such as the United Kingdom, Germany, Canada and the USA etc, students have the opportunity to undertake formal and focused training in the subject. It will be most interesting to see, in due course, how this formalisation of the routes of entry into the discipline will effect interpretations about what it is. Will we still be having the 'what is digital humanities' debate in ten years and how important will it be to understanding the early history of the emergence of the discipline?

(iii) Some interviews, for example, that carried out with Harold Short reveal the importance of the social, intra-personal and administrative work that appears to have been so central to the establishment and institutionalisation of Digital Humanities. Nevertheless, this work tends to be of lower profile that formal scholarly research and is rarely explored in historical studies of the field. The data collected on this will be presented and its wider comparative context with other emerging disciplines will be explored.

Conclusion

The history of Digital Humanities is a research topic that has been neglected not only by those working in long established Humanities subjects but also by the Digital Humanities community itself; much interesting and essential work in this area remains to be done. This paper will present the new findings that the Hidden Histories project has uncovered and will also argue that Oral History has a central role to play in such research. Furthermore, it will draw attention to the pressing need that exists for histories of Digital Humanities.

References

Barnet, B. (2010). Crafting the User-Centered Document Interface: The Hypertext Editing System (HES) and the File Retrieval and Editing System (FRESS). *Digital Humanities Quarterly* 4.1

Barnet, B. (2008). The Technical Evolution of Vannevar Bush's Memex. *Digital Humanities Quarterly* 2.1

Fraser, M. (1996). A Hypertextual History of Humanities Computing. http://users.ox.ac.uk/~ctitext2/history/

Hitchcock, G., and D. Hughes. (1995). *Research and the teacher: a qualitative introduction to school-based research*. Routledge.

Hockey, S. (2004). The History of Humanities Computing. In Schreibman, S., et al. (eds), *A Companion to Digital Humanities*. Oxford: Blackwell. http://www.digitalhumanities.org/companion/

Mahoney, M. S. (1996). Issues in the history of computing. In Bergin Jr., T. J., and Gibson, R. G. (eds), *History of programming languages — II*. New York: ACM.

McCarty, W., J. Nyhan, A. Welsh, and J. Salmon. Questioning, Asking and Enduring Curiosity: an Oral History Conversation between Julianne Nyhan and Willard McCarty. *Digital Humanities Quarterly* [Forthcoming]

McCarty, W. (2003). Humanities Computing. In *Encyclopedia of Library and Information Science*. New York: Marcel Dekker. 1224-1235.

McCarty, W. (2008). Whats going on? *Literary and Linguistic Computing* 23.3: 253-261.

Rockwell, G., J. Nyhan, A. Welsh, J. Salmon. Trading Stories: an Oral History Conversation between Geoffrey Rockwell and Julianne Nyhan. *Digital Humanities Quarterly* [Forthcoming]

Rockwell, G., V. S. Smith, and S. Hoosein, et al. (2011a). Computing in Canada: a history of the incunabular years. *Digital Humanities 2011* http://dh2011abstracts.stanford.edu/xtf/view?docId=tei/ab-311.xml;query=;brand=default

Short, H., J. Nyhan, A. Welsh, and J. Salmon Collaboration must be fundamental or it's not going to work: an Oral History Conversation between Harold Short and Julianne Nyhan. *Digital Humanities Quarterly* [Forthcoming]

Siemens, R., A. Welsh, J. Nyhan, and J. Salmon. Video-gaming, Paradise Lost and TCP/IP: an Oral History Conversation between Ray Siemens and Anne Welsh. *Digital Humanities Quarterly* [Forthcoming]

Unsworth, J., A. Welsh, J. Nyhan, and J. Salmon. Postmodern Culture and more: an Oral History Conversation between John Unsworth and Anne Welsh. *Digital Humanities Quarterly* [Forthcoming]

Joint and multi-authored publication patterns in the Digital Humanities

Nyhan, Julianne
julianne.nyhan@gmail.com
University College London, United Kingdom

Duke-Williams, Oliver
o.duke-williams@ucl.ac.uk
University College London, United Kingdom

Introduction

A frequently claimed hallmark of Digital Humanities is its emphasis on collaborative work and joint publication (see, for example, Moulin, Nyhan et al 2012; Koh 2012; and Deegan and McCarty 2012). Is there any mismatch between the way that the field describes itself and what we find when we examine the evidence of publication patterns and practices? Furthermore, have publication patterns changed since the first journal of the field, *Computers and the Humanities,* was established in 1966? Has joint publication become more or less common or have the proportions of jointly published articles remained the same? Also, how do such patterns compare with other disciplines of the Humanities?

Research context

To the best of our knowledge the empirical evidence of publication practices of Digital Humanities scholars has not, until our research, been systematically investigated. In order to make a first contribution towards addressing this gap in the research literature we focused our research on publication patterns in the leading Digital Humanities journals since 1966. We hope to extend our analysis to other Digital Humanities journals (for example *Computing in the Humanities Working Papers*), identify and analyse publications in non-specialist Digital Humanities journals and contributions to e.g. book collections in a future iteration of this research. We began by harvesting all bibliographical metadata from *Computers and the Humanities* (1966-2004), *Literary and Linguistic Computing* (1986-2011) and *Digital Humanities Quarterly* (2007-2011) and then analysed it in order to explore the following questions:

- What can we observe about patterns of joint publication?
- What percentage of articles per issue and per journal was jointly published?
- What kinds of joint publication patterns existed? What percentage of joint publications had 2 authors? What percentage had 3, 4, 5 or more?
- How many authors contributed to more than one joint publication?
- Of those who published jointly what patterns can be observed? Did they tend to contribute to papers authored by two or more authors?
- Did authors tend to publish with predominately the same people over their careers or do we see a number of shifting constellations?
- What percentage of people published only one article in the journals listed above and based on this do we see large portions of people dropping in and out of the field at particular times?

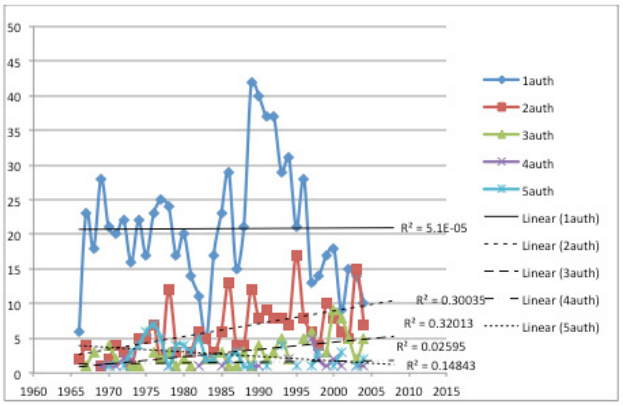
Methodology

Using Zotero we extracted bibliographical metadata from *Computers and the Humanities* (Chum)(1966–2004); *Literary and Linguistic Computing*(LLC) (1986-2011); and *Digital Humanities Quarterly* (DHQ) (2007-2011) and then exported this to Excel for initial viewing. As far as possible the data was cleaned and regularised e.g. a canonical form of personal names chosen where slight differences existed such as E.G. Wills and Edward G. Wills. The cleaned data were then imported into an SQL database and sorted into groups based on the number of authors. The annual observed frequencies of papers with *n* authors were then calculated. For each group, a linear regression was calculated in order to determine within a given journal whether the incidence of *n*-authored papers had changed over time.

For each journal the data were also processed so that dual-authored papers could be analysed using a connectivity index (Bell et al 2002)to determine the extent to which the pool of authors contributing to a given journal were interconnected. A connectivity index was constructed both on a journal-wide basis and on a per-author basis, allowing the distribution of 'well-connected' authors to be compared within and between journals.

Findings related to Chum

The diagram below shows the key findings in relation to *Chum;* space will not allow the findings in relation to *LLC* and *DHQ* to be presented, and all three journals to be compared and contrasted, so this will be done in the full paper. In relation to *Chum*, contrary to what one might expect given digital humanities emphasis on collaborative work, the highest incidence is of single-authored papers, but the frequency of this is variable. We should also remember that this does not necessarily mean that the research was done by a single scholar — all we can say is that publications were predominately by a single scholar. Considering trends over the lifetime of *Chum* we see that single authored papers are flat, whereas frequency of 2 and 3 authored papers increases over time. The strength of this association was examined using regression analysis. For *Chum*, the observed frequency of 2 and 3 authored papers increased over time, and this was significant at the 1% level. The frequency of 5 authored papers decreased over the observed time period, and this relationship was significant at the 5% level. However the overall number of 5 authored papers was low throughout.



Frequency of *n*-authored papers, by year

Findings related to LLC

The same analysis was conducted for LLC (graphs to be shown in the full presentation). Again, the most common form of contribution was of sole-authored papers. However, regression analysis highlighted some notable trends. The frequency of sole-authored papers was found to be decreasing over time, and this relationship was significant at the 5% level. As with *Chum*, the frequency of 3-authored papers was found to be increasing, significant at the 1% level. The frequency of papers with 2, 4 and 5 authors showed some increase over time, but these relationships were not significant.

Conclusion

In relation to *Chum* we have found that single-author publications were predominant for much of the lifetime of the journal. However, this does not necessarily mean that Digital Humanities does not have a higher occurrence of joint publications than other disciplines of the Humanities. Therefore, it is important to not only examine the empirical evidence that exists for publication practices in the Digital Humanities since 1966 at the aggregate level but to attempt also to situate such findings in a wider comparative context.

It should be noted that authorship as reflected in publication credits does not necessarily reflect actual contribution to research: a range of alternative practices might occur, from papers that only carry a single name despite substantial contributions from others, to papers that include 'gift' attributions to persons who have had little or no input(Cronin et al 2003).

The stereotype of the Humanities lone scholar is well known, even if it is increasingly recognised as being an impoverished model(Bulger et al 2011). The rate of publication and productivity in the Humanities has been looked at by Muffo et al(1987); Ramsden(1994); Wanner, Lewis & Gregorio (1981) and Stone (1982) examined, inter alia, "the way humanities scholars work and the materials of their research". Changing publication patterns in the Humanities in Flanders and Belgium have been analysed by Engels et al who found that in the period 2000-09 "The overall growth rate in number of publications is over 62.1%, but varies across disciplines between 7.5 and 172.9%. Publication output grew faster in the Social Sciences than in the Humanities."(2012). In 2003 Kyvik found that in Norwegian Universities co-Authorship has become more common but it is difficult to determine from the article to what degree this applies to the Humanities. Lariviere et al used data from the CD-ROM versions of the *Science Citation Index, Social Sciences Citation Index* and the *Arts & Humanities Citation Index* from 1980 to 2002, to argue that "contrary to a widely held belief, researchers in the social sciences and the humanities do not form a homogeneous category. In fact, collaborative activities of researchers in the social sciences are more comparable to those of researchers in the [natural sciences and engineering] than in the humanities"(2006).

In essence, looking to the scholarly literature in both Digital Humanities and the wider Humanities context relatively few studies have been undertaken based on the empirical data that exists about joint publication patterns.

This short paper will take a first step towards remedying this.

References

Bell, M., M. Blake, P. Boyle, O. Duke-Williams, P. Rees, J. Stillwell, and G. Hugo (2002). Cross-national comparison of internal migration: issues and measures. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 165: 435–464. doi: 10.1111/1467-985X.t01-1-00247

Bulger, Monica, et al. Research Information Network (2011). *Reinventing research? Information practices in the humanities.* Research Information Network, UK.

Cronin, B., D. Shaw, and K. La Barre (2003). A cast of thousands: Coauthorship and subauthorship collaboration in the 20th century as manifested in the scholarly journal literature of psychology and philosophy. *Journal of the American Society for Information Science* **54** 855–871. doi: 10.1002/asi.10278

Deegan, M., and W. McCarty (2012.) *Collaborative research in the digital humanities: a volume in honour of Harold Short, on the occasion of his 65th birthday and his retirement, September 2010.* Farnham: Ashgate.

Engels, T. C. E., T. L. B. Ossenblook, and E. H. J. Spruyt (2012). Changing publication patterns in the Social Sciences and Humanities, 2000-2009. *Scientometrics.*

Koh, A. (2012.) The Challenges of Digital Scholarship. *ProfHacker,* http://chronicle.com/blogs/profhacker/the-challenges-of-digital-scholarship/38103 (accessed 25 January 2012).

Moulin, C., J. Nyhan, et al. (2011) ESF: Science Policy Briefing. *Research Infrastructures in the Digital Humanities*

Kyvik, S. (2003). Changing trends in publishing behaviour among university faculty, 1980-2000. *Scientometrics,* **58:1**.

Lariviere, V., Y. Gingras, and Archambault, E. (2006). Collaboration networks: A comparative analysis of the natural science, social sciences and the humanities. *Scientometrics,* 68(3): 519-533.

Muffo, J. A., S.V. Mead, and A.E. Bayer (1987). Using faculty publication rates for comparing "peer" institutions. *Research in Higher Education,* **27.2** 163-175.

Ramsden, P. (1994). Describing and explaining research productivity. *Higher Education,* 28(2): 207-226.

Stone, S. (1982). Humanities Scholars: Information Needs and Uses. *Journal of Documentation,* 38(4): 292–313.

Wanner, R. A., L.S. Lewis, and D.I. Gregorio (1981). Research Productivity in Academia: A Comparative Study of the Sciences, Social Sciences and Humanities. *Sociology of Education,* 54: 238-253.

Incidental Crowdsourcing: Crowdsourcing in the Periphery

Organisciak, Peter
organis2@illinois.edu
University of Illinois at Urbana-Champaign, United States of America

As the customs of the Internet grow increasingly collaborative, crowdsourcing offers an appealing frame for looking at the interaction of users with online systems and each other. However, it is a broad term that fails to emphasize the use of crowds in subtler system augmentation.

This paper introduces incidental crowdsourcing (IC): an approach to user-provided item description that adopts crowdsourcing as a frame for thinking about augmentative features of system design. IC is intended to frame discussion around peripheral and non-critical system design choices.

A provisional definition of incidental crowdsourcing will be defined in this paper, and then refined based on examples seen in practice. IC will be examined from both the user and system ends, positioned within existing work, and considered in the context of its benefits and drawbacks. This approach allows us to explore the robustness and feasibility of IC, looking at the implications inherent to accepting the provisional definition.

The consequences of considering system design on a scale between IC and non-IC design choices remain to be seen. Toward this goal, the second part of this paper shows a study comparing the participation habits of users in two online systems — one that is representative of IC properties and one that is not. This study finds differences in user engagement between the two systems.

Introduction

Crowdsourcing asks a dispersed group of people to contribute toward a common task. It does not need to be the central feature of a project; it can be used for augments parts of a project. For example, Facebook[1] uses "Likes" to gauge popularity of user-generated content, while photo-sharing website Flickr[2] uses user labeling to improve their search engine.

Incidental crowdsourcing functions in in this way, capturing useful but unobtrusive user input and making sense of it in aggregate. An incidental crowdsourcing feature is supplemental to its site's primary function. Thus, visitors to the website are gently given ways to make a contribution, but not forced into it. IC offers a way to consider the design of online systems through the lens of crowdsourcing, which offers a compelling framework for gathering abstract, perception-based, or conceptual data in a volunteer-driven and often mutually beneficial way.

The value of IC is largely in augmenting existing information, making it valuable in the digital humanities for enriching digital resources. Version 1.0 of Digital Humanities Now, for example, used implicit linking by DH scholars on Twitter to determine the quality of online information (Cohen 2009). Another system dealing with digital resources, citation manager Mendeley, takes an IC approach in improving metadata and predicting research trends (Henning et al. 2010).

Defining incidental crowdsourcing

Incidental crowdsourcing is the gathering of contributions from online groups in an unobtrusive and non-critical way.

It is *unobtrusive* in that it does not cause significant barriers to a user's completion of a task. The corollary to this is that IC exists in a task-driven environment where the user has a primary objective and IC exists alongside it without causing resistance to it.

IC is also *non-critical* to users and systems. For users, making a contribution is not a necessary part of a their use, while systems should not rely on contributions to function, using them for value-added features but degrading gracefully when there are few or unevenly distributed contributions.

This provisional definition is expanded in the full paper by considering complementary characteristics of examples that fit the definition. This refinement expands the definition to note that contributing to IC is *descriptive* — producing data about existing information objects, contributions tend to toward *low granularity*, and systems favor *choices over statements*.

IC is best considered as a scale, where the IC fitness of a crowdsourcing system design element is a mixture of how well it conforms to each part of the above definition.

Action	Examples
Rating	Rating the quality of online content, Rating helpfulness of online comments or reviews
Classification / Curation	Tagging, labeling, adding to lists
Saving / Recommending	Starring, liking/recommending, adding to favorites
Editing	Translating content, correcting grammar or spelling
Feedback	Flagging online comments as inappropriate, "Did you find this helpful?"
Other	Commenting, sharing, encoding

Table 1:
Common forms of incidental crowdsourcing and examples

Following from the provisional definition, Table 1 shows common IC actions, alongside examples of how they are implemented. The full paper outlines these actions relative to their use in digital humanities. These include:

Scoring the quality of an information object. Rating or ranking systems that conform to the definition of IC tend to be on the lower end of granularity, most often using five- or two-point scales. Unary rating mechanisms are also used, for saving or supporting information items in online systems. For example, Facebook's "Like" buttons allow users of the social network can make an assertion on the quality of an item. Rating systems tend to skew upward (Hu et al. 2006, Banjeree and Fudenberg 2004), and single button saving features are generally positive. Implicit recommendation is another valuable indicator of support; for example, it has been used to discover notable web resources through microblogging links (Cohen 2009).

Organizing content. Curatorial features are a way for users to thematically group information objects in a way that can teach a system about the relationships between those objects. For example, newer OPAC replacements encourage IC classification and curation with patron-built book lists, ratings, and tagging (Singer 2008, Spiteri 2011). Such catalogues can be interaction points rather than simply retrieval systems, but participation is non-critical to users.

Editing content. Incidental crowdsourcing is sometimes used to switch user roles from consumer to creator. The Australian Newspaper Digitization Project implemented this approach in corrected OCR transcriptions of old newspapers (Holley 2009), offering a link to the editing interface from the newspaper reading screen.

Feedback. Simply asking users questions which they have the capacity to answer has been noted as a strong motivator for contribution (Kraut and Resnick 2012,

Organisciak 2010), and feedback mechanisms often make use of this with direct questions and easy to choose answers.

User- and System-end considerations

Since IC contributions are non-critical, systems utilizing IC should degrade gracefully when there is a lack of contributions. A system dealing with IC contributions should not be dependent on large or evenly distributed data sets. For example, the transit tracking application *Tiramisu* (Zimmerman et al. 2011, Tomasic et al. 2011) aggregates the location of riders when it is being used, but falls back on historical information when real-time data is unavailable.

Table 2 considers common IC actions and the corresponding value to a system and its users. Notably, in the majority of cases the user's act of contributing is one of description rather than creation. Systems primarily use IC for understanding the content within them, while users primarily contribute to fulfill personal and social needs.

Action	User Use	System Use
Tagging a photo/bookmark	Easy personal retrieval, appeal of collecting, item grouping for easy sharing	Improved search, improved browsing
Rating an product	Sharing opinion	Improved recommendations, prioritize good items
Rating/starring a digital item, starring	Sharing opinion, communicating approval, saving for reference	Identifying quality content
Flagging content	Altruism, catharsis	Higher signal-to-noise in maintenance
Sharing	Showing items to friends, curating content	Identifying popular/interesting content
Feedback	Sharing personal knowledge and opinions, altruism	Correct problem data, discover system issues

Table 2

Comparison of design choice in product ratings

How would an evaluation of systems through the lens of IC look? As an example, I compared the rating patterns of two application marketplaces—Amazon Appstore[3] and Google Play. From each store the lists of best-selling free and paid items were scraped and parsed, and the applications that were on the lists of both sites were matched while others were removed from the data.

The sites were chosen because they are alike in purpose, selling applications for the Android operating system, and much of the same content is represented between them. However, how each store allows users to rate their purchases differs. Google's rating functionality is more aligned with IC, allowing users to rate an item with two clicks on one page. Meanwhile, Amazon is non-IC, asking raters to include title, reviews, and to abide by a codebook.

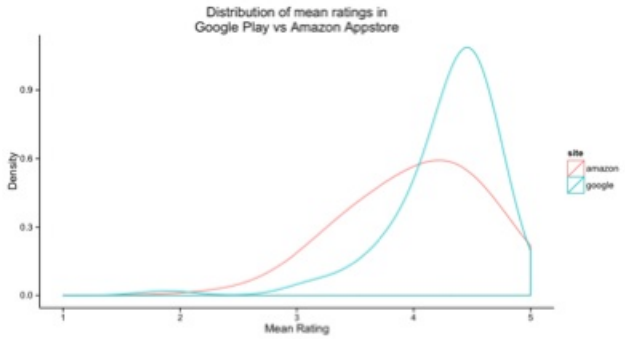


Figure 1

This study found that the distribution of mean ratings skewed higher for Google Play than for Amazon Appstore (Wilcoxon $p<0.001$, see Figure 1). The difference in rating style exists even though there is no difference between the systems in how likely an application is to be rated (T-test $p=0.9873$, $H^o:\mu_{diff}=0$). Breaking the rating distributions down by relative choice frequency (Figure 2) shows a clear pivot at four stars.

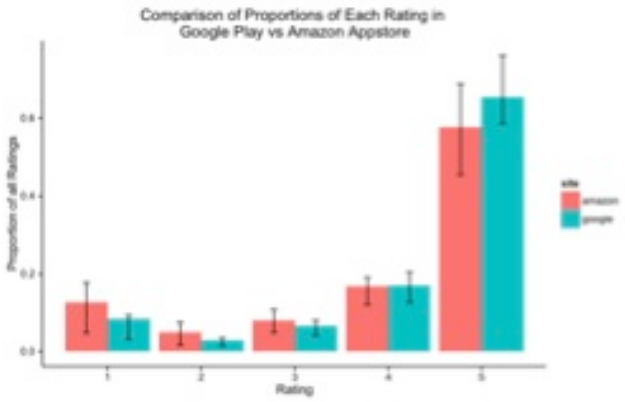


Figure 2-1

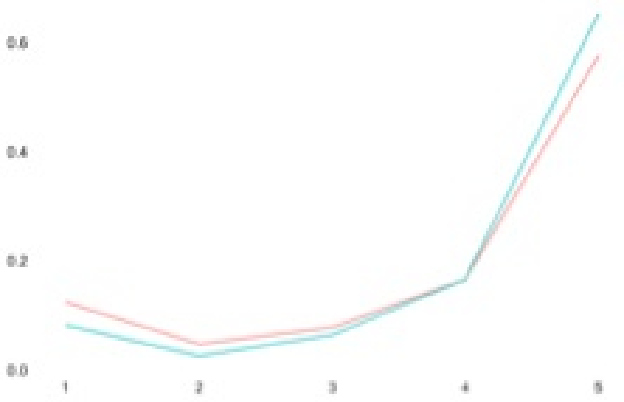


Figure 2-2 (Simplified)

The distinct pivot between the distributions suggests that an adjustment can make Google's distribution — collecting in a more IC appropriate manner — nearly identical to Amazon's. Thus, while the non-IC approach receives more written reviews, Google does not appear to sacrifice rating quality with easier ratings. This could make a difference when looking to measure quality of new or barely-seen items.

Conclusion

This paper introduces the concept of incidental crowdsourcing, a way to crowdsource in a way that is non-critical, descriptive, unobtrusive and peripheral. Incidental crowdsourcing matters as a way to adopt crowdsourcing practices to reflect the subjective 'humanness' of digital object interpretations by consumers.

References

Banerjee, A., and D. Fudenberg (2004). Word-of-mouth Learning. *Games and Economic Behavior* 46(1). Web. 7 Dec. 2011.

D. Cohen. (2009). Introducing Digital Humanities Now. 18 Nov. 2009.

Henning, V., J. J. Hoyt, and J. Reichelt (2010). Crowdsourcing Real-Time Research Trend Data. Raleigh, USA. Web. 1 Nov. 2012.

Holley, R. (2009). Many Hands Make Light Work: Public Collaborative OCR Text Correction in Australian Historic Newspapers. National Library of Australia. National Library of Australia Staff Papers.

Hu, N., P. A. Pavlou, and J. Zhang (2006). Can Online Word-of-mouth Communication Reveal True Product Quality? Experimental Insights, Econometric Results, and Analytical Modeling. *Proceedings of the 7th ACM Conference on Electronic Commerce-2006*. 324–330.

Kraut, R.E., and P. Resnick (2012). Encouraging Contribution to Online Communities. *Designing From Theory: Using the Social Sciences as the Basis for Building Online Communities.*

Organisciak, P. (2010). Why Bother? Examining the Motivations of Users in Large-scale Crowd-powered Online Initiatives. 31 Aug.

Singer, R. (2008). In Search Of A Really 'Next Generation' Catalog. *Journal of Electronic Resources Librarianship*. 20(3). 139–142. Web. 1 Nov. 2012.

Spiteri, L. F. (2011). Social Discovery Tools: Cataloguing Meets User Convenience. *Proceedings from North American Symposium on Knowledge Organization*. 3.

Tomasic, A., et al. (2011). Design Uncertainty in Crowd-Sourcing Systems.

Zimmerman, J., et al. (2011). "Field Trial of Tiramisu: Crowd-sourcing Bus Arrival Times to Spur Co-design." *Proceedings of the 2011 Annual Conference on Human Factors in Computing Systems. CHI '11* held in Vancouver, BC. ACM. 1677–1686. Web. 4 Dec. 2011.

Notes

1. http://www.facebook.com
2. http://www.flickr.com
3. http://www.amazon.com/appstore
4. http://play.google.com

eResearch Tools to Support the Collaborative Authoring and Management of Electronic Scholarly Editions

Osborne, Roger
r.osborne@uq.edu.au
University of Queensland, Australia

Gerber, Anna
a.gerber@uq.edu.au
University of Queensland, Australia

Hunter, Jane
j.hunter@uq.edu.au
The University of Queensland

1. Introduction

The *Australian Electronic Scholarly Editing* project is a collaboration between the University of Queensland, University of NSW, Curtin University, University of Sydney, Queensland University of Technology, Loyola University, Chicago and the University of Saskatchewan. The aim of the project is to develop a set of interoperable services to support the production of electronic scholarly editions by distributed collaborators in a Web 2.0 environment.

One of the fundamental challenges faced by the AustESE project is the development of an interoperable data model. In recent years, research focussed on the production and use of electronic scholarly editions has increasingly involved the development and employment of ontologies (Robinson and Meschini 2010; Romanello et al. 2009). The Text Encoding Initiative has provided the necessary elements to describe the textual and material character of documents, but TEI has addressed neither the naming of components of an edition nor the relationships between these components (Robinson & Meschini 2010). Such frameworks are necessary to support the interoperability of the electronic edition and to facilitate and coordinate greater levels of user engagement with annotations.

One of the most significant contributions that ontologies can make to electronic scholarly editions is to more precisely model and capture the dynamic nature of the 'work'. This provides a more stable framework onto which metadata and annotations can be accurately attached, and also facilitates the efficient execution of workflows that support scholarly editing practices. The importance of acknowledging the complexity and the contingency of the work has been made clear in recent years (Shillingsburg 1997, Eggert 2009). In order to logically integrate the elements of a work into a knowledge-site (Shillingsburg 2006) or a work-site (Eggert 2005), and, at the same time, protect the integrity of transcriptions and images, our approach enables the augmentation of images and transcriptions with stand-off mark-up in the form of annotations. Such an approach contributes not only to theories of scholarly textual editing, but also to theories of knowledge representation in computation (Clement 2011). In this paper we describe the project's overall objectives and the ontology that we are using to underpin the AustESE Workbench, and we will discuss the practical and theoretical implications of the delivery of an integrated workbench that promises to re-invigorate scholarly editing in Australia.

2. Project Objectives and Ontology

The specific objectives of the AustESE project are to provide the Australian scholarly editing community with an online integrated Workbench that provides:

- *collation tools* for automatically detecting, identifying and highlighting variations between different versions of a work and that allow the relationship between texts to be visualized, authored or edited
- *annotation tools* that:
  - enable scholars to create and reply to scholarly commentary attached to texts, variants and images;
  - capture the annotations as stand-off markup that is discoverable, shareable, and re-usable;
  - provide search, browse and visualisation interfaces for annotations;
  - enable both manual and automated migration of annotations between transcriptions and facsimiles.
- a *workflow engine* that captures the sequence of tasks and decision-making steps as well as the provenance of generating an electronic scholarly edition;
- *publishing tools* that automatically compile electronic scholarly editions into standard publication formats;
- and a *repository* that supports the discovery, search, retrieval, exploration and re-use of texts and electronic editions.

AustESE Workbench

The AustESE workbench coordinates the scholarly editing workflow and provides access to online tools that support scholarly editing tasks. For the AustESE project, we have adopted a Service-Oriented Architecture, illustrated in Figure 1, with the aim of developing modular, reusable, and potentially distributed components that can be assembled and substituted according to the requirements of each scholarly edition project. To implement this architecture, we are extending existing scholarly editing tools with REST APIs to enable their integration with our content repository and workflow engine, and implementing new open source software to bridge the gaps between existing tools.

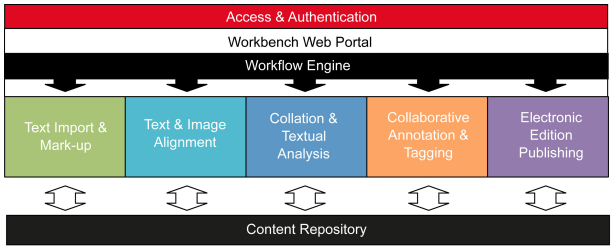


FIGURE 1:
AUSTESE TECHNICAL ARCHITECTURE

Ontologies

The AustESE ontology includes key classes Work, Version, Artefact, Agent and Event, and provides the data model used to organise the metadata and conceptual entities represented within the repository. The AustESE ontology can be mapped onto the IFLA FRBR to link with entities from related databases such as AustLit: The Australian Literature Resource, however it provides additional concepts to those defined by FRBR, to make it easier to model and analyse manuscript materials and fine-grained differences at the level of 'impressions' and 'states'.

The Open Annotation W3C Community Group provides a common data model (Sanderson, Ciccarese & Van de Sompel, 2012) for representing annotations across tools, architectures and collections. The model, which is expressed as an OWL ontology, is intended to be extensible, so that it can be refined to meet the annotation requirements of specific communities. To support the production of apparatus and commentary within electronic editions, we build on the OA core data model with specialised annotation *Motivations*, as illustrated in Figure 2. We categorise annotations as *ExplanatoryNotes*, providing commentary or *TextualNotes*, which provide support for editorial decisions. *VariationAnnotations* are a type of *TextualNote* that describe textual variation between versions of a work. These annotation *Motivations* can be used in search queries and for filtering and sorting annotations to enable selective display and inclusion for print or electronic publication.

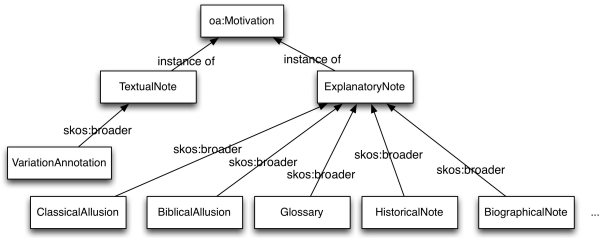


Figure 2:
AustESE Open Annotation Motivation subclasses

We have defined additional properties that may be used within the body of a *VariationAnnotation*, (as shown in the example in Figure 3), to record metadata about the agent, date or cause of the variation as well as documentary evidence including links to manuscript facsimiles. Within our RDF-based annotation tool and annotation repository, we have adopted a Linked Data approach of using HTTP URIs to identify entities that may be referenced within annotations, including *Agents* (*people* or *organisations*) and conceptual entities (*Works, Versions, Artefacts, Events*). We use FOAF and Dublin Core to record annotation provenance, and we apply properties from the AustESE ontology to relate the transcriptions and corresponding facsimile images that are being annotated.

Such a conceptual framework necessarily directs attention to the material artefacts it aims to describe, 'making the archives talk' through the arguments of editors and readers (West 2011). A 'virtual archive' of artefact images provides a foundation upon which transcription and commentary can be overlaid, satisfying the archival impulse of scholarship, and providing a space for multiple, and, perhaps, competing views about how works could and should be represented (Shillingsburg 2010). With the Workbench, the AustESE project aims to facilitate such processes and to support collaborative editorial models that contribute to the development of social editions (Siemens et al 2012).

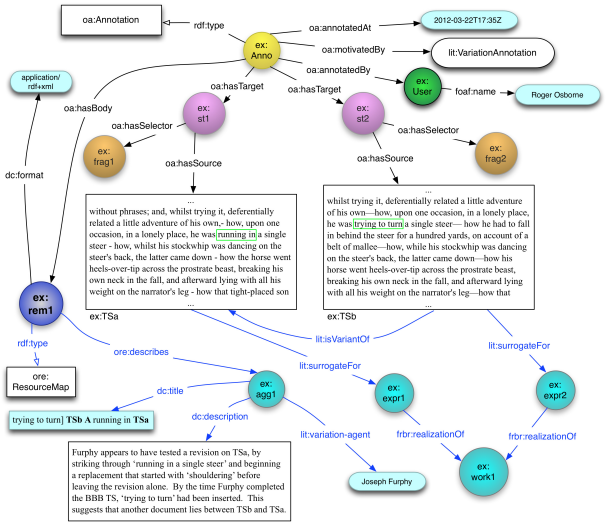


Figure 3:
Annotating Textual Variation

3. Practical and Theoretical Implications

The development of the AustESE Workbench has been informed by several case-studies, particularly the requirements of Paul Eggert's Charles Harpur Critical Archive and Roger Osborne's electronic edition of Joseph Furphy's *Such is Life*. The complex textual and material situations faced by these projects require the type of infrastructure provided by the AustESE Workbench in order to efficiently store, describe, organise and analyse large

numbers of image files and their related transcriptions. Projects such as these can still argue for a particular editorial rationale and contribute new, critically established texts to the system. But with the facility to support solitary and collaborative interpretation in the form of annotations across the archive, the editor's version can be critiqued or ignored if readers object to the editorial approach. Hans Walter Gabler has recently described the emerging phase of electronic scholarly editions as a 'paradigm of a relational interplay of discourses, dynamically correlated both among themselves and with an edition's readers and users: that is, to a paradigm once again of text and ongoing commentary.' (Gabler 2010) While granting due attention to the integrity of the images and transcriptions within the archive, such a paradigm lends itself to the compilation of 'revision narratives' (Bryant 2002) and general commentary that will help to reinvigorate and sustain research on literary works into the future. The AustESE Workbench will achieve this by drawing on Web 2.0 technologies and the Semantic Web to support collaborative social editions and/or finely argued editions produced by solitary editors.

Acknowledgements

The University of Queensland is proud to be in partnership with the National eResearch Collaboration Tools and Resources (NeCTAR) project to create a unique opportunity to develop eResearch Tools that support the Collaborative Authoring and Management of Electronic Scholarly Editions.

References

Australian Electronic Scholarly Editing Project http://itee.uq.edu.au/~eresearch/projects/austese/services.html

Bryant, J. (2002). *The Fluid Text: A Theory of Revision and Editing for Book and Screen*. Ann Arbor: The University of Michigan Press, 157-161.

Clement, T. (2011). Knowledge Representation and Digital Scholarly Editions in Theory and Practice. *Journal of the Text Encoding Initiative* 1. http://jtei.revues.org/203

Eggert, P. (2005). Text-encoding, Theories of the Text, and the "Work-site." *Literary and Linguistic Computing*. 20. 425-35.

Eggert, P. (2009). The Editorial Gaze and the Nature of the Work, in *Securing the Past: Conservation in Art, Architecture, and Literature*. Cambridge: Cambridge University Press. 214-40

Gabler, H. W. (2010). Theorizing the Digital Scholarly Edition. *Literature Compass* 7 (2). 43.

Robinson, P., and F. Meschini (2010). Works, Documents, Texts and Related Resources for Everyone, DH 2010, http://dh2010.cch.kcl.ac.uk/academic-programme/abstracts/papers/html/ab-816.html

Romanello, M. et al. Rethinking Critical Editions of Fragmentary Texts by Ontologies, Elpub, Milano, Italy, http://conferences.aepic.it/index.php/elpub/elpub2009/paper/view/158

Sanderson, R., P. Ciccarese, and H. Van de Sompel (2012). 'Open Annotation Core Data Model', http://openannotation.org/spec/core

Shillingsburg, P. (1997). Text as Matter, Concept and Action. In *Resisting Texts: Authority and Submission in Constructions of Meaning*, Ann Arbor: The University of Michigan Press. 49-103

Shillingsburg, P. (2006). An Electronic Infrastructure for representing Script Acts. In *From Gutenberg to Google*. Cambridge: Cambridge University Press. 80-125.

Shillingsburg, P. (2010). How Literary Works Exist: Implied, Represented, and Interpreted. In McCarty, W. (ed). *Text and Genre in Reconstruction: Effects of Digitalization on Ideas, Behaviours, Products and Institutions*. Cambridge: OpenBook Publishers. 165-182.

Siemens, R., et al. (2012). Toward modeling the *social* edition: An approach to understanding the electronic scholarly edition in the context of new and emerging Media. *Literary and Linguistic Computing* 27 (4). 445-61.

West, J. L. W. (2011). *Making the Archives Talk*. Pennsylvania University Press.

Linked Jazz 52nd Street: A LOD Crowdsourcing Tool to Reveal Connections among Jazz Artists

Pattuelli, M. Cristina
mpattuel@pratt.edu
Pratt Institute, United States of America

Miller, Matt
mmille18@pratt.edu
Pratt Institute, United States of America

Lange, Lea
llange@pratt.edu

Pratt Institute, United States of America

Thorsen, Hilary
hthorsen@pratt.edu
Pratt Institute, United States of America

"There's a bond, a sort of invisible bond between all musicians who play jazz. There is always that bond, it holds them together." — Ian Patterson, 2009

Introduction

Linked Jazz [1] is a Linked Open Data (LOD) project that aims to create methods and tools that reveal the dense fabric of relationships connecting the community of jazz artists who typically practice in rich and diverse social networks. This project takes advantage of the potential of LOD to connect cultural heritage data in new ways and expand traditional access to archival content by making it visible and discoverable in an open information environment. The Linked Jazz project consists of multiple phases and has progressed in an iterative and experimental process. In the first phase, a LOD dataset representing a social network of connections among jazz artists was created through the automatic extraction of personal names from interview transcripts acquired from digital archives of jazz history. Based on this dataset, a visualization tool was developed that offers static and dynamic views of the social network[2] . While the social network is effective in conveying the vast and rich web of interpersonal associations, its connections remain semantically undefined.

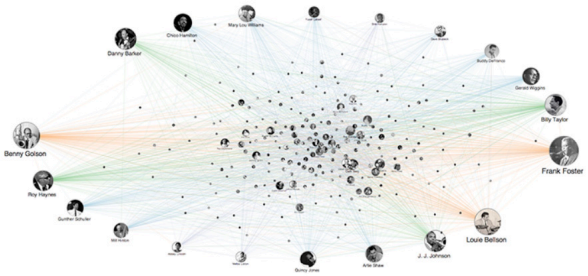


Fig. 1:
Frequency-based network rendering

Enriching the Network

Social network analysis is based on graph theory, in which entities such as individuals are presented as nodes and their relationships are represented by edges between them (Chen & Yang, 2010). While this mathematically-based approach allows for analysis of the network's structure based on parameters like centrality and clustering, typically, social graphs provide little or no explicit information about the nature of the relationship that connects two entities (Chen & Yang, 2010; Scott, 2011). The need for richer descriptions of social relationships within networks and for researchers to be able to find resources more readily has led to projects that focus on enhancing the visibility and interconnectedness of archival resources. Such projects included the development of Encoded Archival Context-Corporate Bodies, Persons, and Families (EAC-CPF)[3] and Social Networks and Archival Context (SNAC)[4] .

The jazz community is characterized by a high degree of interaction and connectivity. A few studies have examined the interconnectivity among jazz musicians. Heckathorn and Jeffri (2003) studied their affiliation patterns and found that this community is highly egalitarian and cohesive. They employed the respondent-driven sampling method, a technique that requires respondents to know and come in regular contact with one another. Approaching the study of personal connections in this community from a different angle, Schubert (2012) adopts discometrics — the application of bibliometric network techniques to discographic data — to reveal how embedded a particular musician is within the jazz community network.

Human-Generated Approach

The complex and dynamic web of interpersonal connections inherent to jazz music is well documented in books and discographies, yet not easy to discover. While a machine-driven approach combining Natural Language Techniques (NLP) techniques and LOD technology has proven effective in revealing basic connections among personal entities (Pattuelli, Weller and Szablya 2011), this approach fell short when attempting to uncover the nature of artists' interpersonal ties and provide a more powerful tool for analysis. We can only assume that jazz artists who cite other jazz artists in their interviews have some kind of association with them, but this relationship could be anything from close friendship and collaboration to just knowing the other person exists.

Identifying and representing the varied and nuanced semantics of social relationships pose significant computational challenges. To perform a deeper analysis of the social network, we complemented the machine-driven approach with a human-driven one that employs crowdsourcing techniques to assist with the interpretation of the interpersonal connections. Crowdsourcing has become increasingly popular as a means to harness human

intelligence to complete small, but crucial tasks within a large-scale project.

Linked Jazz 52nd Street [5] was developed to harness the knowledge of jazz scholars from academic centers for jazz studies as well as jazz enthusiasts from dedicated online forums to assist with the interpretation of the relationships among jazz artists as documented in archived interviews. This tool is a web-based application that asks contributors to classify the relationship between two jazz artists according to a menu of options. This assessment is facilitated by presenting the contributor with interview excerpts referencing the individuals in question. Results are tallied and converted into RDF statements that feed the project's LOD dataset. As part of the development of the tool, a list of social relationships was compiled by selecting suitable predicates from LOD vocabularies including the Relationship vocabulary [6], FOAF[7], and the Music Ontology [8]. The spectrum of relationships ranges from lower degrees of personal closeness (e.g., knows_of, has_met) to deeper levels of social ties (e.g., collaborated_with, influenced_by, mentor_of). This selection was the result of the analysis and mapping of person-centered RDF vocabularies (Pattuelli 2011). Jazz experts were also tapped to help discern which of these relationships would be of most interest to them and the larger community of potential users of jazz archives.

Linked Jazz 52nd Street Design

Crowdsourcing, a term coined by Jeff Howe (2006) in *Wired* magazine, was predominantly seen as an online business model, but recently successful projects, including New York Public Library's *What's on the Menu* [9] and University College London's *Transcribe Bentham* [10], have brought attention to crowdsourcing in the domain of cultural heritage as a method of supporting an array of labor-intensive and error-prone tasks including transcribing, classifying, proofreading, tagging, and annotating digital content.

The design of the Linked Jazz 52nd Street application was informed by research on crowdsourcing, as shown in Figure 2. The overall design of the site is geared towards lowering the barrier for participation through a simple and clean layout (Oomen and Aroyo 2011). Several studies have revealed that while extrinsic factors, such as monetary compensation and recognition, are strong motivators to engage in crowdsourcing projects, so are intrinsic factors, such as the opportunity to contribute to the greater good and learn new skills (Brabham 2010). This suggests that acknowledging user contributions through methods such as feedback and ranking of contributors, as well as providing tutorials and interaction with staff, helps to keep users engaged (Causer, Tonra and Wallace 2012; Huberman, Romero and Wu 2009). To this end, we encourage visitors to begin contributing by having a strong call to action message on the homepage asking visitors to click on a musician's photograph. As the contributor processes a transcript, an ego network visualization is built in real time while a progress bar fills indicating their progress (Fig. 2.1). This extrinsic motivator provides visual feedback indicating the immediate results of their contribution and makes their work transparent and accessible (Holley 2010). We also provide the ability for the contributor to put the task in context by expanding the transcript dialog (Fig. 2.2) to read more of the interview. This intrinsic motivator allows the contributor to break out of their current task and read more of the transcript if they find a compelling story or want to learn more about what is being discussed. Holly (2010) highlighted the importance of offering the contributor options of where to focus their work. We facilitate this choice by providing the contributor with the ability to select which individuals they wish to review (Fig. 2.3). Another important aspect is maintaining the provenance of the source information (Oomen and Aroyo, 2011) which we accomplish by including a link to the contributing institution and metadata such as the interviewer's name (Fig. 2.4). As recommended by Causer, Tonra and Wallace (2012), we provide detailed assistance through a tutorial system and instant popup help tips (Fig. 2.5). We also track and display metrics of the contributors (Fig. 2.6) to acknowledge their work and create a sense of community (Huberman, Romero and Wu, 2009).

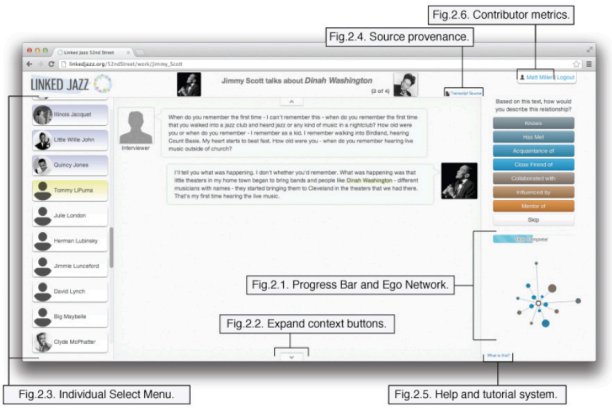


Fig. 2:
Linked Jazz 52nd Street design elements

Future plans for Linked Jazz 52nd Street include usability testing to be conducted both with jazz experts and non-experts. Feedback from these tests will inform

decisions regarding refinements and further development of the tool before its public release.

Conclusion

Scott (2011) points out that advances in social network analysis should not be simply descriptive work, but rather hold substantive significance. In addition to creating a deep network of the jazz community through the description of relationships at a more meaningful level, Linked Jazz 52nd Street will contribute a new LOD dataset representing jazz artists and their relationships that will be freely available for applications developers. Not only does it give new visibility to jazz archival resources, but it has the potential to promote new streams of research, including a socially-driven approach to the study of jazz history.

References

Brabham, D. C. (2010). Moving the crowd at Threadless. *Information, Community, & Society,* 13(8), 112-1145.

Causer, T., J. Tonra, and V. Wallace (2012). Trascription maximized; expense minimized? Crowdsourcing and editing *The Collected Works of Jeremy Bentham. Literary and Linguistic Computing.* 27.2 119-137.

Chen, I. X. and C. Z. Yang (2010). Visualization of social networks. *Handbook of Social Network Technologies and Applications*. New York: Springer. 585-610.

Davis, I. and E. Vitiello (2010). Relationship: A vocabulary for describing relationships between people. Available at: http://vocab.org/relationship/.html

Heckathorn, D. D., and J. Jeffri (2003). Social networks of jazz musicians. In *Changing the Beat: A Study of the Worklife of Jazz Musicians, Volume III: Respondent-Driven Sampling: Survey Results by the Research Center for Arts and Culture.* Washington, D.C.: National Endowment for the Arts Research Division Report #43. 48-61.

Holley, R. (2010). Crowdsourcing: How and Why Libraries Should Do It. *D-Lib Magazine,* 16.3/4). Retrieved from http://www.dlib.org/dlib/march10/holley/03holley.html

Huberman, B. A., D. M. Romero, and F. Wu (2009). Crowdsourcing, attention and productivity. *Journal of Information Science,* 35.6, 758-765.

Oomen, J., and L. Aroyo (2011). Crowdsourcing in the cultural heritage domain: opportunities and challenges. *Initiatives, 29*, 138-149.

Pattuelli, M. C. (2011). Mapping people-centered properties for Linked Open Data. *Knowledge Organization* 38.4, 352-359.

Pattuelli, M. C., C. Weller, and G. Szablya (2011). Linked Jazz: An exploratory pilot. In *DC-2011: Proceedings of the International Conference on Dublin Core and Metadata Applications* (pp. 158-164). The Hague, The Netherlands.

Schubert, A. (2012). Jazz discometrics: A network approach. *Journal of Informetrics* 6: 480-484.

Scott, J. (2011). Social network analysis: Developments, advances, and prospects. *Social Network Analysis and Mining* 1.1, 21-26.

Notes

1. linkedjazz.org
2. linkedjazz.org/network
3. http://eac.staatsbibliothek-berlin.de
4. http://socialarchive.iath.virginia.edu/
5. linkedjazz.org/52ndStreet
6. http://vocab.org/relationship/.html
7. http://www.foaf-project.org/
8. http://musicontology.com/
9. http://menus.nypl.org
10. http://www.ucl.ac.uk/transcribe-bentham/

ChartEx: a project to extract information from the content of medieval charters and create a virtual workbench for historians to work with this information

Petrie, Helen
helen.petrie@york.ac.uk
University of York, United Kingdom

Rees Jones, Sarah
Sarahsarah.reesjones@york.ac.uk
University of York, United Kingdom

Power, Christopher

christopher.power@york.ac.uk
University of York, United Kingdom

Evans, Roger
R.P.Evans@brighton.ac.uk
University of Brighton, United Kingdom

Cahill, Lynne
lynneca@sussex.ac.uk
University of Brighton, United Kingdom

Knobbe, Arno
knobbe@liacs.nl
University of Leiden, the Netherlands

Gervers, Michael
m.gervers@utoronto.ca
University of Toronto, Canada

Sutherland-Harris, Robin
r.sutherland.harris@utoronto.ca
University of Toronto, Canada

Kosto, Adam
ajkosto@columbia.edu
Columbia University, USA

Crump, Jon
jjcrump@uw.edu
University of Washington, USA

Summary

The ChartEx (**Chart**er **Ex**cavator) Project is developing and evaluating an innovative collection of computational methods to assist researchers in searching, extracting, analyzing, linking and understanding the content of medieval charters. The project is using both natural language processing and data mining techniques to establish entities such as locations (location in this context refers to a specific building or piece of land) and actors, events and dates related to those locations. The project is also developing a "virtual workbench" to support historians in working with large corpora of digital charters and the vast amounts of information that can now be extracted from them. These methods could subsequently be applied to other corpora of digitized texts, be they historical or contemporary.

The ChartEx Project

Researchers now have access to a deluge of data in the form of digitized historical records. One example is medieval charters or title deeds which record transfers of land ownership and are a major source for the study of people and places in the past, including the topography, economy and social relationships of pre-modern communities. However, current digital search aids are not sufficiently sensitive to the needs of researchers seeking to exploit the wealth of detail available within this type of record. The ChartEx (**Chart**er **Ex**cavator) Project is developing and evaluating an innovative collection of computational methods to assist researchers in searching, extracting, analyzing, linking and understanding the content of medieval charters. These methods could subsequently be applied to other corpora of digitized texts, be they historical or contemporary.

The ChartEx Project is extracting information from charters, using a combination of natural language processing (NLP) and data mining (DM) components to establish entities such as locations (location in this context refers to a specific building or piece of land) and actors, events and dates related to those locations. The NLP component uses rules derived from the knowledge of experienced researchers, in combination with the semantic meaning of the written language of the charters, to extract these entities. In order to inform these rules a new markup schema was defined for use in the project.

From a sample of manually marked up charters, the NLP component generates automatic markup of entities in a larger corpus of charters. The DM component then uses the output of the NLP component, as refined by researchers if needed, to extract relationships between entities.

However, the rules used by the NLP component and the relationships found by the DM component cannot reflect all the knowledge of experienced researchers. Therefore, the third crucial component of the ChartEx Project is the use of novel instrumental interaction techniques which will allow researchers to both refine the processing of the NLP and DM, and to directly manipulate (visualise, confirm, correct, refine, augment) relationships extracted from the charters to gain new insights of interest about the entities within them. Instrumental interaction emphasizes that both the computational system and human users have knowledge and must work together to achieve the users' goals. ChartEx is developing a highly usable "virtual workbench" for researchers to support this instrumental interaction. The ChartEx Workbench has interactive visualizations of large amounts of data and a range of tools to manipulate data and their relationships.

The ChartEx project is demonstrating its approach by analysing five corpora of charters from the 10th to

16th centuries originating from both the UK and other European countries. Two corpora contain full Latin texts and three contain Latin texts that have been provided with English summaries within digital archival catalogues. The collections derive from The National Archives of the United Kingdom, the Borthwick Institute for Archives in York (United Kingdom), the deeds collected as part of the DEEDS project at the University of Toronto (Canada), and deeds originating from the archive at Cluny (France).

Charters are an abundant and fundamental source for the study of many aspects of medieval societies. While recent scholarship has expanded the range of charter studies to such fields as the history of emotions and performativity, their core usefulness remains their provision of basic data: personal names, place names, and dates. In particular, they help us to trace the ownership and occupation of houses and parcels of land over centuries, providing the basis for many further studies from history to tourism and conservation. Initially in the project we assumed that the formulaic nature of the language in medieval charters would make the NLP quite straightforward. However, the high level formulaic nature of expressions revealed considerable variability when subjected to fine-grained analysis. Nonetheless, by developing a detailed markup schema for the initial hand coding of a set of charters, it was possible to train the NLP component to identify personal names, place names and locations. The Latin (and after c.1300, vernacular) phrases that describe, for example, the location of a property (e.g., 'the tenement in Petergate lying between the tenement once held by John the apothecary and now held by Richard of Huntington on one side, and the church of St Michael on the other'), were the pre-cursors to street-numbers and scientific spatial referencing developed from the 19th century. When researching a particular historical lived environment, the researcher needs to establish links between actors, events, and locations, by recovering and reconstructing the relationships between hundreds, even thousands, of data points, included in different charters and even in different archives.

Most current systems for searching digital charters depend on manual markup systems which identify entities such as place-names and personal-names in the text and include such entities in catalogue metadata. One example of such a project would be 'Paradox of Medieval Scotland' which uses metadata in just this way and then provides full text transcriptions of the locational data within the charter. Another example is the WARD 2 data in The National Archives. ChartEx is using NLP to go beyond the metadata and to explore the actual content of charters, thus enabling researchers to explore the discursive descriptions of locations, actors and events within those locations.

The ChartEx Project has designed a detailed markup scheme that adequately represents the ways in which historians currently read charters and extract spatial means from them. This new markup schema was created through a collaborative process involving researchers from 3 different institutions. This schema includes a set of entities and relationships that are at a level of abstraction above traditional diplomatic markup and specifically identifies actors, locations and events in a way that can be used by the NLP component. This new markup schema is documented in a detailed set of guidelines that have been used annotate a set of 250 charters from across the five different data archival data sets. These manually marked up charters provide act as a set training data for the technical components of the project.

The ChartEx Project has also developed the ChartEx Virtual Workbench to allow historians to work with the large amounts of data becoming available to them. For this we haven taken a highly user-centric approach, using the latest methodologies from human-computer interaction. Eight historians participated in contextual interviews, in which they described in considerable detail how they work with charters on a range of different historic problems and recreated actual pieces of research they had undertaken with charters. These sessions were recorded and analysed in detail to understand the specific tasks that historians are undertaking with charters and how to support them most effectively in doing the same tasks in a purely digital environment, as well as provide them with new functionality that they have not had before. In particular, we also considered the additional information that historians will have available from the results of the NLP and DM components in the project. On the basis of all this information, a number of "low fidelity" prototypes of possible workbench configurations and functionalities have been developed and two co-design workshops have been held with the six historians and two archivist advisors to the to discuss the different possibilities. As a result of these workshops, a more detailed prototype has been developed and is being evaluated by the original set of historians and archivists and a number of newly recruited historians and archivists, who are testing it with realistic research tasks. This will undoubtedly lead to a further iteration of refinement of the Virtual Workbench before the final prototype is developed by the end of the ChartEx Project.

The ChartEx Project is funded under the Digging into Data challenge (www.diggingintodata.org). The research consortium includes, historians and linguists as well as experts in human computer interaction, natural language processing and data mining from six institutions in four countries: Universities of York and Brighton (United Kingdom), University of Toronto (Canada), University of Washington and Columbia University (USA) and University of Leiden (Netherlands).

Markup Beyond XML

Piez, Wendell
wapiez@wendellpiez.com
Piez Consulting Services, United States of America

Markup Beyond XML

A review of the limitations of XML is not necessary here, as they have been debated as long as XML, and SGML before it, have been applied to data in the humanities. (For example, see Barnard et al. 1988, Huitfeldt 1994, Barnard et al. 1995.) Nor is it necessarily helpful, inasmuch as dwelling on the difficulties (while ignoring the considerable strengths) of XML is not in itself a constructive activity.

However, even when conceptualizing a markup technology that does not face these particular limitations, we can also learn from XML's strengths. In particular, from the concept of generic and "descriptive" markup (Coombs 1987) we have learned that document markup need not be strongly bound to application semantics, but can be declarative and oriented to the information and problem domain (rather than the particular platform and toolset in use), providing many advantages both for modeling and for long-term application independence. Similarly, XML's fairly clean separation of **syntax** (the text-based format that constitutes XML formally — angle brackets delimiting markup and distinguishing it from text content) from **model** (most commonly, but not necessarily, the W3C DOM or the related "tree"-shaped model of a document described by the Xpath/XSLT/XQuery family of specifications) has enabled the development of powerful standards-based tools — for both creating and maintaining XML syntax and for processing it — and helped ensure the platform independence and portability of XML technologies.

LMNL, the Layered Markup and Annotation Language, borrows both of these fundamental concepts. LMNL was first described by Jeni Tennison and myself in 2002 (Tennison and Piez 2002). However, its model differs significantly from XML's in two important respects:

I. Where XML stipulates a complete organization of a data set into discrete containers (elements), LMNL simply identifies **ranges** over a text. Ranges, as sequences of characters (or more formally, of **atoms**) may overlap with other ranges. In LMNL syntax, this is legal:

```
[poem}Et [red}l'unique [gold}cordeau{red]
des [green}trompettes{gold] marines{green]
{poem]
```

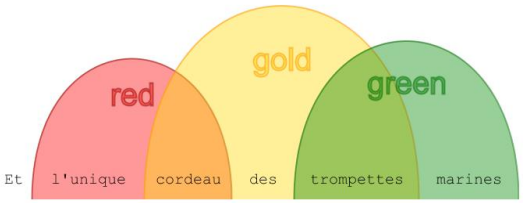


Figure 1.

As identified (in the syntax) by start-tag/end-tag pairs, any range may overlap with any other. Indeed, any relations between ranges — whether one may or does "contain" or "overlap" another — are not conditioned by the rules of LMNL, but provided, when needed, by an application in process. (So an application may say "colors must not overlap", or "red must not overlap with gold", but LMNL itself does not.) Thus LMNL markup is like a set of crayons or markers (for marking a text) as opposed to a pair of scissors (for segmenting it).

II. LMNL ranges may be annotated, in a way similar to attributes assigned to elements in XML, but also with significant differences. While in XML, attributes are provided as name-value pairs, any LMNL **annotation** may have structure: `[poem [author}Guillaume [surn}Apollinaire{surn]{author]}Et l'unique cordeau des trompettes marines{poem]` Here, the author annotation has content that is marked up: `Guilliaume [surn}Apollinaire{surn].` Moreover, in LMNL, annotations may have annotations (as if XML attributes could have attributes), and indeed annotations may encapsulate, comprehend or subsume entire documents (perhaps stored out of line, elsewhere on the network).

Draft specifications for the LMNL model and proposed syntax are at http://www.lmnl-markup.org .

The examples here use markup syntax, but the LMNL model by definition (and in this respect like the W3C XDM) requires no particular serialization format or no serialization at all. The LMNL model is defined in such a way that it can be represented using any capable syntax, or modeled directly in a database or object structure. XML may also be used, and indeed any of the documented XML-based approaches to representing overlap (including so-called milestone elements, or via segmentation and alignment, or using standoff; see TEI P5, chapter 20) may be mapped, usually straightforwardly, into the LMNL model. A simple transformation can rewrite any XML document as a LMNL document; by supplementing this with specialized logic for recognizing any XML-based conventions for representing overlap and expressing them directly as LMNL ranges, any XML that currently represents overlap can be rewritten automatically into LMNL and processed in a LMNL processor. (It is also possible to go back the other way.)

The LMNL model itself is fairly simple. A LMNL instance is defined as a **text layer** (a sequence of characters) with a set of **ranges** over the text. A text layer is a sequence of atoms: an **atom** can be represented using markup, but more commonly Unicode characters (each of which indicates a distince atom) are used. As noted, ranges have **annotations**. Annotations can be anything, from nothing at all (both ranges and annotations can be anonymous in LMNL) to just a name, to entire documents: annotations have their own text layers, so they can marked up. Annotations, like ranges, may be annotated. And their annotations may be marked up (its content marked with ranges), its ranges annotated, etc. While annotations belong to ranges (or annotations), and ranges belong to text layers (in the documents or annotations over which they range), ranges have no necessary relations to one another except as specified in the application. This means that ranges may overlap even other ranges assigned the same (type) name — so single defined sets of tags may be used to indicate ranges in the text for indexing or annotation, even if these ranges sometimes (or often) overlap. Ranges may also be filtered or associated in ways that represent multiple concurrent hierarchies (not single hierarchies alone, one at a time).

Luminescent, in combination with XSLT stylesheets developed to transform its (XML-based) output, is currently capable of all the following:

1. Automatically check for LMNL syntax well-formedness, outside an application.
2. Extract XML dynamically from documents marked with LMNL, given a list of elements to represent as (in) the XML tree.
3. Analyze the content for overlapping ranges.
4. Generate formatted output.
5. Generate alternative renditions and visualizations (e.g. SVG), showing structures of relations (and/or the lack thereof) in the marked up text.
6. Filter and transform.

A few demonstrations of LMNL syntax with outputs from the Luminescent toolchain (implemented as a sequence of XSLT transformations in a pipeline as described in Piez 2012) are offered to supplement this presentation. See: http://www.piez.org/wendell/papers/dh2013/lmnl/index.html

These experiments demonstrate the potential of a model supporting overlap for the study of narrative structure (in which narrative and dialogic structures commonly overlap with the native structures of verse, prose or drama) and for prosody (in which verse structure and sentence structure overlap in interesting ways). In addition, many other uses for LMNL can be readily envisioned, whether for supporting indexing, arbitrary annotation, data retrieval and filtering across arbitrary semantic boundaries, or others.

Finally, it must be noted that in a model that supports both overlapping structures and structured annotations, there are expressive opportunities for the representation of phenomena in literary texts (and any complex text) that are unavailable in XML. Consequently, it is suggestive of other models of text altogether — which can be more easily optimized, in many cases, for certain kinds of processing that are difficult, at best, in XML.

Source code for Luminescent is maintained here on github: https://github.com/wendellpiez/Luminescent

At DH2013 I will be demonstrating the pipeline; describing its design, operations, methods and capabilities; remarking on issues and work remaining to be done; and answering questions.

Some LMNL source code

[octave}[quatrain}[line}[s}[phr}She is as in a field a silken tent{line] [line}At midday when the sunny summer breeze{line] [line}Has dried the dew and all its ropes relent,{phr]{line] [line}[phr}So that in guys it gently sways at ease,{phr]{line]{quatrain] [quatrain}[line}[phr}And its supporting central cedar pole,{phr]{line] [line}[phr}That is its pinnacle to heavenward{line] [line}And signifies the sureness of the soul,{phr]{line] [line}[phr}Seems to owe naught to any single cord,{phr]{line]{quatrain]{octave] [sestet}[quatrain}[line}[phr}But strictly held by none,{phr] [phr}is loosely bound{line] [line}By countless silken ties of love and thought{line] [line}To every thing on earth the compass round,{phr]{line] [line}[phr}And only by one's going slightly taut{line]{quatrain] [couplet}[line}In the capriciousness of summer air{line] [line}Is of the slightest bondage made aware.{phr]{s]{line]{couplet]{sestet]

In an application (in which this document is parsed and transformed into an SVG representation), a graphical view shows how cleanly nested the sentence/phrase structure is with the verse structure of this sonnet. The only case of overlap is at the end of line 9 (at the start of the sestet), where it enjambs with line 10 after the only mid-line caesura in the poem:

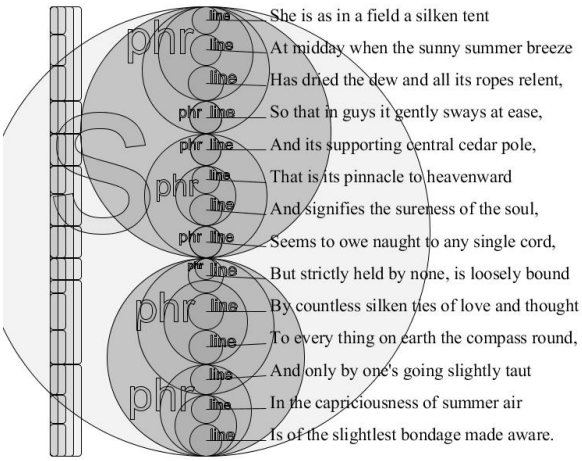


Figure 2:

References

Barnard, D., R. Hayter, M. Karababa, G. Logan, and J. McFadden (1988). SGML-Based Markup for Literary Texts: Two Problems and Some Solutions. *Computers and the Humanities*. 22(4): 265-276.

Barnard, D., L. Burnard, J. P. Gaspart, L. A. Price, C. M. Sperberg-McQueen, and G. B. Varile (1995). Hierarchical Encoding of Text: Technical Problems and SGML Solutions. The Text Encoding Initiative: Background and Context. *Computers and the Humanities*. 29(3). *The Text Encoding Initiative: Background and Context*. 211-231.

CATMA: Computer Aided Textual Markup and Analysis. http://www.catma.de/ .

Coombs, J. H., A. H. Renear, and S. J. DeRose (1987). Markup Systems and The Future of Scholarly Text Processing. *Communications of the ACM*. 30(11): 933-947

Huitfeldt, C. (1994). Multi-Dimensional Texts in a One-Dimensional Medium. *Computers and the Humanities*, 28(4-5). *Humanities Computing in Norway*. 235-241.

Tennison, J., and W. Piez (2002). The Layered Markup and Annotation Language (LMNL). In *Extreme Markup Languages 2002*.

Piez, W. (2004). Half-steps toward LMNL. In *Proceedings of Extreme Markup Languages*. http://conferences.idealliance.org/extreme/html/2004/Piez01/EML2004Piez01.html .

Piez, W. (2008). *LMNL in Miniature: An introduction. Amsterdam Goddag Workshop* held December 2008 http://piez.org/wendell/LMNL/Amsterdam2008/presentation-slides.html .

Piez, W. (2010). *Towards Hermeneutic Markup: an Architectural Outline. Digital Humanities 2010* held July 2010 at King's College. London http://piez.org/wendell/dh2010/index.html .

Piez, W. (2012). "Luminescent: parsing LMNL by XSLT upconversion." Presented at *Balisage: The Markup Conference 2012 (Montréal, Canada), August 2012. In Proceedings of Balisage: The Markup Conference 2012. Balisage Series on Markup Technologies*, 8. doi:10.4242/BalisageVol8.Piez01.

Schmidt, D. (2010). The inadequacy of embedded markup for cultural heritage texts. *Literary and Linguistic Computing* 25(3): 337-356. doi: 10.1093/llc/fqq007.

Sperberg-McQueen, C. M. (1991). Text in the Electronic Age: Textual Study and Text Encoding, with Examples from Medieval Texts. *Literary and Linguistic Computing* 6(1).

Text Encoding Initiative (TEI). *P5: Guidelines for Electronic Text Encoding and Interchange*, chapter 20, "Non-hierarchical Structures". http://www.tei-c.org/release/doc/tei-p5-doc/en/html/NH.html .

Stührenberg, M., and D. Goecke (2008). SGF — An integrated model for multiple annotations and its application in a linguistic domain. Presented at *Balisage: The Markup Conference 2008 (Montréal, Canada), August 2008. In Proceedings of Balisage: The Markup Conference 2008. Balisage Series on Markup Technologies*, 1. doi: 10.4242/BalisageVol1.Stuehrenberg01.

Stührenberg, M., and D. Jettka (2009). A toolkit for multi-dimensional markup — The development of SGF to XStandoff. Presented at Balisage: The Markup Conference 2009 (Montréal, Canada), August 2009. In *Proceedings of Balisage: The Markup Conference 2009. Balisage Series on Markup Technologies*, 3. doi: 10.4242/BalisageVol3.Stuhrenberg01.

MESA and ARC, developing disciplinary metadata requirements in a multidisciplinary context

Porter, Dot

dot.porter@gmail.com
Indiana University, United States of America

The widespread adoption of digital media and technologies in almost every facet of humanities research and scholarly communications, including discovery and repurposing of information, writing, publication, peer review, curation, and dissemination, has brought with it both

great opportunities and significant challenges. Medievalists were among the first humanities scholars to adapt digital tools for their work; indeed, Father Roberto Busa's *Index Thomasticus*, a digital index begun in 1949 that comprises more than 11 million words of medieval Latin found in the works of St. Thomas Aquinas, is widely acknowledged as the first digital humanities project. Since that time medievalists have a strong record of building a wide array of digital repositories, using electronic tools and textual encoding to advance new methods for editing texts and constructing scholarly editions, and building tools for the analysis of textual and visual data (Unsworth 2011).

Despite such progress, the community still faces significant needs in the area of discoverability of digital scholarship. Many scholarly initiatives that take advantage of the possibilities of digital media are not indexed or discoverable via traditional library and information sciences resources, or even via mainstream search engines such as Google. In recent years two federations of electronic scholarly projects, the Networked Infrastructure for Nineteenth-Century Electronic Scholarship (NINES) and 18thConnect, have made significant progress towards addressing this problem in the fields of 18th- and 19th-cenutry Anglo-American literary studies. Together they provide aggregated searching of more than 1.6 million digital objects from 128 member projects, and users of the web portals provided by each federation may search the content of both federations individually or simultaneously.

In 2010, Dot Porter and Timothy Stinson were awarded a planning grant by the Andrew W. Mellon Foundation to explore the feasibility of and need for establishing a similar federation in the field of medieval studies. In May 2011, we organized a planning meeting held in Baltimore, MD in order to take the first steps towards a dialogue with this community. In order to learn from our peers in other disciplinary areas, as well as to explore potential avenues of collaboration with them, we also invited representatives from NINES, and 18thConnect. The outcomes of this meeting included the formation of a steering committee to direct its creation.[1]

After evaluating possibilities for achieving aggregation, including adapting known software or building our own, there was unanimous support from the steering committee for creating an implementation of MESA utilizing Collex, the open-source collections builder that facilitates collecting, tagging, analyzing, and annotating digital sources and provides the framework upon which both NINES and 18thConnect are built (Nowviskie 2007). Collex already has a proven track record in meeting the goals identified by MESA, as evidenced by both of these extant federations. Furthermore, adoption of Collex is efficient because it avoids duplication of tools and facilitates not only aggregation within our field, but also aggregation between fields and between extant aggregators. This has already been demonstrated by NINES and 18thConnect, which are currently linked and may be searched either separately or in tandem. In addition, Laura Mandell, director of 18thConnect, is currently organizing the Advanced Research Consortium (ARC), a meta-federation that will span disciplinary areas from the medieval era through the twentieth century, and MESA has joined NINES and 18thConnect as a member of ARC. More recently, REKn (focused on Renaissance studies) and ModNets (focused on modernist studies) have also joined ARC (Robideau 2011). Thus, all our decisions about how to develop MESA must be informed by, and take into account, the needs of our partner "nodes" in ARC.

The Mellon Foundation awarded MESA an implementation grant in June 2012. We knew, even as we wrote the implementation proposal, that a number of adaptations to Collex and the underlying metadata schemas would be necessary due to the fact that, unlike NINES and 18thConnect, projects in MESA will have medieval and modern languages, as well as a variety of objects including manuscripts, maps, architectural drawings and images, and statuary. In the months since funding was awarded the MESA team has been focusing primarily on the specifics of how to aggregate the first twelve projects into the MESA Collex instance.[2] Aggregation of metadata from member projects is achieved by generating Resource Description Framework (RDF) files from extant metadata (e.g., transcriptions, catalogue records, or descriptions of images) and making these RDF files discoverable and cross-searchable within Collex.

Modifications to MESA RDF have been determined primarily by the MESA co-directors, with significant input from the Steering Committee. Over several meetings and through email correspondence, the Steering Committee developed a list of what they saw as the most important, most basic, metadata requirements for a medieval digital federation to include. We were aware of the metadata requirements already defined for NINES and 18thConnect (detailed on the Collex wiki: http://wiki.collex.org/index.php/Submitting_RDF), which include Dublin Core Metadata Terms, RDF terms, and custom Collex terms. However, rather than use the existing Collex metadata as our starting point, we first brainstormed (from our own extensive experience as scholars and project developers) by what fields and terms users of MESA would expect to be able to search in the federation. We also had to take into account the practicalities of the first twelve projects, and future projects to be accepted into the federation (what metadata these projects have, how is it formatted, and would it be practical for us to request additional metadata of them before they are aggregated into MESA).

In addition to Title, Author, and Date of Creation, we discussed the need to be able to include descriptions of people responsible for creating objects (scribe, artist, etc.); the format of the objects (for example manuscript codex, printed sheet, sculpture, painting, stained glass window, etc.); the provenance of the objects (including where they were created, and where they are now); languages inscribed in and on objects, as well as cultures reflected in objects.

In some cases, NINES and 18thConnect did include the fields we needed, notably a flexible field for Author, <role:***>, where *** can equal one of several roles (AUT for author, ART for Visual Artist, EDT for Editor, etc.) and a very flexible system for noting dates. We added <dc:type> for describing the general format of objects (defined as "The nature or genre of the resource.").[3] Although the formal definition of <dc:format> seems better ("The file format, physical medium, or dimensions of the resource."),[4] we discovered that in practice other projects that we are federating or hope to federate (including e-codices and Europeana Regia) use <dc:type> for describing the general format of objects, and <dc:format> for more specific information (including medium and measurements), so we decided to follow their practice. We do plan to include only general format in MESA fields, and not detailed information on medium or measurements of objects.

As neither NINES nor 18thConnect have needed a field for describing provenance, we have added <dc:provenance> as a recommended field in the MESA RDF specifications.[5] At one point we discussed requiring projects to further specify data within <dc:provenance> — for example, one <dc:provenance> for each change in status to the object, and each <dc:provenance> must have one <dc:date>, one <dc:name>, and one <dc:event> defined within it. However it became clear, once we started mapping actual project metadata, that the more highly specified metadata was simply not going to be realistic.

As of November 2012, we are still finalizing our methods for specifying languages and cultures of objects, but we hope to have those finalized soon.

Although NINES and 18thConnect had well-developed RDF specifications and workflows, we knew coming into the project that MESA would need to make changes, some slight and some more radical, in order to ensure that the federation would be usable and useful for an audience of scholars and students in medieval studies. This abstract has detailed some of those modifications, and a full paper at the conference would detail all the changes, the argumentation for including the fields we include, and discussion of how the changes required by MESA have influenced the development of the metadata specifications for the whole of ARC. Indeed, while we knew that MESA would require changes to the existing specifications, what we did not expect was that so many of our requested changes would be welcome, and adopted by, the larger group of ARC nodes. Although our relationship with ARC is ongoing MESA has already made substantial contributions and we look forward to more years of fruitful collaboration.

References

Dublin Core Metadata Terms http://dublincore.org/documents/dcmi-terms/

Eighteenth Century Scholarship Online www.18thconnect.org

Medieval Electronic Scholarly Alliance Blog http://www.dlib.indiana.edu/projects/mesa/

Nineteenth Century Scholarship Online www.nines.org

Nowviskie, B. (2007). A Scholar's Guide to Research, Collaboration, and Publication in NINES, Romanticism and Victorianism on the Net. 47. http://www.erudit.org/revue/ravon/2007/v/n47/016707ar.html

Resource Description Framework http://www.w3.org/RDF/

Robideau, R. Texas A&M's College of Liberal Arts to house digital literary research consortium, Texas A&M College of Liberal Arts (press release) http://liberalarts.tamu.edu/html/news-texas-a-m-s-college-of-liberal-arts-to-house-digital-literary-research-consortiu.html

Unsworth, J. (2011). Medievalists as Early Adopters of Information Technology. *Digital Medievalist Journal* 7 http://www.digitalmedievalist.org/journal/7/unsworth/

Notes

1. Members of the Steering Committee: Dot Porter, co-chair (Indiana University), Timothy Stinson, co-chair (North Carolina State University), James Cummings (University of Oxford), Christoph Flüeler (Institut d'études médiévales, University of Fribourg and e-codices), Will Noel (University of Pennsylvania), Dan O'Donnell (University of Lethbridge), Lynn Ransom (University of Pennsylvania), Peter Robinson (University of Saskatchewan), Torsten Schaßan (Herzog August Bibliothek Wolfenbüttel), and Stephen Shepherd (Loyola Marymount University).
2. The projects to be aggregated into MESA during the first year of the project are: Digital Image Archive of Medieval Music (DIAMM), e-codices: Virtual Manuscript Library of Switzerland, Gothic Ivories Project, InteLex, Online Froissart, Parker Library on the Web, Petrus Plaoul, Roman de la Rose Digital Library, St. Gall Monastery Plan, sermones.net, University of Pennsylvania Libraries Penn in Hand, Walters Art Museum
3. http://dublincore.org/documents/dcmi-terms/#terms-type

4. http://dublincore.org/documents/dcmi-terms/#terms-format
5. http://dublincore.org/documents/dcmi-terms/#terms-provenance

Building the Social Scholarly Edition: Results and Findings from A Social Edition of the Devonshire Manuscript

Powell, Daniel James
djpowell@uvic.ca
Electronic Textual Cultures Lab, University of Victoria, Canada

Crompton, Constance
ccrompto@uvic.ca
Electronic Textual Cultures Lab, University of British Columbia-Okanagan, Canada

Siemens, Ray
siemens@uvic.ca
Electronic Textual Cultures Lab, University of Victoria, Canada

1. Introduction and Background

a. History and Context

Social media technologies can extend and enhance scholarly conversation while challenging traditional notions of textual authority and peer review. Twitter facilitates resource and idea sharing with a speed and ease formerly only possible at conferences; Facebook allows the formation of communities of interest founded not on geography but affinity; blogs disseminate research for widespread discussion; and, most significantly, Wikipedia has become the most popular and largest single reference resource in history, with more than 14 million articles in over 250 languages produced by 1 million monthly contributors (Wikimedia Report Card, 2012). This long paper reflects on the construction of a *social scholarly edition* of the Devonshire Manuscript that attempted to harness emerging social media environments to produce a new type of scholarly edition, one that allows multiple stakeholders to access, contribute, and discuss its construction.[1]

In this paper we recount the incipient formation of a new type of editing community, one that we argue is defined by iterative publication of material, multiple communities of interest contributing to a single project, the use of technology to facilitate these contributions, and the growing importance of self-directed learning to scholarly editing. Our successes and, just as importantly, our moments of failure, offer insight into best practices for a type of "facilitative scholarship" that will likely become increasingly common as comfort with social media technologies grows within the academy. As outlined in a DH2012 poster session (Crompton and Siemens, 2012), we designed the public editing process for the social edition from the start to encourage communication across editorial communities while preserving the peer review process. These communities included the Electronic Textual Cultures Lab team, the project advisory board, the online Iter Community (http://www.itergateway.org/), early modern critics and scholars operating in the blogosphere, Wikibook and Wikipedia users, Tudor enthusiasts, and the general public. 2

b. Materials of the Project

The Devonshire Manuscript (British Library Additional MS 17,492) contains approximately 200 items (Southall, 1964: 143, Remley, 1994: 47), including poems, verse fragments, excerpts from longer works, anagrams, jottings, and doodles by a coterie of men and women centered on the court of Queen Anne Boleyn. Inscribed in over a dozen hands, the manuscript has long been valued as a source of Sir Thomas Wyatt's poetry. In addition to 129 of his poems, the volume contains other transcribed lyrics and original work by numerous court figures, including Mary Shelton, Lady Margaret Douglas, Mary (Howard) Fitzroy, and Lord Thomas Howard (Southall, 1964: 143). These multiple contributors often comment and evaluate each other's work through marginal notation and in-line interjection. In addition to a consideration of the volume as "a medium of social intercourse" (Love and Marotti, 2002: 63), the multi-layered and multi-authored composition of the Devonshire Manuscript make it an ideal text for experimentation in social editing.

The *Social Edition of The Devonshire Manuscript* project manifests Ray Siemens' earlier argument that social media environments might enable new editing practices (Siemens et al., 2012a). In building an edition of an early modern text on the principles of open access and editorial

transparency in both production and dissemination, we have integrated scholarly content into environments maintained by the social-editorial communities that have sprung up on the web; most notably, these include the Wikimedia suite of projects (Wikipedia, Wikibooks, Wikisource). We have run an experiment to see how one might build an edition which is scholarly in a traditional sense, but which extends the editorial conversation into multiple pre-existing social media platforms including blogs, wiki discussion pages, dedicated Renaissance and early modern online community spaces, Skype-enabled interviews with our advisory group, and Twitter.

2. The Complexity of the New Scholarly Editing Community

a. Iterative Publication

Perhaps more than any other editorial choice, the iterative publication of the social edition of the Devonshire Manuscript departed most clearly from traditional scholarly editing practices. We have, in effect published (or are in the perpetual process of publishing) two versions of the edition in two mediums: a fixed PDF version, distributed to the project's advisory board, and a version housed on the publicly-editable Wikibooks. We are currently working with multiple publishing partners to produce a second online edition, an e-reader edition, and a print edition to meet the needs of a broad and varied readership. These versions were planned to productively inform and influence each other's development, with cross-pollination of editorial input across platforms. Although they did so, each medium also engendered difficulties in communication, coordination, and expectations to be overcome or accommodated—with varying results.

b. Communities of Interest and Technologies of Communication

As outlined above, a central aim of the project was to facilitate knowledge transfer and creation between multiple editorial communities, all of whom were invested differently in the project. These ranged from individual academics giving feedback as advisors to interested members of the public in contact with project staff via Twitter. These groups adopted, considered, and, at times, rejected different types of communication technologies in fascinating ways. Wikibook discussion pages were considered by established academics to be spaces meant for peer review; wiki editors explained that they were in fact where confrontations over edits usually occurred. Wiki editors were very helpful with questions of coding and technical production of content, while other communities felt deeply uncomfortable editing posted content. Sustained discussions in the Iter Community space proved difficult, while members of the public interested in Tudor culture followed our work avidly and often interacted with us on Twitter. Bloggers focused on the early modern period helped to generate discussion and disseminate reports as our edition building progress, but chose to limit their direct involvement with producing the edition. In often surprising ways, the technologies of communication each group used came to define, in some cases, the communities of interest and their respective investments. Considered as a whole, our project suggests that social media technologies can be harnessed for productive interaction and discussion by those scholars invested in a content area or project, but that they require comprehensive oversight by dedicated staff to develop and maintain participation in knowledge construction and dissemination.

c. Self-Directed Learning

Wikimedia content is openly editable by any individual. Project staff quickly reconsidered this theoretically nonexistent barrier to entry, though, when coding of the edition began in Wikibooks. Resembling a cross between HTML, XML, and CSS, Wikitext language is idiosyncratic and required a great deal of time and experimentation on the part of project staff to use effectively. Given the central importance of lab staff to the production of this edition, we have realized that this ad-hoc program of self-directed study produced a new community: young scholars, mostly masters level and younger doctoral students, who have shown interest in digital scholarly production. In other words, those usually construed as "assisting" in large projects here took on increasingly centralized roles in coordinating community input, coding the social edition in Wikibooks, discussing the project with various communities, and writing and disseminating critical research on the project as a whole.

3. Conclusion

a. The Open Source Edition?

The basic structures of the social edition are completely open for manipulation and repurposing. The formation, maintenance, and oversight of multiple communities, however, is central the success of any such open edition. Community investment provides a foundation for a technologically facilitated, process-driven approach. As

our full paper will discuss in more detail, developing such communities is often difficult, with success depending on intensive and regular engagement and oversight. It is difficult for disparate communities, even when facilitated by social media technologies, to effectively come together for intellectual production. As even the well-regarded *Transcribe Bentham* project has widely discussed, crowdsourcing textual transcription—much less scholarly editing and production—is fraught with difficulties we are only beginning to navigate (Cause et al., 2012a; Causer et al., 2012b). In this reconfigured landscape of scholarly production, where we are likely "witnessing the nascent stages of a new 'social' edition existing at the intersection of social media and digital editing" (Siemens et al., 2012a: 446), however, we are not without models: the open source community, especially those groups devoted to general tool building and knowledge construction (OpenOffice, Wikimedia, Linux, Mozilla) is a powerful articulation of possible ways the technologically facilitated social production of intellectual content may fruitfully develop—given a robust and vibrant community of interest.

b. Ways Forward

The past two years of work suggests that some blend of intensive oversight and engagement with defined communities, along with a receptivity to spontaneously formed communities of affinity—as supported by both the *Transcribe Bentham* project (Causer, 2012b) and our own observations—is necessary to effectively implement social scholarly production. Only by becoming effective promoters, facilitators, and instigators can digital humanists provide an effective locus around which multiple communities can cohere. Although we encountered certain difficulties in facilitating knowledge exchange among various communities, on the whole we learned how to effectively facilitate community interaction across and between mediums and communities to produce scholarly knowledge in new ways.

References

Causer, T., J. Tonra, and V. Wallace (2012a). Transcription maximized; expense minimized? Crowdsourcing and editing *The Collected Works of Jeremy Bentham*. *Literary and Linguistic Computing* 27(2). 119-137.

Causer, T., and V. Wallace (2012b). Building a Volunteer Community: Results and Findings from Transcribe Bentham. *Digital Humanities Quarterly*. 6(1). http://www.digitalhumanities.org/dhq/vol/6/2/000125/000125.html .

Crompton, C., R. Siemens, and the ETCL and INKE Research Groups. (2013) 'Vertues Noble & Excelent'? Digital Collaboration and the Social Edition. *Digital Humanities Quarterly*. In consideration. [internally circulated for comment and revision]

Crompton, C., and R. Siemens (2012). The Social Edition: Scholarly Editing Across Communities. In *DH2012*. 16-22 July 2012. Abstract available at http://www.dh2012.uni-hamburg.de/conference/programme/abstracts/the-social-edition-scholarly-editing-across-communities/ .

Love, H., and A. F. Marotti (2002). Manuscript Transmission and Circulation. In Loewenstein, D., and J. Mueller. (eds).*The Cambridge History of Early Modern English Literature*. Cambridge: Cambridge University Press. 55-80.

Remley, P. G. (1994). Mary Shelton and Her Tudor Literary Milieu. *Rethinking the Henrician Era: Essays on Early Tudor Texts and Contexts*. In Herman, P. C. (ed). Urbana: University of Illinois Press. 40-77.

Siemens, R., C. Warwick, R. Cunningham, T. Dobson, A. Galey, S. Ruecker, S. Schreibman, and the INKE Team (2009). Codex Ultor: Toward a Conceptual and Theoretical Foundation for New Research on Books and Knowledge Environments. *Digital Studies/Le champ numérique* 1(2). http://www.digitalstudies.org/ojs/index.php/digital_studies/article/view/177/220 .

Siemens, R., M. Timney, C. Leitch, C. Koolen, A. Garnett, with the ETCL, INKE, and PKP Research Groups (2012a). Toward modeling the *social* edition: An approach to understanding the electronic scholarly edition in the context of new and emerging social media. *Literary and Linguistic Computing.* 27(4). 445-461.

Siemens, R., M. Timney, C. Leitch, C. Koolen, and A. Garnett (2012b). Understanding the Electronic Scholarly Edition in the Context of New and Emerging Social Media: Selected, Annotated Bibliographies. *Digital Humanities Quarterly*, 6(11). http://www.digitalhumanities.org/dhq/vol/6/1/000111/000111.html .

Southall, R. (1964). The Devonshire Manuscript Collection of Early Tudor Poetry, 1532–41. *RES* 15, 142-50. WikiMedia Report Card. (2012). http://reportcard.wmflabs.org/.

Notes

1. For an overview of pertinent critical contexts surrounding the modeling of the social edition, see Siemens et al., "Understanding the Electronic Scholarly Edition in the Context of New and Emerging Social Media."

2. These efforts are in keeping with the aims of the Implementing New Knowledge Environment (INKE) Project, a $2.5 million, 7-year Major Collaborative Research Initiative (MCRI) grant from the Social Sciences and Humanities Research Council (SSHRC) of Canada devoted to "exploring the future of the book from the perspective of its history." See the INKE website http://inke.ca/ and Siemens et al., "Codex Ultor: Toward a Conceptual and Theoretical Foundation for New Research on Books and Knowledge Environments."

Against the Binary of Gender: A Case for Considering the Many Dimensions of Gender in DH Teaching and Research

Radzikowska, Milena
mradzikowska@gmail.com
Mount Royal University, Canada

Sostar, Tiffany
possibly.t@gmail.com
University of Calgary, Canada

Ruecker, Stan
sruecker@id.iit.edu
Illinois Institute of Design

There are genders beyond masculinity and femininity. Genderqueers, bois, gurls and otherwise gender variant people explore and express a wide variety of non-binary gender identities. Joy Johnson and Robin Repta note that "gender is typically theorized as a multidimensional, context-specific factor that changes according to time and place", but in research practice, gender "is routinely assumed to be a homogeneous category … measured by a single check box" (28). The non-binary gender communities are untapped, underserved, and underrepresented by most scholarly activities. Digital humanities scholars have the potential to move beyond assumptions and to engage with non-binary gender in meaningful and tangible ways. This engagement can help the digital humanities live up to its promise, to "enhance the impact of [scholars'] work and engage with new audiences" (Prescott 63).

Though gender variance is not new, "exposure to new gender norms and social scripts has transformed the ways that some young individuals make sense of gender and gender non-conformity" (Shapiro 21). Shapiro cites technology as a tool that offers "new physical and social possibilities for gender" (24). The technology used in the field of digital humanities is part of this process of renegotiation and understanding, but it is necessary for digital humanities scholars to understand and accommodate these news ways of knowing and being. Mussell (writing about periodicals) notes that "there has been a fundamental transformation in the terms of access" (202) in the digital humanities classroom. This "fundamental transformation" applies to all aspects of knowledge compiled within the digital humanities, and it is perhaps the right moment now for the field to undertake the next step toward a comprehensive and nuanced engagement with new understandings of gender.

This paper is not only a call to catch up to complex understandings of non-binary gender in other disciplines such as queer and feminist theory (Butler), medical research (Johnson and Repta), child psychology (Wiseman and Davidson) and marketing and consumer research (Bettany, Dobscha, O'Malley and Prothero). It is a call to exceed the reach of these existing engagements. We must answer the same challenge Wiseman and Davidson have issued to clinical psychologists, "to continually reflect and discuss the multiplicity that is possible within human experience, recognizing the ways of thinking and knowing that we are embedded in" (536). Because the digital humanities influence how scholars do their work and how students frame their research questions, we are in a strong position to create a comprehensive discipline-wide conceptualization of non-binary gender. This is asking much of the digital humanities. The difficulty of engaging meaningfully with non-binary gender is highlighted by contradictions such as statements that "gender is to be considered as a two-sided coin, as constructions of masculinity, and what it is to be male, inevitably generate and constitute constructions of femininity, and what it is to be female" (Bettany et al 16) resulting from a conference entitled "Moving Beyond Binary Opposition: Exploring the Tapestry of Gender in Consumer Research and Marketing." What is demonstrated here is the embeddedness of the binary construction of gender that the conference sought to explore and move beyond. It is that very embeddedness that digital humanities should seek to confront and overcome. We have the power to shape views through our construction of digital humanities classrooms and research projects. We should try to answer the challenge offered by Ryka Aoki; "[w]e should provide [genderqueer, trans, and gender variant people] the chance and even some guidance to find their own answers…

let us give people the support and affirmation that they may never have experienced" (sec. 4).

Our paper proposes four ways that the digital humanities can engage scholars, research participants, and students in considering the non-binary nature of gender. First, we suggest that the design and development of tools for digital humanities scholarship should enable the exploration of texts according to gender identity, gender expression, biological sex, and sexual orientation. In addition, such exploration should be considered a multi-dimensional spectrum, not a series of on-off switches. Second, when constructing engagements with potential users of DH tools, we should employ design testing methods that move beyond assumptions of the gender binary. Third, in designing DH experiments, we might choose to privilege content from non-binary gender sources. Fourth, we can acknowledge non-binary gender in DH classes and seminars and, where possible, include relevant readings and explore the re-imagining of existing DH tools according to non-binary gender. Though these measures cannot provide a comprehensive response to the many issues that Digital Humanities scholarship must address regarding non-binary gender and gender inclusivity, they do represent important first steps towards an inclusive and collaborative solution.

References

Aoki, R. (2010). "On Living Well and Coming Free." In **Bornstein, K. and Bergman, S. B.** (eds). *Gender Outlaws*. Berkeley: Seal Press.

Bettany, S., S. Dobscha, L. O'Malley, and A. Prothero (2010). "Moving Beyond Binary Opposition: Exploring the Tapestry of Gender in Consumer Research and Marketing." *Marketing Theory* 10 3-28. Sage. (accessed 14 Oct. 2012).

Butler, J. (2004). *Undoing Gender*. New York: Routledge.

Johnson, J. L. and R. Repta (2012). "Sex and Gender: Beyond the Binaries." In **Oliffe, J. L. and Greaves, L.** (eds). *Designing and Conducting Gender, Sex, and Health Research*. New York: Sage Publications: 17-37. Sage. (accessed 14 Oct. 2012.)

Mussell, J. (2012). Teaching Nineteenth-Century Periodicals Using Digital Resources: Myths and Methods. *Victorian Periodicals Review* 45 201-209. Project Muse. (accessed 25 Oct. 2012).

Prescott, A. (2011). "Consumers, Creators or Commentators?: Problems of Audience and Mission in the Digital Humanities." *Arts and Humanities in Higher Education* 11 61-75. Sage. (accessed 24 Oct. 2012.)

Shapiro, E. (2010).*Gender Circuits: Bodies and Identities in a Technological Age*. New York: Routledge.

Wiseman, M. and S. Davidson (2011). Problems With Binary Gender Discourse: Using Context to Promote Flexibility and Connection in Gender Identity. *Clinical Child Psychology and Psychiatry* 17 528-537. Sage. (accessed 14 Oct. 2012).

Slave Biographies: Atlantic Database Network

Rehberger, Dean
rehberge@msu.edu
Michigan State University, United States of America

Hawthorne, Walter
walterh@msu.edu
Michigan State University, United States of America

Midlo Hall, Gwendolyn
ghall1929@gmail.com
Michigan State University, United States of America

LaChance, Paul
lachance1943@rogers.com
University of Ottawa, Canada

Foley, Catherine
Catherine.Foley@matrix.msu.edu
Michigan State University, United States of America

1) Overview

The NEH supported *Slave Biographies: Atlantic Database Network* project provides a platform for researchers of African slaves in the Atlantic World to upload, analyze, visualize, and utilize data they have collected, and to link it to other datasets, which together complement each other in such a way as to create a much richer resource than the individual datasets alone. During the past two decades, there has been a seismic change in perception about what we can know about African slaves and their descendants throughout the Atlantic World (Africa, Europe, North and South America). Scholars have realized that, far from being either non-existent or extremely scarce, various types of documentation about African slaves and their descendants abound in archives,

courthouses, churches, government offices, museums, ports, and private collections spread throughout the Atlantic World. Since the 1980s, a number of major databases were constructed in original digital format and used in major publications by their creators, but they lacked a platform for preservation and therefore are at risk of being lost as their creators retire. Also, a number of collections of original manuscript documents are beginning to be digitized and made accessible online free of charge. However, our task as historians is more than to preserve images of primary sources; it is to interpret those sources by finding new ways to organize, share, mine and analyze as well as to preserve original materials which might otherwise be discarded or lost.

The *Slave Biographies* digital repository has developed a core metadata schema describing enslaved peoples. The fields have been defined based on six datasets and a distinguished group of historians, serving on the project's Advisory Board. The digital repository schema, fields, and search interface will be made available in English and shortly thereafter in Spanish, Portuguese, and French.

Slave Biographies addresses two major challenges that historians increasingly face: first, to create models for collaborative research in a field that has been dominated by a methodology of—and rewards for—individual research and, second, to analyze vast quantities of data that can now be accessed digitally. The project makes tools available to perform calculations and visualize the data to encourage and assist collaborative, international studies of these numerous but widely scattered collections of materials. The stories about lives of slaves as well as the analyses of slavery emerging from this network will be a unique resource for linguists, creolists, anthropologists, economic historians, sociologists, geographers, cartographers, creative writers, and genealogists searching for their African ancestors as well as for historians of slavery.

This DH2013 presentation on the *Slave Biographies* project will not only give a comprehensive overview of the project but will outline the challenges of dealing with large historical datasets. However, what is most exciting about the project is that by bringing together large sets of data, we cannot only do large calculations and visualizations about the historical practices of slavery, but we can drill down to find rich representations of individuals and families, giving life to the individual biographies of slaves and their relationships and contexts. That is, the promise of big data may not be to bury people in blizzards of numbers but to recover the lives of individuals.

2) Significance

Never before has it been more important to the humanities to try, as NEH Chairman Jim Leach has recently noted, “to manage a deluge of data and turn bits of information into useful knowledge.” This is particularly true of history. Historians, their students and the public at large are awash in the materials available in digital archives and databases, a flood of data enhanced by global scholarly networks and better access to archives and collections around the world. More than ever, it is crucial to find ways to preserve and manage large stores of quantitative and qualitative data and to make it accessible in ways that important research questions can be asked and well-formed answers derived. To be sure, the task of accumulating, organizing, and making sense of mountains of information from scattered corners of the globe cannot be handled by any one researcher. If the humanities are to advance in transformative ways in this age of globalization, humanists must find new ways to collaborate—to work together on large, international projects. If international collaboration is to occur, humanists in the wealthy countries with best access to new networking and data-analyzing technologies must find ways to make them available to their colleagues in poorer countries. *Slave Biographies*, to this end, provides both a collaborative, international project that is building a data network available to scholars, teachers, students, and genealogists in the U.S. and abroad, and a platform for addressing the difficult practical, ethical, methodological and, especially, hermeneutic problems scholars face when turning their attention to collecting and analyzing data about African slaves and their descendants.

Answering important historical questions

For several decades after the publication in 1969 of Philip D. Curtin's seminal book, *The Trans- Atlantic Slave Trade: A Census*, much scholarly attention and resources have been focused on quantitative studies of trans-Atlantic slave trade voyages. This research has yielded considerable important scholarship, discussed in the following section about related databases. However, these sources contain relatively little information about individual enslaved Africans.

Recently, a growing number of scholars have been unearthing important data from other sources, such as notarial documents; plantation inventories; police reports; testimony by runaway slaves, conspirators and rebels against slavery, church books of baptism, death and marriage, church Inquisition testimony, government and church censuses, which reveal much about slave life in the New World *and* about African slaves’ lives in parts of the Old World. These sources focus on individual slaves. When

records about many individuals are combined, patterns can be discerned. Data about ethnicities tell us from where within Africa many slaves hailed; data about slave residence in the Americas tell us where members of particular groups ended up and where and how they were housed; data about marriages tell us with whom Africans and their descendants chose to partner; data about skills tell us what slaves did and their contributions to agriculture, trade and the economy beyond brute labor. And this list could go on for pages.

Among the questions that might be asked and answered from multiple, large-scale datasets are:

- What percentage of people by African ethnic group was skilled in X?
- On X plantation, what was the gender ratio of slaves by African ethnic group?
- What percentage of Africans married people of the same ethnic group?
- What were the gender ratios of slaves identified as being of XXXX ethnicity?
- What injuries did people performing X type of work most commonly have?
- In X period, what was the percentage of slaves in Y place by ethnic group?
- In what records does the slave named XXXX appear? What were XXXX's professions?
- What places did he live? Who were his/her children and children's children?
- What was the value of slaves by ethnicity in X period? By skills and gender?

Making statistical data easily available and securely preserved is, then, one aspect of the project. Making that data understandable is another. Scholars and students—and anyone with Internet access—can search and browse (or download) individual datasets or the entire collection of datasets. Users can formulate questions (like those spelled out in section I.A. above), and get calculated answers not only in the form of numbers but also visual graphs: pie and line charts and histograms. The project also has a visualization platform that connects slaves to family members creating a complex web of social and kinship networks. This functionality affords nuanced views of the interconnectedness of individual slaves within the larger data collection.

By aggregating multiple datasets about slaves and developing tools that allow users to visualize and analyze this information, *Slave Biographies* empowers users to see broad patterns within a big set of data and identify small stories about individual slaves and their families. Ultimately with *Slave Biographies* scholars will be able to ask and answer unexpected questions that arise from the mass of biographic data in the repository and give life to slave experiences in the Atlantic World through their biographies contain in the data.

Into the future, *Slave Biographies* intends to enhance the visualization layer of the platform to allow users to map the data; maps of the Americas and Africa will illustrate from the places where individual slaves (based on ethnic identifications) hailed and to where they were finally brought. A time scrubber will enable scholars and students to see temporal and spatial shifts in patterns, a visualization of the slave trade attached to names and individuals. Another frontier for the project marries the geo-spatial mapping layer with richly illustrated slave biographies in an iPad application as well as a Windows Surfaces version for museum exhibition. Prototyping is already underway for this experience, which we hope will provide an engaging access point into the repository for general audiences.

Inspired by DH: The Day of Archaeology

Richardson, Lorna-Jane
l.richardson@ucl.ac.uk
UCL Centre for Digital Humanities, United Kingdom

Inspired by DH: The Day of Archaeology

The Day of Digital Humanities has become part of the Digital Humanities landscape: Inspired by the success of the Day of Digital Humanities project, the Day of Archaeology (DoA) was established as a voluntary project by a group of professional digital archaeologists and PhD students in 2011. The project is organised and run for free, and server space and staff time is donated and voluntary. The aim of the DoA was to utilise digital and participatory technologies, using a simple Wordpress-based platform, that would enable even the least digitally-minded archaeologist to share their work within the archaeology community and with the wider public. The web project aims to collate archaeological experiences and connect archaeologists across the world, using a variety of digital technologies. The participants record and share their Day on the Wordpress-based DoA website: www.dayofarchaeology.com, alongside photo-sharing sites, Facebook, YouTube and Twitter.

The Day of Archaeology www.dayofarchaeology.com has subsequently developed into an annual Public Archaeology event, which offers a unique insight into the working day of archaeologists worldwide. Participant-archaeologists have come from Asia, North and South

America, Europe, Australia and Africa. They have written about their day from excavations, offices, museums, community projects, the tourist industry, local government and voluntary groups. The project participants aim to answer a simple question in their contribution; “what do archaeologists do?”. Participants have contributed blog posts, films, photos, Tweets, Facebook pages, archive 'bingo' and 'ask an archaeologist'. The first event was held on the July 29th 2011, where some 500 people working, studying or volunteering in archaeology projects around the world contributed blog posts describing their day. The published posts and text are not scripted by the organisers, and only minimally edited to avoid defamation or incorrect information being shared.

The resulting website presents a behind-the-scenes view of archaeology that incorporates not only the exciting discoveries often showcased in Public Archaeology, but also everyday details of archaeological work in the real world. This project aims to move the public understanding of archaeology away from the 'Indiana Jones' model of excitement and object-oriented discovery, to one that can appreciate the painstaking and vital work undertaken by professionals and volunteers to protect, preserve and interpret our shared pasts.

This paper will explore how the Day of Digital Humanities model has translated into a more defined discipline; what the DoA has learned from Day of Digital Humanities; it will presents details of the project, how it was organised and who participated; the difficulties and benefits of the web-based model when applied to archaeology; how social media has been used; critical reflections on how the project has engaged with different audiences and what impact the DoA has had, and will have as it develops, for participants, the archaeology community and the wider public.

Five desiderata for scholarly editions in digital form

Robinson, Peter
p.m.robinson@bham.ac.uk
University of Saskatchewan, Canada

Scholarly editions have received considerable attention from the digital humanities over the last decades. There are several reasons for this: the fundamental place in the academy occupied by scholarly editions, with many humanities disciplines basing their work on texts which require establishment; the amenability of scholarly editions to computer methods, in terms of the digital representations of primary sources from which editions derive, and in terms of the highly-structured nature of the editions themselves, lending itself naturally to complex computer encoding. Accordingly, we have seen many digital editions made, in many different forms, and we are beginning to see too the first attempts towards a theory of digital editions, as a phenomenon distinct from print editions (which are, indeed, the subject of many theoretical debates): thus articles by Kiernan, Gabler, Pierazzo, Siemens and Robinson.

This paper will use the knowledge we have gained from our experiences, and the first competing theoretical discussions of digital editions, towards the statement of a set of desiderata, expressed as five propositions, collectively declaring the principles upon which scholarly editing in the digital medium should proceed. At the least, these will serve as starting points for useful discussion.

Proposition 1: **A digital edition should encode both the text of the document, line by line and page by page, and the text of the work which the document text instances, chapter by chapter, paragraph by paragraph (or, poem by poem, line by line).** One should be able to examine the text of the document, a page at a time; one should be able to read the text of the work, a chapter at a time, a poem at a time. This might seem obvious: yet several recent editions (for example, Sutherland’s edition of the Jane Austen manuscripts online), the thirty or so online transcription tools listed by Ben Brumfield, and the genetic transcription system proposed by the Text Encoding Initiative (Burnard et al), all assert that one need encode only the disposition of the text on the document. Thus, not one of the separate works contained in the Austen manuscripts is accessible as a work in the Sutherland edition. It is a mistake not to encode both document and work. It is certainly more difficult (because of the all-too-familiar overlapping hierarchy problem) to encode both. But it can be done, and it should be done.

Proposition 2: **Every act of editing in a digital edition should be attributed explicitly to the person who did it.** Any act of editing, in any medium, requires knowledge and effort: an edition is made from thousands, millions of such acts. Every such act should be recognized, and explicitly linked to the person who did it. Our confidence in editions comes from knowing who was responsible for each act. In the digital medium, as in the print medium, attribution is everything. But in the digital medium, we can go further: we can label who did what. And we should.

Proposition 3: **Everything in digital editions should, by default, be made available under a Creative Commons Attribution Share-alike licence.** Editing in the digital medium is profoundly collaborative. Even if one scholar does all the work of transcription, encoding, interface design and publication on his or her own: others will want to take elements of that edition and reuse it

in ways that the original scholar could never anticipate. Further, we may expect, as the movement to 'social editing' gathers pace, that we will see more and more informed readers becoming editors: contributing transcriptions, identifying documents, enriching their encoded texts by labelling persons, places, events. It is not quite true to say that digital editions belong to us all: but it is nearly true enough for us to make us much as possible as free as possible to all: hence, the Attribution Share-alike licence. We should not impose the 'non-commercial' restriction (which is too often a back-door way of maintaining the old worst habits of academic culture, to reward our friends and punish our enemies). Indeed, we should welcome and encourage commercial interests to provide the best interfaces they can to our editorial materials; the 'share-alike' provision will foreclose any commercial attempt to monopolize the text. Nor should we require the 'no-derivative works' restriction: we should welcome the scholar who wants to take what we did and use it as the starting point for his or her work — so long as this scholar acknowledges our work.

Proposition 4: **All the materials in a digital edition should be available independent of any one interface.** It should be possible (for example) for a scholar interested in the Greek New Testament to take the text of the transcription of Codex Sinaiticus given on the British Library website, combine it with other texts taken from other places, and present it in a distinct interface, offering tools and facilities nowhere else available. To make this possible, it will not be enough for editors to provide text: they must provide the facility (whether through metadata or an ontology or an API) to allow that text to be taken up and given out through an interface completely independent from the original digital publication. Of course, this cannot happen unless the materials are free of any restrictive licence, as argued in proposition 3.

Proposition 5: **All the materials in a digital edition should be held in a long-term sustainable data store.** Large scale data storage facilities, maintained in perpetuity as part of an institution's core mission ("in perpetuity" that is, as long as the institution lasts), have become commonplace in the last decade, thanks to the success of the institutional repository movement. Yet, very few editions ground their data in an institutional repository or similar facility. Institutional repositories are mature, well able to dispense scholarly edition materials, of every kind, and are increasingly recognized by universities and other memory institutions as the digital equivalent of a print library. We can and should use them.

A survey of existing digital editions against these propositions would yield interesting reflections. It appears that not one of the many digital editions so far made satisfies all five propositions; and rather many fail to satisfy even one. It is possible, of course, that these five propositions are flatly wrong. It is also possible that much of what we have been doing, under the dizzying spell of the technologies that press upon us, is wrong, and needs to change.

References

Brumfield, W. B. (2012). Crowdsourced Transcription Tool List. Blog entry for April 11, 2012. http://manuscripttranscription.blogspot.co.uk

Burnard, L., F. Iannidis, E. Pierazzo, and M. Rehbein (n.d.) An Encoding Model for Genetic Editions. http://www.tei-c.org/Activities/Council/Working/tcw19.html .

Codex Sinaiticus Online. http://www.codex-sinaiticus.net/en/ .

Gabler, H. W. (2007). The Primacy of the Document in Editing. *Ecdotica*, 4. 197–207.

Gabler, H. W. (2010). "Theorizing the Digital Scholarly Edition." *Literature Compass*, 7. 43-56.

Kiernan, K. (2006). "Digital facsimiles in Editing." In Burnard, L., K. O'Brien O'Keeffe and J. Unsworth (eds.) *Electronic Textual Editing*. New York: Modern Language Association of America, 262-268.

Mingana Collection online. http://vmr.bham.ac.uk/Collections/Mingana/

Pierazzo, E. (2011). A Rationale of Digital Documentary editions. *Literary and Linguistic Computing*, 26. 463-477.

Robinson, P. M. W. (2013). "Towards a theory of digital editions." *Variants* 10. 105-132.

Siemens, R., M. Timney, C. Leitch, et al. forthcoming. Toward Modeling the Social Edition: An Approach to Understanding the Electronic Scholarly Edition in the Context of New and Emerging Social Media. *Literary and Linguistic Computing*.

Sutherland, K. (2010). *Jane Austen's Fictional Manuscripts Digital Edition*. http://www.janeausten.ac.uk/index.html

A social network analysis of Rousseau's autobiography "Les Confessions"

Rochat, Yannick
yannick.rochat@epfl.ch
EPFL, Switzerland

Bornet, Cyril
cyril.bornet@ozwe.com
OZWE, Lausanne, Switzerland

Kaplan, Frédéric
frederic.kaplan@epfl.ch
EPFL, Switzerland

Introduction

We propose an analysis of the social network composed of the characters appearing in Jean-Jacques Rousseau's autobiographic *Les Confessions*, with existence of edges based on co-occurrences. This work consists of twelve volumes, that span over fifty years of his life.

Having a unique author allows us to consider the book as a coherent work, unlike some of the historical texts from which networks often get extracted, and to compare the evolution of patterns of characters through the books on a common basis. *Les Confessions*, considered as one of the first modern autobiographies, has the originality to let us compose a social network close to the reality, only with a bias introduced by the author, that has to be taken into account during the analysis. Hence, with this paper, we discuss the interpretation of networks based on the content of a book as social networks. We also, in a digital humanities approach, discuss the relevance of this object as an historical source and a narrative tool.

Literature review

Prior to this work, comparable studies have been published (Elson et al. 2010; Elson 2012), with edges based on conversations. After presenting a model to build "conversational networks" from classic novels, the authors conduct a social network analysis from which they can conclude that "as the number of characters in a novel grows, so too do the cohesion, interconnectedness and balance of [its] social network".

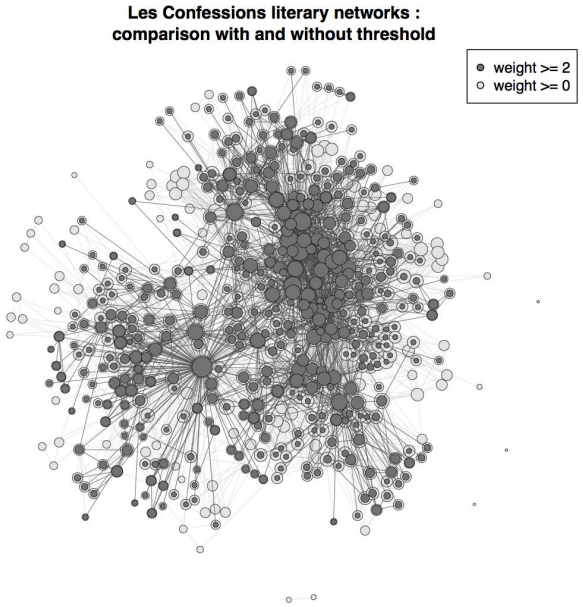
In (Moretti 2011), the author proposes the use of network theory to analyse the plot of Shakespeare's Hamlet. Finally, the study consists of re-telling the story via networks in order to sensitize the reader to this problematic, but doesn't develop any tool or concrete methodology.

Another recent paper (Carron, et al. 2012) proposes a statistical method invoking concepts of small-world, centrality and assortativity, with the objective of detecting real facts from fictional ones in mythological narratives.

Methodology

We propose a method that allows building a network from an index of names, and pages on which they occur (528 names on 672 in the selected – Slatkine, 2012 – edition). Vertices in the network represent the characters. To determine the existence of an edge between two characters, we have to deal with two constraints : the page is a restriction we have to get around, and some co-occurrences may be too weak to mean anything. Therefore, we take into account co-occurrences on same and adjacent pages (a name on last line of page n and a name on first line of page n+1 are closer than two names on first and last line of the same page), and then restrict the meaningful links to those reaching a certain level of significance. In this study, an edge with a coefficient of 1 means it links two adjacents nouns. An edge with a coefficient of 2 means two times adjacents nouns, or two names occuring on the same page. With that in mind, we choose to impose a threshold of 3 on links for them to be considered, so that no two characters are linked by error.

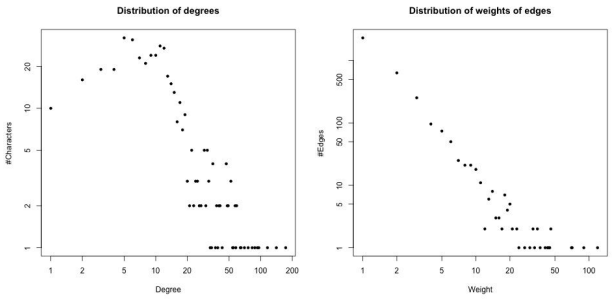
The resulting network is composed of 435 vertices and 3572 edges situated in a single connected component instead of 528 vertices and 4138 edges without threshold (links incident to non-contemporary characters like Plato or Copernic, cited together by Rousseau because of the influence of their work on his, have disappeared). It is undirected, i.e. relations are reciprocal. For comparison, the network with co-occurrences per page only, without threshold, is composed of 528 vertices and 2047 edges.



Analysis

Average path length is 2.48 steps, a small number but equal to what is obtained from random graphs generated with same order and density. However, diameter is equal to 10, which is high compared to 4 in the random case. The fifty years of Rousseau's life covered by *Les Confessions* lead to characters of the beginning and end of his life far away one from the other in terms of network distance. The comparison with random cases also yields an interesting result in the case of transitivity (closure of triplets of characters), which is equal to 0.299 against 0.038 in the random case, and global clustering coefficient equal to 0.724 against 0.038. These two results lead to assert that Rousseau links strongly characters between them in his narratives. According to (Watts et al., 1998), the network satisfies conditions to be considered as a "small-world" network (with possible discussion because of the high value of diameter).

Minus some noise on both sides, distribution of degrees of vertices and weights of edges show obvious power-law shape. Distribution of degrees has mode equal to 6, which is an interesting result since such a shape is common with many networks, but not a known results for literary or narrative networks. This implies that the author usually cite characters at least a few times, or with many other characters at the same time.



In (Newman et al. 2003), the authors define assortativity as the correlation of degrees of adjacent nodes. They conclude that social networks have positive assortativity, which is due to the frequent group structure observed on networks of this type. Assortativity of degrees computed on the network equals -0.114, while in the random case we obtain -0.006. In this work, we explore the potential explanations, from a possible bias introduced by the author, to a criticism of our method of creating the edges.

In the rest of the work, we still plan to show how the roles of protagonists can be identified, followed and compared via centrality indices like eigenvector centrality (centrality measure of a character depends on the ones from his neighbours, as it does with theirs). The question of dynamics of a literary network, linked to the chronological way Rousseau wrote the book, will also be considered.

References

Barabási, A.-L., and R. Albert (1999). Emergence of scaling in random networks. Science, 286(5439): 509–512.

Brandes, U., and T. Erlebach (2005). Introduction. In *Network Analysis*, volume 3418, pages 1–6. Berlin Heidelberg: Springer.

Mac Carron, P., and R. Kenna (2012). Universal properties of mythological networks. *EPL (Europhysics Letters)*, 99(2): 28002, July 2012.

Elson, D. K. (2012). *Modeling narrative discourse*. Ph.D. thesis, Columbia University, New York City.

Elson, D. K., N. Dames, and K. R. McKeown (2010). Extracting social networks from literary fiction. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL 2010)*, Uppsala, Sweden.

Marina, H., B. Ulrik, P. Jürgen, and M. Ines (2012). Studying Social Networks: A Guide to Empirical Research. Campus Verlag.

Moretti, F. (2011). Network theory, plot analysis. http://litlab.stanford.edu/?page_id=255

Newman, M. E. J., and J. Park (2003). Why social networks are different from other types of networks. *Phys. Rev. E*, 68(3): 036122, September 2003.

Wasserman, S., and K. Faust (1994). *Social Network Analysis: Methods and Applications*. Cambridge University Press, November 1994.

Watts, D. J., and S. H. Strogatz (1998). Collective dynamics of 'smallworld' networks. *Nature*, 393:440–442, June 1998.

From Anecdote to Data: Humanities Scholars Beyond the Tenure Track

Rogers, Katina Lynn
katina.rogers@virginia.edu
Scholarly Communication Institute, University of Virginia, United States of America

Though humanities graduates have long engaged in a range of stimulating careers, little data has been collected on humanities scholars working outside the professoriate. Consequently, discussions about alternative academic careers—those within the orbit of universities and cultural heritage institutions, but off the tenure track—have been

largely anecdotal. In order to ground the conversation, the Scholarly Communication Institute (SCI) initiated a study in 2012 to investigate perceptions about career preparation provided by graduate programs. The study was a directive from SCI's ninth summer meeting in 2011, which identified graduate education reform as an area of critical importance to the current humanities landscape.

The main goal of the study, which focused primarily on the context of North American higher education, was to establish a body of data that can serve as a foundation on which to base recommendations for new and revised methodological training. The results of the study reveal clear patterns that highlight the current strengths of graduate education relative to non-professorial employment, as well as significant opportunities for improvement.

The changing nature of career paths for humanities scholars is an issue of particular concern to digital humanities practitioners, who have long been working in the kinds of hybrid roles that the term "alternative academic" has come to describe. Many of the skills implicit in digital humanities scholarship and work products—including collaboration, project management, and technological fluency—are becoming increasingly important in new models of graduate training, even among programs not specifically allied with the digital humanities.

BACKGROUND

While doctoral study is a time of intense focus, it is also deeply exploratory. This exploration traditionally takes shape through the research process, as candidates follow the winding labyrinth of a line of inquiry, its antecedents, and its significance. Universities understand and value freedom of this nature; indeed, the fundamental structure of academic employment—tenure—is built around the importance of protecting the freedom of academic inquiry.

Increasingly, though, students need space for another kind of exploration, one more directly related to their future employment opportunities. The myth of a single (academic) job market persists in graduate programs today, perpetuated by departments that measure prestige by the tenure-track placements of their graduates. However, the convergence of increased casualization of the academic work force, a period of high unemployment, and steady enrollment in graduate programs means that people with advanced humanities training increasingly seek intellectually satisfying positions outside the professoriate. Following the 2011 launch of #Alt-Academy, a collection of essays edited by Bethany Nowviskie, the neologism and Twitter hashtag #alt-ac became a widely used shorthand to describe these kinds of careers, together with the excitement and challenges that accompany them.

In addition to the rich narrative material gathered under the #alt-ac umbrella, several other studies provide groundwork for SCI's recent work. In particular, the 2012 report by the Council of Graduate Schools and the Educational Testing Service titled "Pathways Out of Graduate School and Into Careers," provides a valuable look at graduate education and employment in the U.S. across all disciplines. An earlier study, "Ph.D.'s—Ten Years Later" (Nerad and Cerny, 1996), explores the experiences of Ph.D. holders working in business, government, and non-profits. It provides incredibly useful context, but the data from the study no longer accurately reflects the current academic or employment environments.

While both of these studies provide useful baseline information and analysis, the disciplinary scope of each is quite broad, making it difficult to assess finer-grained issues particular to the humanities. By focusing on a narrower segment of the academic population—humanities scholars working outside the tenure track—SCI's study can probe more deeply into issues that concern that group.

METHODS

The study consisted of two main phases: one public, one confidential. The first phase involved creating a public database of self-identified alternative academic practitioners. The database was built within the framework of the #Alt-Academy project in order to leverage the energy of existing conversations.

The second phase of the study comprised two confidential surveys. The primary survey targeted people with advanced humanities degrees who self-identify as working in alternative academic careers, while a second targeted employers that oversee employees with advanced humanities degrees. Because we were working with a somewhat nebulous population, our subsequent distribution focused on "opt-in" strategies—especially social media, listervs, and traditional media coverage. While this method has limitations, we hoped to learn something not only from the content of the responses, but from the number and type of respondents.

Data collection extended from July to October, 2012. Overall, the surveys had a very strong response, though the response rate also highlighted an important discrepancy. Nearly 800 people completed the main survey—almost four times our initial goal of 200 respondents. The employer survey, however, fell slightly short of our more modest goal of 100 respondents, totaling about 80 responses. The uneven response rate underscores the significant difference in engagement level on the part of job seekers compared to employers.

FINDINGS

Analysis is still in progress; the final report will be completed and published by August 2013, at the end of SCI's current phase of funding. The preliminary results of the surveys strongly suggest that while humanities graduates can and do apply their knowledge and skills to wide assortment of careers, there are many ways in which graduate programs could better equip them for the paths they take. Further, many of the skills employers report as desirable for alternative academic roles—such as project management, collaboration, and communicating with varied audiences—would also enhance the research, teaching, and service of those who do pursue academic roles.

Unsurprisingly, the data shows that a large majority of students enter graduate school expecting to pursue careers as professors—a total of 74%. What is perhaps more interesting is their level of confidence: of that 74%, 80% report feeling fairly certain or completely certain that this was the career they would pursue. These expectations are not aligned with the actual career outcomes of the respondents, or with humanities graduates more broadly.

Deepening the problem, students report receiving little or no preparation for careers outside the professoriate, even though the need for information about many different careers is acute. Only 18% reported feeling satisfied or very satisfied with the preparation they received for careers other than the professoriate. The responses are rooted in perception, so there may well be resources available that students are not taking advantage of—but whatever the reason, it is clear that students do not feel that they are being adequately prepared.

NEXT STEPS

Through a series of conversations with experts in the coming year, SCI will explore strategies to better equip students for a variety of careers without sacrificing disciplinary rigor. Based on the outcomes of the meetings, SCI plans to draft recommendations encouraging humanities departments to consider evaluating and modifying required aspects of their graduate-level curricula in ways that best serve students and the health of the discipline.

One way to move toward curricular change is to encourage humanities departments to form more deliberate partnerships with the inter- and para-departmental organizations that are already engaging in this kind of work. Traditional and digital humanities centers have jump-started excellent training programs, research projects, and public-facing work, though opportunities frequently take the form of extracurricular fellowships or informal training programs (such as the Digital Humanities Summer Institute and THATCamps, which both provide short-format, non-credit training opportunities). If departments that wish to move in similar directions connect with these centers, there may be opportunities to share infrastructure (physical and digital), expertise, time, and funding.

While informal programs have been a good starting point, incorporating successful training elements into the structures and core curricula of departments is an important move, especially in terms of sustainability and increased access (for all graduate students, not only those who win competitive spots in small programs). When individuals and small centers are supported by robust partnerships with traditional academic departments, the possibility for sustainable change becomes even greater.

CONCLUSION

This study represents an important step in the path of rethinking graduate education and academic employment, and we hope it helps to lay the groundwork for further study and concrete action. By making our data publicly available, we hope that other scholars will deepen the analysis of the responses that we have received. We also hope that an increasing number of departments will accurately track —and publish—data on the career paths of their former students. Increased information and transparency are critical to fostering an academic community that recognizes the value of permeable boundaries. Finally, we hope that the humanities community will strengthen its efforts to engage with the public. If, rather than feeling constrained by the exclusivity of a tenure-track career path, students instead feel free to explore ways to apply their humanities training to a broad spectrum of paths, their work would enrich both the academic community and the broader public.

References

Council of Graduate Schools (2012). *Pathways Through Graduate School and Into Careers*. 19 April 2012. http://pathwaysreport.org .

Nerad, M., and J. Cerny (1999). From Rumors to Facts: Career Outcomes of English PhDs. ADE bulletin 32.7. 11. (30 July 2012). http://www.mla.org/bulletin_124043 .

Nerad, M., and J. Cerny (2012). "Ph.D.'s—Ten Years Later." (30 July 2012). http://depts.washington.edu/cirgeweb/c/research/phd-career-path-surveys/phds-ten-years-later/ .

Nowviskie, Bethany, (ed). (2012). #Alt-Academy. MediaCommons (2011). (1 Oct. 2012). http://mediacommons.futureofthebook.org/alt-ac/ .

Scholarly Communication Institute (2011). New-Model Scholarly Communication: Road Map for Change. July 2011. http://uvasci.org/past-institutes/new-model-scholarly-communication/sci-9-report/ .

Mapping Editions: Literary Editions and GIS (a field report)

Roland, Meg
mroland@marylhurst.edu
Marylhurst University, United States of America

This paper will grapple with the relationship between geography and literature or, to borrow a phrase from Kevin Bartoy, "between dirt and discussion."[1] Based on key foundational questions and a brief review of current humanities-based GIS projects, I propose that GIS (Geographical Information Systems) offers the potential to re-imagine the form and practice of literary editions and affords pedagogical opportunities for spatial analysis in literary studies.

This past spring, students in my Literature and Maps class experimented with Google Earth to create a small literary edition of Malory's Roman War account (from *Le Morte Darthur*) as a means to explore how textuality and ambiguity might be accommodated in a system based on locational coordinates and quantitative principles. With GIS projects already changing the practice of *analyzing* literary texts; does it also hold the potential to change the practice of *producing* literary texts such that they foster spatial thinking and new means of reading and analyzing literary texts? Annotated editions within a geographical platform offer the potential for student projects as contingent, learning-based editions which explore digital and imaginative terrain in the production of literary editions and digital humanities projects.

Theoretical and historical contexts that arise from the "spatial turn" in literary studies and the humanities[2] are key foundational questions that ideally should inform the editorial principles of GIS-based literary editions prior to the technical work of layering a text into a GIS system. Such spatially-based projects in the digital humanities hold significant potential for collaboration between geographers and humanists, between the "the poets and the geeks[3] .

Researchers in the discipline of history have enthusiastically embraced the potential of GIS and, indeed, Richard White, in his introduction to *Placing History: How Maps, Spatial Data, and GIS are Changing Historical Scholarship* has predicted that GIS databases "are going to change the practice of history[4] ." It is significant, however, that in the seminal text *The Spatial Humanities: GIS and the Future of Humanities Scholarship,* literary scholars and editors are notably absent as contributors. The Scholars Lab, of course, has begun to generate work such as *Mapping St. Petersburg* which is experimenting with "literary cartography" to explore the representation of space in literary texts[5] .

There has been a strong pedagogical interest in "literary mapping" via web sites such as Google Lit Maps, but itinerary-based mapping provides only a rudimentary means by which to investigate the potential intersection of GIS and literary study. Indeed, "lit-maps" run the risk of sidestepping Michel de Certeau's concern that "surveys of routes miss what was: the act itself of passing by[6] and thereby reducing spatial analysis of literary texts to the reductive reproduction of itineraries.

GIS projects offer the potential for a more complex engagement with spatial methodology than mere itinerary mapping, including, as David Bodenhamer has noted, "text-based geographical analysis, multimedia, animated maps, deep contingency, deep mapping, and the geo-spatial semantic web[7] ." The very concept of "deep contingency" offers striking potential for editors and readers of GIS-based literary editions, a potential that has yet to be explored. Such digital humanities-based editions, I would argue, need to accommodate fluidity if they are to engage readers in the kind of analysis Bodenhamer envisions.

Currently, there are a few digital humanities projects that explore the possibility of a GIS-based literary edition. By virtue of the presence of a full version of the text within the GIS system, *Mapping the Lakes: A Literary GIS* comes closest to creating a true GIS-based literary edition, although the data analysis is less than fully integrated with a readerly analysis of the texts. David Wallace's project, *Europe: A Literary History* [8] smartly utilizes key principles of GIS to produce a collaborative approach to literary history that also explores geographic relation in light of what Iain Chambers has conceptualized as "fluid cartographies[9] ." Many contemporary geographers have chafed under the ascendency of GIS in geography departments and the limitations of qualitative analysis within GIS systems; in this regard, literary scholars bring both naiveties and methodological approaches that can critically engage in GIS reception and use. For students, the technical and editorial issues in creating a GIS-based edition offers an opportunity to engage in spatially-attuned literary practices.

This "field report" will engender a conversation about possible methodologies for literary GIS editions and share my students' foray into the creation of a Google Earth-

based edition. I hope to initiate a discussion of pedagogical opportunities for students to engage with literary texts as editors, analytic cartographers, and readers in the process of producing digital and spatial humanities projects.

Notes

1. Kevin Bartoy, *Between Dirt and Discussion: Methods, Methodology, and Interpretation in Historical Archeology* (2006).
2. For example, foundational work by Henri Lefebvre, *The Production of Space* (1974) and Edward Soja, *Postmodern Geographies: The Reassertion of Space in Postmodern Critical Theory* (1989).
3. Cohen, Patrica, "Digital Keys for Unlocking the Humanities' Riches," *The New York Times,* November 16, 2010.
4. Richard White, *Placing History: How Maps, Spatial Data, and GIS are Changing Historical Scholarship* (2008).
5. See Mapping Saint Petersburg.
6. Michel de Certeau, "Walking in the City."
7. David Bodenhamer, *The Spatial Humanities: The Future of Humanities Scholarship* (2010).
8. Europe: A Literary History
9. Iain Chambers, "Maritime Criticism and Theoretical Shipwrecks," *PMLA* 2010.

The DARIAH Approach to Interdisciplinary Interoperability

Romanello, Matteo
matteo.romanello@gmail.com
Deutsches Archaeologisches Institut; King's College London, United Kingdom

Beer, Nikolaos
nikolaos.beer@uni-paderborn.de
Musikwissenschaftliches Seminar Detmold/Paderborn

Herold, Kristin
kristin.herold@uni-paderborn.de
Musikwissenschaftliches Seminar Detmold/Paderborn

Kolbmann, Wibke
wk@dainst.de
Deutsches Archaeologisches Institut

Kollatz, Thomas
kol@steinheim-institut.org
Salomon Ludwig Steinheim-Institut für deutsch-jüdische Geschichte

Rose, Sebastian
sebastian.rose@uni-koeln.de
Universität Köln - Historisch-Kulturwissenschaftliche Informationsverarbeitung Universität zu Köln

Walkowski, Niels
walkowski@bbaw.de
Berlin-Brandenburgische Akademie der Wissenschaften

Introduction

In this paper we present the preliminary results of the work that is being carried out in the framework of the DARIAH-DE [1] project on the topic of interdisciplinary interoperability. DARIAH-DE is the German branch of the EU-funded Digital Research Infrastructure for the Arts and Humanities (DARIAH) project with 17 institutions belonging to a wide range of disciplines as project partners. This fortunate circumstance compelled us to tackle the problem of how to achieve greater interoperability between the digital collections of institutions coming from different disciplines.

We first present our pragmatic approach to interoperability and then discuss some use cases that were developed to complement a set of recommendations on interdisciplinary interoperability. These recommendations are meant, on the one hand, to help Humanities institutions to integrate their collections into the DARIAH data federation and, on the other hand, to let them benefit fully of the services provided by the infrastructure.

The importance of interoperability when building a research infrastructure becomes evident as soon as one considers, for example, even a basic search mechanism over several collections, also known as federated search. Building such a federated search requires institutions to expose, at least, an end-point that can be harvested using a given protocol (Aschenbrenner, et al. 2010). The Collection Interoperability group in Bamboo — an analogous infrastructure project in the United States — was faced with the same problem of making collections interoperable and decided for developing adapters to make collections compliant with the OASIS Content Management Interoperability Services (CMIS) standard. Let us see now what has been DARIAH's approach to interoperability.

Methodology

If interoperability is difficult, true interoperability across disciplines is perhaps even more so as — particularly when talking about semantic interoperability — the narrower is the application domain the higher are the chances of achieving some results. This is the case, for example, when using ontologies for this purpose as shown by (Marshall and Shipman 2003).

Therefore, given the number of domains and disciplines that DARIAH is trying to cater for, the solution of mapping the meaning of content in different collections onto the same ontology or conceptual model appeared soon to be not a viable one. As Bauman makes clear while discussing the topic of interoperability in relation to the goal and mission of TEI (Bauman 2011), the drawback of adhering to standards for the sake of interoperability is the consequent loss in terms of expressiveness.

Instead,DARIAH position on this respect is of allowing for crosswalks between different schemas: a sort of "crosswalk on demand". Infrastructure users will be able to use the Schema Registry — a tool which is being developed in DARIAH-DE — to create crosswalks between different metadata schemas so that they are made interoperable.

Our main goal was to devise a set of guidelines that is realistically applicable by partner institutions as part of their policies. Therefore, the first preliminary step was to gather and analyze information about the digital collections of the partners with regard to interoperability. We identified the following key aspects to guide our analysis:

1 **Identifiers:** two aspects of identifiers were considered: on the hand, their persistence over time, which is a crucial aspect for any infrastructure project, and on the other hand the use of common, shared identifiers(e.g. controlled vocabulary URIs)to express references to the same "things", that is one of the core ideas of Linked Data.

2 **APIs and Protocols:** APIs and protocols are essential as they allow for workflows of data access and exchange not necessarily dependent on human agents. This idea is implied in the notion of "blind interchange" discussed by Bauman with the only difference being that, in our own vision, as little human intervention as possible should be required.

3 **Standards:** using the same standard is in some, if not many cases, not enough to achieve real semantic, not only syntactic, interoperability. Therefore we discuss further aspects of standards in relation to interoperability such as multiple serializations of the same scheme, and the problem of adoption and adaption of schemes to different contexts.

4 **Licences:** licences, and specifically their machine-readability, play — perhaps not surprisingly — a crucial role within an interoperability scenario: not only should a licence be attached to any collection as soon as it is published online, but such licence should also be readable and understandable, for example, to an automated agent harvesting that collection.

These four aspects define the context for the use cases that are described in the next section and also define the core aspects that will be covered in the recommendations.

Use Cases

The main aim of the use cases was to show how basic, minimal standards, such as the OAI-PMH protocol, can be exploited together with already existing and re-usable tools in order to devise more advanced interoperability scenarios. The requirements of such use cases were that they could be implemented in a limited amount of time, without requiring much programming effort and using openly available tools and data coming from partner institutions.

Static OAI-PMH

The first use case focussed on tools that can be used to create an OAI-PMH end-point for collections without having to set up an own repository as this could be a problem particularly for small institutions or poorly funded disciplines. Since OAI-PMH will be part of the recommendations as the minimal protocol for accessing collections, we wanted to explore ways so to overcome the barriers to adopting it. We used as data collection a sample of *Kalonymos* [2] , a journal published by the Steinheim-Institut (STI), as it does not provide yet such an interface. The Dublin Core metadata for the articles in *Kalonymos*, stored in a static file, were then served by an OAI-PMH end-point created out of such static file by using the OAI-PMH Static Repository Gateway[3] , a piece of software written in C and openly available.

```
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
xmlns:ow="http://www.ontoweb.org/ontology/1#"
xmlns:dc="http://purl.org/dc/elements/1.1/">
      <ow:Publication rdf:about="oai:kobv.de-opus-bbaw:905">
            <dc:title>Eine Analyse des Kontextes wäre
                  hilfreich</dc:title>
            <dc:creator>Riedmüller, Barbara</dc:creator>
            <dc:subject>Wissenschaftsfreiheit</dc:subject>
            <dc:subject>Akademische Freiheit</dc:subject>
            <dc:subject>Forschungsfreiheit</dc:subject>
            <dc:subject>Genforschung</dc:subject>
            <dc:subject>General serials and their indexes</dc:subject>
            <dc:publisher>Berlin-Brandenburgische Akademie der
                  Wissenschaften</dc:publisher>
            <dc:publisher>BBAW. Interdisziplinäre Arbeitsgruppe
                  Gegenworte - Hefte für den Disput über Wissen
         </dc:publisher>
            <dc:date>1998</dc:date>
            <dc:type>Article</dc:type>
            <dc:format>application/pdf</dc:format>
      </ow:Publication>
</rdf:RDF>
```

Fig. 1
Sample output of RDFizer.

OAI to LOD

The OAI-PMH protocol is central also in the second use case, which seeks to show how the metadata served by an OAI-PMH end-point can be harvested and programmatically transformed into an RDF representation, more suitable for example in a Linked Data scenario.

The data we used come from a document repository of the Berlin-Brandenburg Academy of Science[4] containing Open Access publications by the academy members. We were able to transform into RDF the Dublin Core metadata served via the OAI-PMH interface of the repository by running the open source OAI-PMH RDFizer[5] .

The so obtained output, see fig. 1, could be enriched by adding links for example to the subject headings contained in the Gemeinsame Normdatei (GND) catalogue of the German National Library (DNB) [6] that are already available as RDF. However possible, this enrichment was not implemented for this use case as it required some extra work to overcome the lack of a lookup API or SPARQL end-point on the DNB side.

Marc21 to SKOS/RDF

The third use case consisted in transforming the thesaurus of the German Archeological Institute, currently encoded in Marc21XML and accessible via an OAI-PMH end-point, into an RDF representation of the same data encoded in SKOS — the W3C standard to publish Knowledge Organization Systems in the Semantic Web[7] .

Such transformation was made possible by the Stellar Console, an open source tool developed by Ceri Binding and Doug Tudhope(Keith et al. 2012) in the framework of the AHRC-funded project “Semantic Technologies Enhancing Links and Linked Data for Archaeological Resources”(STELLAR) [8] . The 80,00 Marc21 records of the thesaurus, after being harvested via the OAI-PMH interface, were transformed into an intermediate CSV file, which is in turn fed into the Stellar Console in order to produce a SKOS/RDF output consisting of slightly less than 1 million triples (Romanello 2012).

TEI to CIDOC-CRM

The last use case is another example of extracting more deeply structured semantic information from legacy data, in this case some letters, encoded in TEI, that are part of the project Carl Maria von Weber — Collected Works (WeGA)[9] .

We ran the TEI documents through an XSLT transformation written by Sebastian Rahtz[10] and included as part of OxGarage[11] , an online toolkit for the conversion of TEI-compliant documents into other formats. What this transformation does is to look for elements of specific types and to transform them into equivalent statements expressed by means of the CIDOC-CRM ontology and encoded in RDF/XML.

Straight out of the box, the tool created the corresponding semantic statements for the elements `<date>`, `<persName>` and `<placeName>`. As we observed also for use case 3, it is technically possible to enrich the RDF output by adding the URIs of the person or place that is being referred to in the document. This could be realized by leveraging the links to authority lists that are often included directly into the TEI elements, but the required effort goes beyond the scope of our current work.

```
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#">
   <E31_Document xmlns="http://purl.org/NET/crm-owl#"
                 rdf:resource="http://www.example.com/place/detmold
                 http://www.example.com/place/d http://www.example.com/place/dresden
                 http://www.example.com/place/prag"/>
   <P4_has_time-span xmlns="http://purl.org/NET/crm-owl#">
      <E52_Time-Span>
         <P82_at_some_time_within>
            <E61_Time_Primitive>
               <rdf:value>26.-27. Januar 1817</rdf:value>
            </E61_Time_Primitive>
         </P82_at_some_time_within>
      </E52_Time-Span>
   </P4_has_time-span>
   <E21_Person xmlns="http://purl.org/NET/crm-owl#"
               rdf:about="http://www.example.com/person/carolinabrandt">
      <P131_is_identified_by>
         <E82_Actor_Appellation
            rdf:about="http://www.example.com/persname/carolinabrandt">
            <rdf:value>Carolina Brandt</rdf:value>
         </E82_Actor_Appellation>
      </P131_is_identified_by>
   </E21_Person>
   <E21_Person xmlns="http://purl.org/NET/crm-owl#"
               rdf:about="http://www.example.com/person/grafenvizthum">
      <P131_is_identified_by>
         <E82_Actor_Appellation
            rdf:about="http://www.example.com/persname/grafenvizthum">
            <rdf:value>Grafen Vizthum</rdf:value>
         </E82_Actor_Appellation>
      </P131_is_identified_by>
   </E21_Person>
   <E53_Place xmlns="http://purl.org/NET/crm-owl#"
              rdf:about="http://www.example.com/place/prag">
      <P2_has_type rdf:resource="http://www.tei-c.org/type/place/settlement"/>
      <P87_is_identified_by>
         <E48_Place_Name rdf:about="http://www.example.com/placename/prag">
            <rdf:value>Prag</rdf:value>
         </E48_Place_Name>
      </P87_is_identified_by>
   </E53_Place>
</rdf:RDF>
```

Fig. 2
Sample output of OxGarage

Conclusions

We hope that the use cases above have shown how the use of open standards and protocols combined with open source, thus adaptable and reusable, tools can allow us to achieve some greater interoperability in a cost- and time-effective way.

References

Aschenbrenner, A., W. P. Andreas, T. Blanke, and M. W. Küster Towards an Open Repository Environment. *Journal of Digital Information* **11:1**. http://journals.tdl.org/jodi/article/viewArticle/758. (accessed 22 März 2010).

Bauman, S. (2011). 'Interchange vs. Interoperability'. In *Proceedings of Balisage: The Markup Conference. Balisage Series on Markup Technologies.* held 2011 in Montréal, Canada. http://www.balisage.net/Proceedings/vol7/html/Bauman01/BalisageVol7-Bauman01.html.

Marshall, C. C., and F. M. Shipman 'Which Semantic Web?' In *Proceedings of the Fourteenth ACM Conference on Hypertext and Hypermedia.* held 2003 in Nottingham, UK. Nottingham, UK: ACM, 57–66. http://portal.acm.org/citation.cfm?id=900063.

May, K., C. Binding, D. Tudhope, and S. Jeffrey 'Semantic Technologies Enhancing Links and Linked Data for Archaeological Resources'. In Zhou, M., I. Romanowska, W. Zhongke, X. Pengfei, and P. Verhagen (eds.), *CAA Proceedings 2012.* Amsterdam: Amsterdam University Press, 261–272. http://dare.uva.nl/aup/nl/record/412958.

Romanello, M. Skosifying an Archaeological Thesaurus. *Computers for the Classics*. http://c4tc.wordpress.com/2012/10/08/skosifying-an-archaeological-thesaurus/. (accessed 8 October 2012).

Notes

1. http://de.dariah.eu/
2. http://sourceforge.net/projects/srepod/
3. http://d-nb.info/025276212
4. http://edoc.bbaw.de/oai2/oai2.php
5. http://simile.mit.edu/repository/RDFizers/oai2rdf/README.txt
6. http://www.dnb.de/EN/Standardisierung/Normdaten/GND/gnd_node
7. http://www.w3.org/TR/skos-reference/
8. http://hypermedia.research.glam.ac.uk/resources/STELLAR-applications/
9. http://www.weber-gesamtausgabe.de/en/
10. http://www.zde.uni-wuerzburg.de/tei_mm_2011/abstracts/abstracts_papers#c249172
11. http://www.tei-c.org/oxgarage/

Widening the Big Tent: Amateurs and the "Failure of the Digital Humanities"

Rowberry, Simon
simon.rowberry@winchester.ac.uk
University of Winchester, United Kingdom

The Failure of the Digital Humanities

Mark Sample's "Unseen and Unremarked On: Don DeLillo and the Failure of the Digital Humanities" argues that post-1922 literary texts are being left behind as a part of the Digital Humanities (Sample 2012). This is a direct result of the Sonny Bono, or "Mickey Mouse," Copyright Term Extension Act, another apparent move towards perpetual copyright. These difficulties are compounded by other obstacles including closed access or disorganised archives, insufficient preservation tools for early computer usage, and authors who simply refuse to embrace the digital. Without the necessary permissions or archival material, scholars of these twentieth century scholars are becoming increasingly envious of their colleagues, who develop tools that would equally aid interpretation of these more recent authors. Mid-twentieth century literature is of particular relevance to Digital Humanities research, since many frequently cited precursors of electronic literature including Vladimir Nabokov's *Pale Fire* (1962), Julio Cortázar's *Rayuela* (1963), and the short stories of Jorge Luis Borges's *Ficciones* (originally published c.1960s), are still protected by the Sonny Bono Copyright Term Extension Act. Many of the theoretical issues that have been teased out of these texts — especially early hypertext theory (see Landow 1992; Bolter 1991; Joyce 2002) — perhaps can only truly be tested once many of these texts have been the subject of digital experimentation. This paper argues that although these projects are often not being carried out by faculty members, the need and potential uses for such tools among non-academic readers is demonstrated through the samizdat distribution of online versions and tools readily available

for all those who wish to conduct a Google search. The launch of the first authorized Pynchon e-books (Flood 2012) was met with dismissive claims that better samizdat copies had been in circulation for many years beforehand. These projects are coming into fruition externally to traditional (digital) humanities departments, spreading out to computer scientists' extracurricular projects or the work of those outside of the academy who build digital tools and resources for the love of the original literary artefact. A few examples of the diverse work being undertaken includes wikis (for authors such as Thomas Pynchon (Ware 2006) and Terry Pratchett (Anon. 2005)) databases (*Finnegans Wake* Extensible Elucidation Treasury (FWEET)) (Slepon 2005), interpretations of literary texts through social media on both a single platform, and a dense and complex ecosystem of literary engagement and reception (such as the recently organized group read of William Gaddis's JR centralised around the Twitter hashtag #occupygaddis) and many other forms that demonstrate potential platforms for further research and development.

Literature Review

This study fits into a wider field of readership and reception studies, an interdisciplinary research subject, which has had some crossover within the Digital Humanities. Anouk Lang's edited collection, *From Codex to Hypertext: Reading at the Turn of the Twenty-First Century*, which includes chapters on how reader recommendation systems are changing in the digital age (Wright 2012), the community of LibraryThing (Pinder 2012), and the network of reader reviews on Amazon (Finn 2012). Furthermore, the present study runs parallel to crowdsourcing in the Digital Humanities, most recently exemplified by the Transcribe Bentham Project (Causer and Wallace 2012), as many projects involve large numbers of volunteers to organize materials. Moreover, as Henry Jenkins *et al.* have recently suggested, the easy transmission and manipulability of media in the early twenty-first century is essential to ensure the text's viability, and the evidence of fan communities exploring literary texts suggests a desire for these more of these platforms. (Jenkins, Ford, and Green 2013) There have also been more specific papers exploring the use of particular social network platforms for literary reception (see Schroeder and den Besten 2009; Ketzan 2012) and how the use of these tools reflect the development of underlying software through the way users build on the platform (Howison and Crowston 2011).

Do "Amateurs" Fit into the Big Tent of the Digital Humanities?

There has been a considerable debate concerning the purview of the Digital Humanities, particularly the extent to which building tools is essential to being described as Digital Humanities. (Svensson 2012) This paper asserts that the Big Tent should be widened to include a broader spectrum of scholars, amateur or professional, who engage with the transformational nature of digital tools, whether engaging with new methods of collaborating and presenting interpretive data or building databases to explore the manipulable nature of the original texts. These pockets of activity demonstrate a potential audience for these tools and push the boundaries of what counts as fair use in ways that academic institutions typically shy away from for fear of lawsuits. The deformative acts (Samuels and McGann 1999) these projects often engage in can thus reveal the ways in which these texts reflect a Digital Humanities agenda despite their marginalized status as both amateur projects and remediated texts (Bolter and Grusin 2000) still protected by copyright. Furthermore, there is evidence of the acceptance of these projects through examining the number of citations to some of the most prominent projects such as FWEET, which has been cited as both an exemplar of hypertextuality (Krapp 2005) and a reference guide for Joyce's enigmatic text comparable to Roland McHugh's authoritative *Annotations to Finnegans Wake.* (Conley 2007) Thus, we can witness how these projects engage with the academy.

Case Studies

The present study focuses on two case studies to illustrate the range of productivity that has engaged the non-Digital Humanities community for two twentieth-century authors: James Joyce and Vladimir Nabokov. These two authors represent polar opposites regarding their respective estates' view of intellectual property rights and digital media. The Joyce estate has been involved in a couple of high profile copyright disputes leading to the dissolution of some major digital editions of Joyce's work, most prominently, Michael Groden's "Digital *Ulysses*." On the other hand, FWEET, maintained by Raphel Slepon, a former medical researcher and programmer, runs counter to the usually aggressive policies of the Joyce estate. FWEET collates allusions from McHugh's *Annotations to Finnegans Wake* (1980) and other major reference guides to Joyce's novel, as well as material collected from a range of independent contributors, into a database which allows the user to sift through a taxonomy of references, view all the noted allusions on a line-by-line basis, or search for particular tropes. The original text is obfuscated by the database's interface and thus the website acts as a reference

guide primarily rather than a readable digital edition of the text.

Meanwhile, the Nabokov estate has occasionally granted the use of his texts for digital work despite taking an aggressive policy towards intellectual property rights in post-Soviet Russia. Two digital Nabokov projects have been sanctioned since 1967: Ted Nelson's demonstration of *Pale Fire* as a hypertext in the late 1960s and Brian Boyd's *Ada* Online. Alongside these official projects, there have been a plethora of hypertext experiments with the whole or parts of *Pale Fire*. These examples of remediation begin to explore the generative network of Nabokov's most complex novel and demonstrate the novel's effectiveness as a precursor of hypertext literature. Both case studies highlight how two respected authors' works are being transformed by digital media without the intervention of digital humanists. Through careful study of the digital reception of the texts, we can not only learn how these texts are being transmitted and circulated by a popular audience, but also start to understand how these texts, currently protected by strict copyright laws, can and will be part of a wider Digital Humanities ecology.

References

Anon. (2005). Annotations — Discworld & Pratchett Wiki. http://wiki.lspace.org/mediawiki/index.php/Main_Page (accessed 30 October 2012).

Bolter, J. D. (1991). *Writing Space: The Computer, Hypertext, and the History of Writing.* Hillsdale: Lawrence Erlbaum Associates.

Bolter, J. D. and R. Grusin (2000). *Remediation: Understanding New Media.* Cambridge, MA: The MIT Press.

Causer, T., and V. Wallace (2012). Building A Volunteer Community: Results and Findings from Transcribe Bentham *Digital Humanities Quarterly* 6.2 http://www.digitalhumanities.org/dhq/vol/6/2/000125/000125.html .

Conley, T. (2007). Annotations to 'Finnegans Wake' (review). *James Joyce Quarterly* 1 (2). 363–366.

Finn, E. (2012). New Literary Cultures: Mapping the Digital Networks of Toni Morrison. In *From Codex to Hypertext: Reading at the Turn of the Twenty-First Century*, ed. Anouk Lang, 177–202. Amherst and Boston: University of Massachusetts Press.

Flood, A. (2012). Thomas Pynchon Finally Gives in to Gravity as Digital Backlist Is Published. *The Guardian.* http://www.guardian.co.uk/books/2012/jun/13/thomas-pynchon-digital-backlist-published (accessed 25 October 2012).

Howison, J., and K. Crowston (2011). Collaboration Through Superposition: How the IT Artifact as an Object of Collaboration Affords Technical Interdependence Without Organizational Interdependence. *Institute for Software Research. Paper 491.* http://repository.cmu.edu/isr/491 (accessed 30 October 2012)

Jenkins, H., S. Ford, and J. Green (2013). *Spreadable Media: Creating Value and Meaning in a Networked Culture.* New York and London: New York University Press.

Joyce, M. (2002). *Of Two Minds: Hypertext, Pedagogy and Poetica.* Ann Arbor: University of Michigan Press.

Ketzan, E. (2012). 'Literary Wikis: Crowd-sourcing the Analysis and Annotation of Pynchon, Eco and Others'. *Digital Humanities 2012*. held 16-22 July in Hamburg, Germany.

Krapp, P. (2005). Hypertext Avant La Lettre. In Hui Kyong Chun, W. and Keenan, T. (eds). *New Media Old Media: A History and Theory Reader*. 359–373. New York: Routledge.

Landow, G. P. (1992). *Hypertext: The Convergence of Contemporary Critical Theory and Technology.* Baltimore and London: The Johns Hopkins University Press.

McHugh, R. (1980). *Annotations to Finnegans Wake.* London: Routledge & Kegan Paul.

Pinder, J. (2012). Online Literary Communities: A Case Study of LibraryThing. In Lang, A. (ed). *From Codex to Hypertext: Reading at the Turn of the Twenty-First Century*, 68–87. Amherst: University of Massachusetts Press.

Sample, M. (2012). Unseen and Unremarked On: Don DeLillo and the Failure of the Digital Humanities. In *Debates in the Digital Humanities*, ed. Gold, M. K., 187–201. Minneapolis and London: University of Minnesota Press.

Samuels, L., and J. J. McGann (1999). Deformance and Interpretation. *New Literary History* 30(1). 25–56.

Schroeder, R., and M. den Besten (2009). Literary Sleuths Online: e-Research Collaboration on the Pynchon Wiki. http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1086671 (accessed 9 October 2012).

Slepon, R. (2005). Love's Old Fweet Fong. *FWEET*. http://fweet.org/pages/fw_prlg.php (accessed 30 October 2012).

Svensson, P. (2012). Beyond the Big Tent. In Gold, M. K. *Debates in the Digital Humanities*. 36–49. Minneapolis and London: University of Minnesota Press.

Ware, T. (2006-). Thomas Pynchon Wiki — A Literary/Literature Wiki. http://pynchonwiki.com/ (accessed 30 October 2012).

Wright, D. (2012). Literary Taste and List Culture in a Time of 'Endless Choice'. In Lang, A. (ed). *From Codex to Hypertext: Reading at the Turn of the Twenty-First Century*. 108–123. Amherst: University of Massachusetts Press.

Collaborative Authorship: Conrad, Ford and Rolling Delta

Rybicki, Jan
jkrybicki@gmail.com
Jagiellonian University, Krakow, Poland

Hoover, David L.
david.hoover@nyu.edu
New York University, USA

Kestemont, Mike
mike.kestemont@gmail.com
University of Antwerp, Belgium

Burrows's "Delta" is a popular authorship attribution algorithm (Burrows 2002). Suppose that we have an anonymous text which has to be attributed to one of a series of candidate authors for whom we have a number of reference samples as training material. Delta computes a dissimilarity score between the test item and all reference samples and attributes the anonymous text to the author of the sample to which it is most similar. We propose a procedure called "Rolling Delta", reminiscent of a number of earlier applications (e.g. Van Dalen-Oskam & Van Zundert 2007; Burrows 2010; Van Zundert & Van Dalen-Oskam 2012; Hoover 2012). The general goal is to visualize stylistic shifts in texts, for instance, in order to pinpoint authorial takeovers in the case of collaborative authorship. An implementation of Rolling Delta is freely available (Eder, Kestemont & Rybicki 2012).

First, each reference text is segmented into equal-sized, partially overlapping samples. If we specify a 'window size' of 5,000 and a 'step size' of 100, for example, the first sample of a text contains words 1-5,000, the second 101-5,101, etc. The procedure uses the relative frequencies of the *n* most frequent words in the reference collection. Subsequently, we compute a centroid (*C*) for each reference text, containing the mean relative frequency for each word in its windows, and the standard deviation. Then we divide the test text into windows and compute the 'Delta' between each test window *W* and each reference centroid, using the following formula— see Argamon (2008) for more details:

$$\Delta(C,W) = \sum_{i=1}^{n} \frac{1}{\sigma_i(C)} |\mu_i(C) - f_i(W))|$$

After "rolling" through the test text, we plot the resulting Delta series for each reference text. The lower the Deltas for a reference text, the more similar the style in the test windows — and vice versa. If the curve for a text shows a sudden drop, this may indicate a stylistic change in the test text, caused, for instance, by one author taking over from another. One can use vertical lines in the plot to mark the position of certain events in the test text as an aid in interpretation (e.g. chapter beginnings).

This method seems the perfect tool to study the notable literary collaboration between Joseph Conrad (1857-1924) and Ford Madox Ford (1873-1939): their three joint works, *The Inheritors* (1901), *Romance* (1903) and *The Nature of a Crime* (1909, 1924), and the various authorship claims by Ford, including that concerning a fragment of *Nostromo* (1904). The Conrad/Ford controversy was enhanced by the two authors' neurotic behavior that eventually took its toll on their collaboration and friendship. Physical evidence about authorship is complicated by the fact that Ford (and others, including John Galsworthy) took Conrad's dictation when he was sick or indisposed or could not make a deadline.

The Inheritors is Ford's second published novel, coming out nine years after *The Shifting of the Fire*. He did most of the writing himself, though he discussed it extensively with Conrad, whose role, he said, was "to give each scene a final tap" (Saunders 1996: 135-36). For *Romance*, based on Ford's earlier unfinished *Seraphina*, however, the consensus seems to be that it is about two thirds Conrad and one third Ford. According to Conrad:

> We collaborated right through, but it may be said that the middle part of the book is mainly mine with bits by F.M.H. — while the first part is wholly out of "Seraphina": the second part is almost wholly so. The last part is certainly three quarters MS. F.M.H. with here and there a par. by me.

According to Ford, "parts one, two, three and five are a mosaic of alternately written passages while part four is entirely Conrad's work" (Karl 1997: 147). Najder further comments that "the change in numbering the parts of *Seraphina* has caused some trouble for Conrad's and Ford's biographers. As late as the summer of 1901, the novel consisted of four parts, but ended, as it does now, with Kemp's trial. While continuing to write part 3, Conrad expanded it into another, which became part 4, and the last, part 5, written by Ford" (Najder 2007: 317). The third

collaborative work, *The Nature of a Crime*, was written almost exclusively by Ford and heavily edited by Conrad.

Ford's possible contribution to *Nostromo* — mostly based on the one large part of the manuscript in his hand — is limited to the novel's second part. Brice quotes a letter from Ford to Keating (1923 or 1925) saying he wrote 10,000 words of *Nostromo* that he remembers and that he "could place my finger on fairly substantial passages" (Brice 2004: 79), and another 20,000 that he only faintly remembers and would find difficult to trace. Later, in *Return to Yesterday*, Ford himself minimizes his contribution, saying that what he "wrote into Conrad's books was by no means great in bulk" (Brice 2004: 78) and was "so frequently emended out of sight that they could not make as much difference to the completion and glory of his prose as three drops of water poured into a butt of Malmsey" (Brice 2004: 79). This study tries to find the drops of Ford's water in Conrad's Malmsey.

Figure 1 shows a bootstrap consensus tree of works and collaborations by Conrad and Ford, produced from multiple cluster analyses of the most frequent word frequencies for this collection of texts (Eder & Rybicki 2011). It places *Romance* decisively among works by Conrad, but the two other collaborative texts among those by Ford.

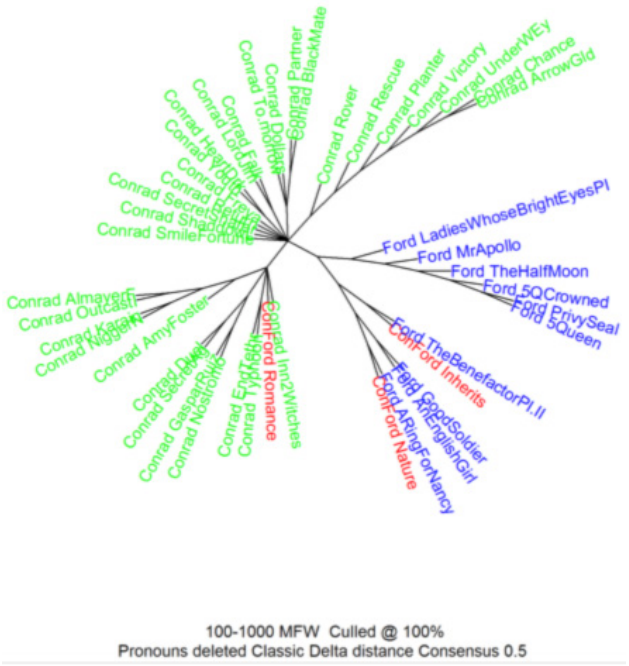


Figure 1.

Figure 2 is a 'Rolling Delta' diagram produced for *The Inheritors*, which is compared to four novels by Conrad (*The Nigger of the Narcissus* 1897, *Lord Jim* 1900, *Chance* 1913, and *Victory* 1915) and four by Ford (*The Fifth Queen* 1906, *Privy Seal* 1907, *The Fifth Queen Crowned* 1908, and *The Good Soldier* 1915). The analysis was performed for the 1,000 most frequent words appearing in all the texts with a 'rolling window' of 5000 words, stepping by 1,000. Throughout the collaborative novel, Ford's style (part that of *The Benefactor*, part that of *Privy Seal*) dominates over that of Conrad's — except for a short fragment that coincides very closely with Chapters 16 and 17 of *The Inheritors* (contained within vertical lines *a* and *b*); here, the greatest similarity is to *Chance* and *Heart of Darkness*. Interestingly, however, Conrad's earliest work among the four used for comparison, *The Nigger of the Narcissus*, is the outlier almost invariably in *The Inheritors*.

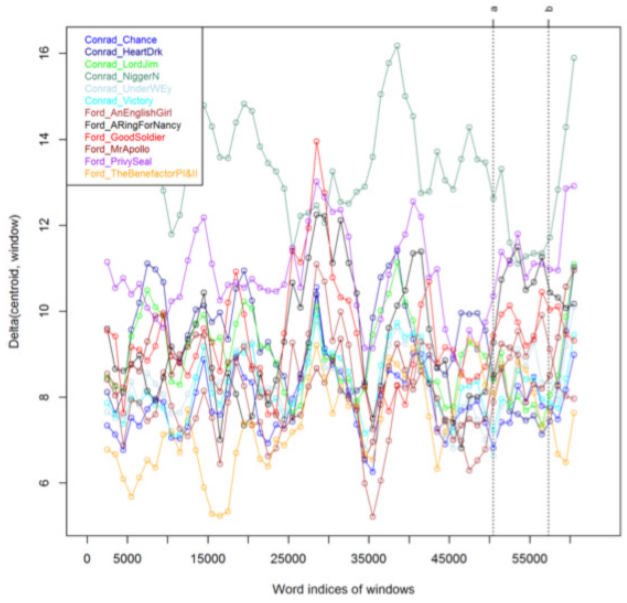


Figure 2

Figure 3 presents the same comparison for *Romance*. This novel, in good agreement with Figure 1, exhibits a domination of Conradian style (mostly that of *Lord Jim*, *Heart of Darkness* and *Chance*). Ford's idiom (this time, in its *The Benefactor* variety) makes itself seen in a single long fragment: Part 1, Chapter 5 (between lines *a* and *b*) and in a much shorter one (Part 2, Chapter 7: *c*).

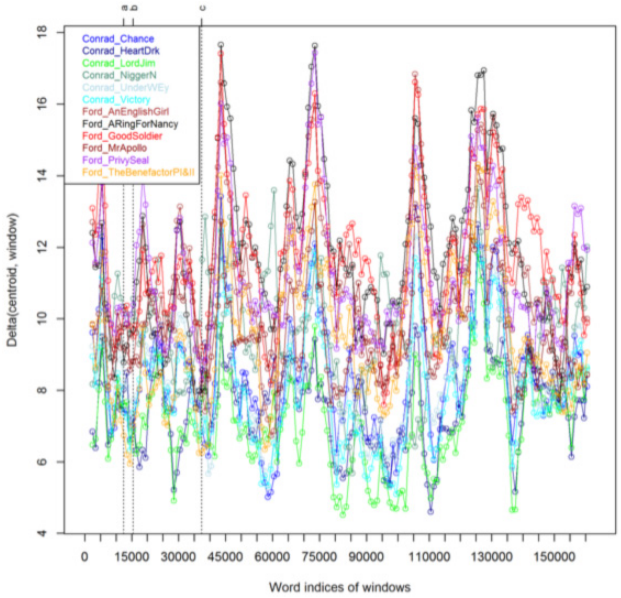


Figure 3

Figure 4 shows the same results for *The Nature of a Crime*, which strongly diverge from the previous collaboration. Ford's style (mostly that of *The Good Soldier*) dominates the final joint effort of the two writers, with two minor and non-decisive interventions of Conrad's style (esp. *Under Western Eyes*).

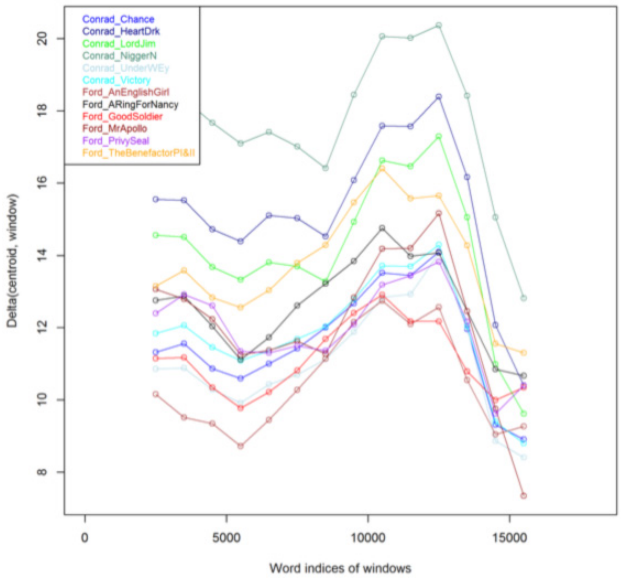


Figure 4

Figure 5 tests the hypothesis that Ford did indeed contribute to Conrad's *Nostromo*, but provides little to support it: Conrad's style dominates Ford's throughout.

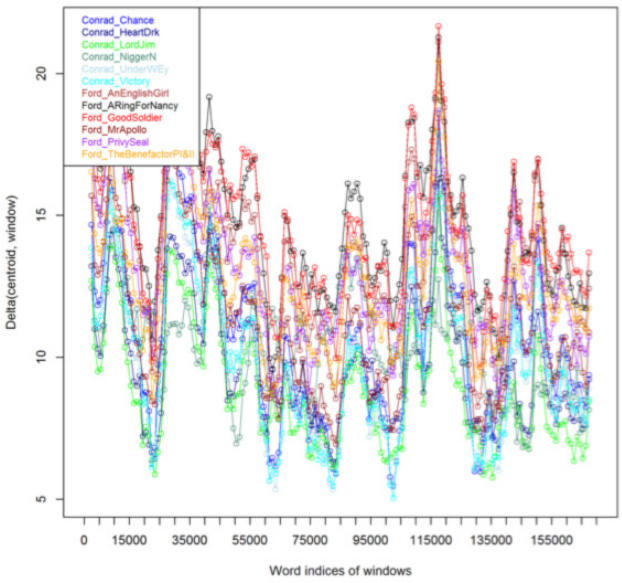


Figure 5

The application above of Rolling Delta produces interesting results. Chief among them is a confirmation of the usual (if uncertain) consensus about the proportions of the styles of both writers in their three collaborations. The decisive domination of Ford's style over Conrad's in *The Inheritors* and *The Nature of A Crime* is interesting, as it seems to have survived Conrad's extensive editing, confirmed by biographical evidence. A similar stylometric visibility of the underlying authorial personality that persists despite subsequent editing has been reported in a study of an edited translation (Heydel & Rybicki 2012). From a methodological point of view, "Rolling Delta" for *R* (devised by Kestemont) is a welcome addition to the latest stylometric tools, with its potential to pinpoint the change(s) from author to author in collaborative works. Further experiment are required to explain why, in some cases, the Rolling Delta curves tend to rise and fall in unison. It is possible that these fluctuations are simply indicative of minor stylistic fluctuations in the reference text, or, when these differences are larger, indeed its "originality" or rather its "deviation" with respect to the rest of the texts.

References

Argamon, S. (2008). Interpreting Burrows's Delta: Geometric and probabilistic foundations. *Literary and Linguistic Computing*. 23. 131-147.

Brice, X. (2004). Ford Madox Ford and the Composition of Nostromo. *The Conradian: the Journal of the Joseph Conrad Society* (Autumn) 75-95.

Burrows, J. (2002). 'Delta': A measure of stylistic difference and a guide to likely authorship. *Literary and Linguistic Computing*. 17. 267-287.

Burrows, J. (2010). Never Say Always Again: Reflections on the Numbers Game. In McCarty, W. (ed), *Text and Genre in Reconstruction. Effects of Digitalization on Ideas, Behaviours, Products and Institutions*. Cambridge: Open Book Publishers. 13-36.

Eder, M. and J. Rybicki (2011). Stylometry with R. In *Digital Humanities 2011: Conference abstracts*, Stanford University. 308-311.

Eder, M., M. Kestemont, and J. Rybicki (2012). *Computational Stylistics*. https://sites.google.com/site/computationalstylistics/

Heydel, M., and J. Rybicki (2012). The Stylometry of Collaborative Translation. Woolf's *Night and Day* in Polish. *Digital Humanities 2012 Conference Abstracts* 212-217.

Hoover, D. (2004). Testing Burrows's Delta. *Literary and Linguistic Computing* 19. 453-475.

Hoover, D. (2012). *The Tutor's Story*: A Case Study of Mixed Authorship. *English Studies* 93. 324-339.

Karl, F. (1997). *A Reader's Guide to Joseph Conrad*, Syracuse, NY: Syracuse University Press.

Najder, Z. (2007) *Joseph Conrad: A Life*. trans. by Najder, H. New York: Camden House.

Saunders, M. (1996). *Ford Madox Ford: A Dual Life*. 1. New York: Oxford University Press.

Van Dalen-Oskam, K., and J. Van Zundert (2007). Delta for Middle Dutch — Author and copyist distinction in *Walewein*. *Literary and Linguistic Computing*. 22. 345-362.

Van Dalen-Oskam, K., and J. Van Zundert (2012). Delta in 3D: Copyist distinction by Scaling Burrows's Delta. *Digital Humanities 2012 Conference Abstracts*. 402-404.

Simulating Plot: Towards a Generative Model of Narrative Structure

Sack, Graham Alexander
gas2117@columbia.edu
Columbia University, United States of America

Introduction

This paper explores the application of computer simulation techniques to the fields of literary studies and narratology by developing a model for plot structure and characterization. Using a corpus of 19th Century British novels as a case study, this author begins with a descriptive quantitative analysis of character names, developing a set of stylized facts about the way narratives allocate attention to their characters. I show that narrative attention in many novels appears to follow a "long tail" distribution. I then construct an explanatory model instantiated in a JAVA-based simulation, demonstrating that basic assumptions about plot structure are sufficient to generate output consistent with the real novels in the corpus.

This study differs from prior computational work in literary criticism in two crucial respects. First, rather than style, this paper is concerned with plot and characterization. As critic Franco Moretti has argued, plot is the crucial element that must be quantified if computational methods are to gain traction in mainstream literary criticism. Second, the overwhelming majority of prior computational studies in literary criticism have been *descriptive*—counting and classifying the surface features of a text. This study, however, is focused on *generative models*. Although I make use of descriptive analysis, the main intent is to motivate a computer simulation that I will show is sufficient to reproduce several key stylized facts about actual narratives.

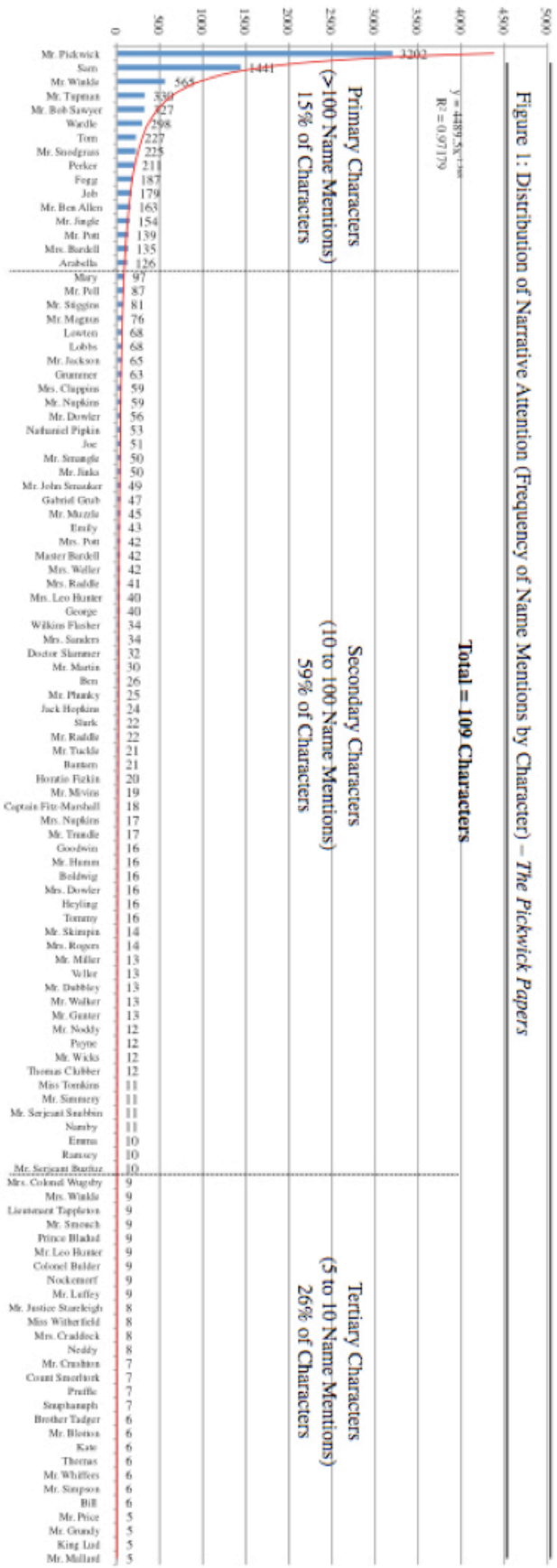
Part 1: Descriptive Analysis

In *The One vs. The Many* (2003), literary critic Alex Woloch repositions the questions of plot and characterization with which narratologists and formalists have traditionally been concerned in terms of the concept of "narrative attention." Woloch argues that "narrative attention" is a scarce resource that authors must choose how to allocate amongst the characters populating their stories.

Taking a cue from *The One vs. The Many*, this paper begins by applying quantitative rigor to the concepts of "distribution" and "apportioning of narrative attention," terms that Woloch uses qualitatively. By way of example, Figure 1 depicts the statistical distribution of character name mentions in Charles Dickens' *The Pickwick Papers*. The distribution of name mentions (an observable metric) can be used as an instrumental variable for the distribution of narrative attention (a latent, unobservable variable). The result is striking—109 characters organized into what one might term "the long tail": a small set of central characters represented by the spike on the left followed by a steep drop off to a long but shallow tail consisting of dozens of characters who are mentioned fewer than 10 times.

A wide range of phenomena are also known to follow a long tail: wealth distribution, website hits, and online books sales, for example, all obey a power law. The data for the novels sampled suggests that character name mentions and, by extension, narrative attention, are similarly distributed. That the distribution of attention within a novel should closely resemble the distribution of wealth within a nation

is a provocative fact that calls for explanation. Do narratives exhibit power-law behavior because of rich-get-richer effects such as preferential attachment or are character names merely a special case of Zipf's law regarding word frequency?



Part 2: Generative Models

Computer simulation techniques can play a valuable role in elucidating the dynamics driving narrative attention. This paper takes a structuralist approach that envisions narrative as composed of sub-structures with combinatorial rules. I assume that a plot structure is composed of a set of interwoven "plot strands" each with an internal hierarchy of characters. A JAVA-based simulation is used to implement this model. The user specifies the number of characters, plot strands, and scenes. The model then progresses sequentially through the plot, instantiating each strand as a scene in the predetermined order.

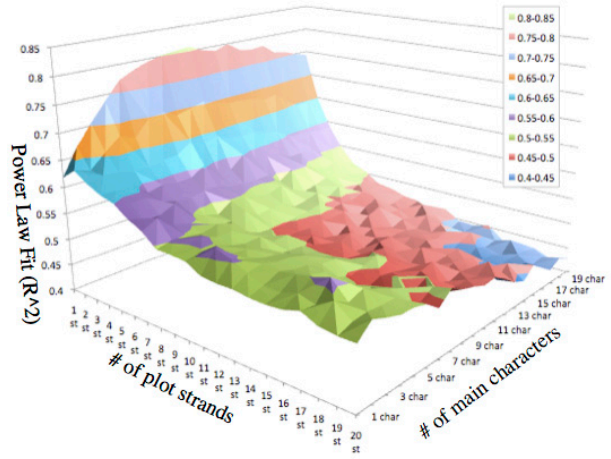
Although simplistic in its assumptions, this simulation is sufficient to reproduce a number of the salient features of narrative attention in the novels sampled. If the number of plot strands and main characters are set low—corresponding to a narrative that is tightly focused around one or a few characters in a single story line—the results closely resemble those observed for a *Bildungsroman* such as *Pride and Prejudice*. If the number of plot strands and main characters are set high—corresponding to a narrative focused around a large ensemble of characters across many subplots or parallel plots—the results closely resemble those observed for a sweeping social problem novel such as *Bleak House*.

Figure 2 shows a sweep of the model's output in parameter space. The z-axis is the average goodness of fit of a power-law distribution. The x-axis represents the number of main characters (from x = 1 to x = 20) and the y-axis the number of plot strands (from y = 1 to y = 20). The number of characters is held constant at 50 and the number of scenes is held constant at 30. The model is run 40 times for each (x,y) pair, for a total of 16,000 runs. As the graph shows, the distribution of narrative attention fits a power law well for a low number of plot strands. As the number of plot strands increases, the fit erodes, particularly if the number of characters is increased along with the strands.

The simulation developed is intentionally simplistic: I have modeled plot structure and characterization only in terms of combinatorial rules for plot strands. Nevertheless, this simple model of plot structure is sufficient to generate results consistent with the way narrative attention is allocated in actual novels. This suggests the explanatory power of the "strand"—an historically under-theorized narrative unit that is worthy is further investigation.

Figure 2: Parameter Sweep of Model Output

Constants: # of characters = 50; # of scenes = 30



References

Anderson, C. (2008). *The Long Tail*. New York: Hyperion.

Easley, D., and J. Kleinberg (2010). Power Laws and Rich Get Richer Phenomena. *Networks, Crowds, and Markets*. Cambridge UP.

Moretti, F. (2010). Network Theory, Plot Analysis. *New Left Review*.

Woloch, A. (2003). *The One vs. the Many*. Princeton, NJ: Princeton University Press.

Centre and Circumference: Modelling and Prototyping Digital Knowledge Environments as Social Sandboxes

Saklofske, Jon
jon.saklofske@acadiau.ca
Acadia University, Canada; INKE

This past year, the Modeling and Prototyping team within INKE (Implementing New Knowledge

Environments), a SSHRC Major Collaborative Research Initiative headed by Dr. Ray Siemens (UVic), has focused its research efforts on modeling the digital scholarly edition by producing a number of prototype editing environments. One such environment, NewRadial, is a reimagining of a 2009 Java-based data visualization prototype that was designed by Jon Saklofske and Jean-Marc Giffin for the purposes of reimagining the ways that scholars could work with William Blake's composite art in an electronic environment. As stated in articles published in the *Poetess Archive Journal* and *European Romantic Review*, the 2009 single-user version of NewRadial was intended to facilitate perceptual encounters with Blake's work that were not bound by the technology and formal constraints of the traditional book, and to provide the opportunity to centralize a user's secondary scholarship and commentary on Blake's work in a space that included the original archival material but did not directly alter that material. In an effort to adapt these ideas and arguments to INKE's focus on digital edition spaces, Saklofske and Jake Bruce have recreated NewRadial as browser-based software that makes use of an HTML5 frontend, an adapter system to ensure compatibility with a variety of database types, and a server-based backend. This new prototype has allowed INKE's Modeling and Prototyping team to engage with some of the larger questions that have motivate this year's digital scholarly edition research:

1. How do we model and enable context, such as prosopography and placeography, within the electronic scholarly edition?
2. How do we engage knowledge-building communities within the space of the electronic edition, and capture process, dialogue, and connections in and around such editions?

Material print editions are records, artifacts that efface the process of their formation, version-objects that assert an argument and establish a historical position through the printed finality of their collation and production. If digital editions are to take full advantage of their environments (rather than simply emulating print traditions) they need to visibly include both process and product, and offer opportunities for editorial diligence, contribution, perspective, control and debate to their users. Top-down forms of authoritative and exclusive editorial selectivity become ironic and anachronistic in dynamic digital environments which privilege "a new kind of scholarly discourse network that eschews traditional institutionally-reinforced hierarchical structures" (Siemens 2011).

NewRadial's collaborative space is a reimagining of the digital scholarly edition as a transparent workspace layer in which established primary objects from existing databases can be gathered, organized, correlated, annotated, and augmented by multiple users in a dynamic environment that also features centralised margins for secondary scholarship and debate. It performs Jerome McGann's idea that "the fundamentally dynamical character of the textual condition can be digitally realized: the dialectic of the field relations between the history of the text's transmission and the history of its reception" (para. 34).

The INKE NewRadial prototype is being designed as an effective way for knowledge communities to work with all types of media objects, to aggregate search results from multiple databases using meta-adapters, and to share RDF-based secondary scholarship and annotation data over HTTP for use in other tools and workspaces. Currently, our prototype installation has successfully used adapters to import NINES/ARC data, the Archbook image repository, Google image search results and other scholarly database holdings.

This INKE prototype re-presents database material in a sandbox environment, encouraging iterative experimentation, hosting methodological and interpretative debate and supporting innovative combinations and connections. These opportunities can serve as the raw processes, the activated complex from which more traditional scholarly print projects (collaborative or otherwise) can precipitate. In summary, NewRadial is a social edition space that encourages three types of work:

1. A simple search, sorting and manipulation of database objects in a visual field for the purposes of early scholarly inquiry and curiosity-based research.
2. Initial, raw and in-process commentary on connections and associations between database objects. Within the database's visual field, scholars can add comments on such correlations, thus starting conversations, discussions and debates relating to such ideas. These discussions are hosted and archived by the NewRadial server.
3. Larger edition projects in which a community is able to centralize and sort specific selections from a larger database. NewRadial can be used to construct these edition environments, browse such environments, and (if desired) encourage secondary scholarship to proliferate in and around such projects.

NewRadial is thus an example of Stuart Moulthrop's idea of "intervention," which is "a practical contribution to a media system (e.g., some product, tool or method) intended to challenge underlying assumptions or reveal new ways of proceeding" (212). Its affordances introduce a dynamic multiplicity of vision into what has traditionally been a reductive, oppositional and snail's pace process of inter-edition debate and evolution. The development of

this digital edition environment prototype is the first step towards creating inclusive editorial workspaces which draw from broad data foundations and which encourage knowledge-building communities to actively reimagine edition-building processes.

Neil Fraistat recognizes the importance and potential impact of digital textuality's massive addressability on the future of editing, and acknowledges the potential for interactivity that digital environments facilitate. Drawing on his own edition work, he suggests that digital editions should be interoperable, layered and modular, multimodal, dynamic, scalable, curatable, everted and sustainable (331-32). NewRadial prototypes these characteristics in a manner that is distinct from Fraistat's own work on the Shelley-Godwin Archive project, and in doing so, offers an alternative and comparative environment in which such ideas can be implemented, and participates (prototypically and argumentatively) in a larger, theoretical discussion about the nature of new knowledge environments and their impact on traditional ideas of editing.

If we agree with Alan Galey and Stan Ruecker that a prototype is, in essence, an argument (405), then one of NewRadial's main arguments arises from its reimagining the space of an edition as an inclusive, virtual, visual environment. It builds on Jan Holmevik's recognition that "visual/auditory modes of delivery of knowledge…. [allow] philosophy to divest itself of literate shackles, of narrow-minded thinking, and print-based archives where philosophy feels most at home" (10). Further, by promoting the idea of the social edition through its affordances, NewRadial prototypically engages with the possibility raised by Gregory Ulmer that "while the entire administrative superstructure of literate specialized knowledge will be translated into cyberspace, once there much of it will evaporate" (5). In 2003, Ulmer anticipated that "the practices that will replace specialized knowledge remain to be invented" (5). Jerome McGann reiterated this idea in 2006, suggesting that "at some point books and their technology will cease to be our encompassing informational environment; they will get incorporated into the digital network of artifacts and information" (McGann, para. 19). Ten years after Ulmer's observation and seven years after McGann's confirmation of this potential trajectory, in the midst of tools and applications that — for the most part — continue to sustain the dominant paradigms of literacy, NewRadial's incorporation of the networked inventiveness of INKE's collaborative practices into its own developmental situation answers Ulmer's invitation to invent new, digitally-situated practices of knowledge acquisition, creation and exchange.

In the particular context of the INKE project, NewRadial is becoming what it beholds, and is primarily emerging as a social edition prototype. However, beyond the specific implementations within the INKE frame, NewRadial raises broader questions and offers possible directions involving relational data models, scalable data browsing, and crowd-sourced descriptive frameworks in humanities research and scholarship. NewRadial's knowledge environment both models and prototypes alternative ways of working with and contributing to data-centric humanities research online. It offers a unique lens through which a diverse collection of digital humanities objects can be re-imagined, freely explored and iteratively prototyped.

References

Fraistat, N. (2012). Textual Addressability and the Future of Editing. *European Romantic Review*. 23 (3). 329-333.

Galey, A., and S. Ruecker (2010). How a Prototype Argues. *LLC* 25 (4). 405-424.

McGann, J. (2006). From Text to Work: Digital Tools and the Emergence of the Social Text. In *Romanticism on the Net.* 41-42. (2 November 2013.)

Moulthrop, S. (2005). After the Last Generation: Rethinking Scholarship in the Days of Serious Play. In *Proceedings of Digital Arts and Culture Conference.* held at IT-University. Copenhagen, Denmark. 208-215.

Saklofske, J. (2011). Remediating William Blake: Unbinding the Narrative Architectures of Blake's Songs. European Romantic Review 22 (3). 381-88.

Saklofske, J. (2010). NewRadial: Re-visualizing the Blake Archive. In *Poetess Archive Journal.* 2 (1).

Saklofske, J., and J. M. Giffin (2009). *NewRadial*. 2 November 2012. http://sourceforge.net/projects/newradial/

Saklofske, J. and J. Bruce (2012). *NewRadial* (INKE) . 2012. 2 November 2012. http://sourceforge.net/projects/newradial-inke/ and http://inke.acadiau.ca/newradial/

Siemens, R., et al. (2012). Toward Modelling the Social Edition: An Approach to Understanding the Electronic Scholarly Edition in the Context of New and Emerging Social Media. Accepted for publication in *Literary and Linguistic Computing.* 70.

Ulmer, G. (2003). *Internet Invention: From Literacy to Electracy.* New York: Longman.

Made to Make: Expanding Digital Humanities through Desktop Fabrication

Sayers, Jentery

jentery@uvic.ca
University of Victoria, Canada

Boggs, Jeremy
jeremy@clioweb.org
University of Virginia, USA

Elliott, Devon
devonelliott@rogers.com
Western University, Canada

Turkel, William J.
william.j.turkel@gmail.com
Western University, Canada

Contributing Labs: The research has been conducted at the Scholars' Lab (University of Virginia), the Humanistic Fabrication Lab (Western University), and the Maker Lab in the Humanities (University of Victoria).

Introduction

This paper presents substantive, cross-institutional research conducted on the relevance of desktop fabrication to digital humanities research. The researchers argue that **matter is a new medium for digital humanities**, and —as such—the field's practitioners need to develop the workflows, best practices, and infrastructure necessary to meaningfully engage digital/material convergence, especially as it concerns the creation, preservation, exhibition, and delivery of cultural heritage materials in 3D. Aside from sharing example workflows, best practices, and infrastructure strategies, the paper identifies several key growth areas for desktop fabrication in digital humanities contexts. Ultimately, it demonstrates how digital humanities is "made to make," or already well positioned to contribute significantly to desktop fabrication research.

Terminology

Desktop fabrication is the digitization of analog manufacturing techniques (Gershenfeld 2005). Comparable to desktop publishing, it affords **the output of digital content (e.g., 3D models) in physical form (e.g., plastic)**. It also personalizes production through accessible software and hardware, with more flexibility and rapidity than its analog predecessors. Common applications include using desktop 3D printers, milling machines, and laser cutters to prototype, replicate, and refashion solid objects.

Literature Review

To date, desktop fabrication has been used by historians to build exhibits (Elliott, MacDougall, and Turkel 2012); by digital media theorists to fashion custom tools (Ratto and Ree 2012); by scholars of teaching and learning to re-imagine the classroom (Meadows and Owens 2012); by archivists to model and preserve museum collections (Terdiman 2012); by designers to make physical interfaces and mechanical sculptures (Igoe 2007); and by well-known authors to "design" fiction as well as write it (Bleecker 2009; Sterling 2009). Yet, even in fields such as digital humanities, **very few non-STEM researchers know how desktop fabrication actually works**, and research on it is especially lacking in humanities departments across North America.

By extension, humanities publications on the topic are rare. For instance, **"desktop fabrication" never appears in the archives of *Digital Humanities Quarterly*** . The term and its methods have their legacies elsewhere, in STEM laboratories, research, and publications, with Neil Gershenfeld's *Fab: The Coming Revolution on Your Desktop* (2005) being one of the most referenced texts. Gershenfeld's key claim is that: "Personal fabrication will bring the programmability of digital worlds we've invented to the physical world we inhabit" (17).

This attention to digital/material convergence has prompted scholars such as Matt Ratto and Robert Ree (2012) to argue for: 1) **"physical network infrastructure"** that supports "novel spaces for fabrication" and educated decisions in a digital economy, 2) **"greater fluency with 3D digital content"** to increase competencies in digital/material convergence, and 3) an **established set of best practices**, especially as open-source objects are circulated online and re-appropriated.

To be sure, digital humanities practitioners are well equipped to actively engage all three of these issues. The field is known as a field of makers. Its practitioners are invested in knowing by doing, and they have been intimately involved in the development of infrastructure, best practices, and digital competencies (Balsamo 2009; Elliott, MacDougall, and Turkel 2012). They have also engaged digital technologies and content directly, as physical objects with material particulars (Kirschenbaum 2008; McPherson 2009). The key question, then, is **how to mobilize the histories and investments of digital humanities to significantly contribute to desktop fabrication research** and its role in cultural heritage.

Research Questions

To spark such contributions, the researchers are asking the following questions: 1) What are the best procedures for digitizing rare or obscure 3D objects? 2) What steps should be taken to verify the integrity of 3D models? 3) How should the source code for 3D objects be licensed? 4) Where should that source code be stored? 5) How are people responsible for the 3D objects they share online? 6) How and when should derivatives of 3D models be made? 7) How are fabricated objects best integrated into interactive exhibits of cultural heritage materials? 8) How are fabricated objects best used for humanities research? 9) What roles should galleries, libraries, archives, and museums (GLAM) play in these processes?

Findings

In response to these questions, the three most significant findings of the research are as follows:

I) WORKFLOW: Currently, there is no established workflow for fabrication research in digital humanities contexts, including those that focus on the creation, preservation, exhibition, and delivery of cultural heritage materials. Thus, the first and perhaps most obvious finding is that such **a workflow needs to be articulated, tested in several contexts, and shared with the community**. At this time, that workflow involves the following procedure: 1) Use a DSLR camera and a turntable to take at least twenty photographs of a stationary object. This process should be conducted in consultation with GLAM professionals, either on or off site. 2) Use software (e.g., 3D Catch) to stitch the images into a 3D scale model. 3) In consultation with GLAM professionals and domain experts, error-correct the model using appropriate software (e.g., Blender or Mudbox). What constitutes an "error" should be concretely defined and documented. 4) Output the model as an STL file. 5) Use printing software (e.g., ReplicatorG) to process STL into G-code. 6) Send G-code to a 3D printer for fabrication.

If the object is part of an interactive exhibit of cultural heritage materials, then: 7) Integrate the fabricated object into a circuit using appropriate sensors (e.g., touch and light), actuators (e.g., diodes and speakers), and shields (e.g., wifi and ethernet). 8) Write a sketch (e.g., in Processing) to execute intelligent behaviors through the circuit. 9) Test the build and document its behavior. 10) Refine the build for repeated interaction. 11) Use milling and laser-cutting techniques to enhance interaction through customized materials.

If the object and/or materials for the exhibit are being published online, then: 12) Consult with GLAM professionals and domain experts to address intellectual property, storage, and attribution issues, including whether the object can be published in whole or in part. 13) License all files appropriately, state whether derivatives are permitted, and provide adequate metadata (e.g., using Dublin Core). 14) Publish the STL file, G-code, circuit, sketch, documentation, and/or build process via a popular repository (e.g., at Thingiverse) and/or a GLAM/university domain.

When milling or laser-cutting machines are used as the primary manufacturing devices instead of 3D printers (see step 6 above), the workflow is remarkably similar.

II) INFRASTRUCTURE: In order to receive feedback on the relevance of fabrication to the preservation, discoverability, distribution, and interpretation of cultural heritage materials, humanities practitioners should actively consult with GLAM professionals. For instance, the researchers are currently collaborating with libraries at the following institutions: the University of Virginia, the University of Toronto, York University, Western University, McMaster University, the University of Washington, and the University of Victoria.

By extension, desktop fabrication research extends John Unsworth's (1999) premise of "the library as laboratory" into all GLAM institutions and suggests that new approaches to physical infrastructure may be necessary. Consequently, the second significant finding of this research is that **makerspaces should play a more prominent role in digital humanities research**, especially research involving the delivery of cultural heritage materials in 3D. Here, existing spaces that are peripheral or unrelated to digital humanities serve as persuasive models. These spaces include the Critical Making Lab at the University of Toronto and the Values in Design Lab at University of California, Irvine. Based on these examples, a makerspace for fabrication research in digital humanities would involve the following: 1) training in digital/material convergence, with an emphasis on praxis and tacit knowledge production, 2) a combination of digital and analog technologies, including milling, 3D-printing, scanning, and laser-cutting machines, 3) a flexible infrastructure, which would be open-source and sustainable, 4) an active partnership with a GLAM institution, and 5) research focusing on the role of desktop fabrication in the digital economy, with special attention to the best practices identified below.

III) BEST PRACTICES: Desktop fabrication, especially in the humanities, currently lacks articulated best practices in the following areas: 1) attribution and licensing of cultural heritage materials in 3D, 2) sharing and modifying source code involving cultural heritage materials, 3) delivering and fabricating component parts of cultural heritage materials, 4) digitizing and error-correcting 3D models of cultural artifacts, and 5) developing and sustaining desktop fabrication infrastructure.

This finding suggests that, in the future, **digital humanities practitioners have the opportunity to actively contribute to policy-making related to**

desktop fabrication, especially as collections of 3D materials (e.g., Europeana and Thingiverse) continue to grow alongside popular usage. Put differently: desktop fabrication is a disruptive technology. Governments, GLAM institutions, and universities have yet to determine its cultural implications. As such, this research is by necessity a matter of social importance and an opportunity for digital humanities to shape public knowledge.

Collation on the Web

Schmidt, Desmond
desmond.allan.schmidt@gmail.com
University of Queensland

Collation comes from the Latin *confero* (perfect participle *collatum*) meaning 'bring together'. There are several meanings in English, among them 'bring together for comparison … in order to ascertain points of agreement and difference' (OED, 2012). Even here collation may refer to a mechanical, manual or computerised process of comparing texts. My focus is on the latter, because it derives from an originally manual process as described, for example, by West (1973, 66f) and Dearing (1962, 14ff). Collation was a key part of the preparation of a critical edition because it supplied the raw differences between a chosen copy text and the other versions that aided the establishment of a single text suitable for printing.

Vinton Dearing in 1962 described what is perhaps the world's first collation program (1962, 18-19). It compared two texts, one line at a time, within a window of 10 lines in either direction. Once a line (or later a word) was matched in the two versions being compared, the window was moved on. This allowed it to recognise insertions, deletions, substitutions and transpositions over short distances. The window was used probably because memory on the IBM 7090 for which it was written, was limited. This basic design was then followed in all subsequent collation programs. For example the collation program of Froger (1968 234), 'EDIT' (Silva and Bellamy 1969, 41-25), 'OCCULT' (Petty and Gibson 1970), the collation program of Gilbert (1973), 'UNITE' (Marín 1991), 'PC-CASE' (Shillingsburg 1996, 144-148), 'TUSTEP-Collate' (1979), 'URICA!' (Cannon and Oakman 1989), 'DV-Coll' (Stringer and Vilberg 1987) and 'Collate' (Robinson 1989, 1994) all appear to use the same 'sliding window' technique. The size of the window varies, and in various programs extra features are added such as the ability to embed references, define transposed blocks and perform spelling normalisation (Collate), or the ability to merge collation output from each run (TUSTEP, PC-CASE).

One point often mentioned in these early collation programs is that they were developed to automate the manual process of producing a print edition. As Cannon explains: 'automatic collation should proceed as it would be performed manually' (1976, 33). Robinson also admits, when talking of the automatic treatment of variants that 'most electronic editions do the same as book editions: they just do more of it, perhaps with marginally more convenience' (2003).

The sliding window technique has come to define what automatic collation is, but it has some serious technical limitations. For example, it cannot see alignments of words outside of the window, and this makes it prone to mistakes, which must be manually corrected. However, modern computers have no need of a window, as they can easily load into memory the entire text for comparison.

One may also ask whether a print-based collation technique is really suited to a modern fluid medium like the Web. Differences discovered by a machine are not always suitable for display on a screen. As Robinson points out: 'Some differences will be just, well, noise: only a few ... are real variants, of real interest to real scholars.' (2009, 349). Hence all the early collation programs employ filtering, whether some kind of fuzzy matching, or a normalisation table to discount minor spelling variants. However, an apparatus generated in this way cannot subsequently be recombined with the base text to produce the faithful text of another version, because after filtering it contains only a tiny fraction of the true differences. It thus can only be attached to a base version as a series of notes, which limits the possibilities for display and interaction between user and text.

Another problem arises from the use of embedded markup. When SGML and then XML became popular from around 1990, there was a notable decrease of interest in collation programs. Existing programs were not updated, and replacements that fully handle XML have not yet emerged. The reason seems to be that if the computed differences between two texts contain disconnected start or end-tags, how does one supply the missing tags? In the case of an apparatus entry generated from TEI-XML such as: 'word`</hi>`', what is the format of 'word'? It could be anything because the start-tag and its attributes have been lost. So markup must be stripped out before collation can take place, as is done, for example, in Juxta Commons (2012). But stripping out markup is prone to error: how, for example, does one deal with embedded notes, and interpretations, or alternatives like 'sic' and 'corr' or embedded variants? (Schmidt 2012a). It also makes it difficult to compare formatting differences, and to later restore the markup, because the differences only refer to the stripped text.

A further problem is whether people really want to see a print apparatus on the screen. Although it may be

defended as a traditional form of variant display, what the modern user ultimately wants is interactivity. The essence of the modern Web is animation or the ability to edit and contribute in real time, not statically formatted data.

From collation to merging

What is needed for the medium of the Web is a thorough reassessment of the collation process. As a first step the difficulties in comparing embedded markup can be avoided by separating the text from its properties. 'Standoff properties' (Schmidt 2012a), which are modelled on LMNL (Piez, 2010), can be used in place of embedded XML, and may be generated from plain text or XML files. For each version this produces one version of the text and one or more markup files. The text and markup can then be merged separately, using the nmerge program (Schmidt 2009), into multi-version documents, which record the differences between all the versions globally – not merely between the base version and the rest. Because it doesn't use a sliding window, but looks for differences over the entire text, nmerge doesn't lose its way. The separately computed differences in the markup and the text are merged with the text's own structural properties and then formatted into HTML, without the need for XML. This new platform for digital editions facilitates various techniques for displaying variation (Figure 1). Each display is generated as a partial web-page so it can be incorporated into any kind of Web-delivery system:

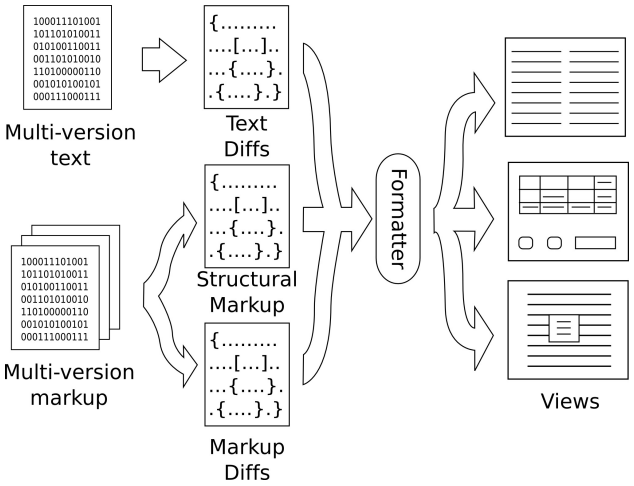


Figure 1:
Collation workflow using MVD+standoff properties

The most popular of these is the side-by-side display. Examples include MEDITE (Bourdaillet and Ganascia 2006), the MVD-GUI (Schmidt et al. 2008), Juxta Commons (2012), the Trein der Traagheid electronic edition (Roelens et al. 2012), the Versioning Machine (Schriebman 2011), etc. Some of these programs have synchronised scrolling, which helps keep compared versions in alignment. Side by side view is more suited to programs like MEDITE or nmerge that compute character-level differences as opposed to word-level differences, because the user can see at a glance how two similar words differ. And multi-version documents already contain all the differences between versions, which don't need to be recomputed each time, resulting in a much faster response, as can be seen in the AustESE (Australian electronic scholarly editions) test web interface (Schmidt 2012b).

Another popular type of variant display is the table, as found in CollateX (Dekker et al. 2011), and in the Cervantes hypertext edition (Urbina 2008). This is particularly useful in textual criticism because it presents much the same information as the old apparatus, but in a native digital form. In the AustESE test interface, table view (Figure 2) offers several options to reduce variant clutter without resorting to filtering. Character-level granularity can be easily extended to word-level, which is more useful for this type of display. Table view has the advantage over side-by-side in that it allows the user to explore the differences between a larger set of versions. Combining a horizontally scrolling table of variants with a synchronised vertically scrolling main text even produces a credible replacement for the print critical edition in digital form (Schmidt 2012b).

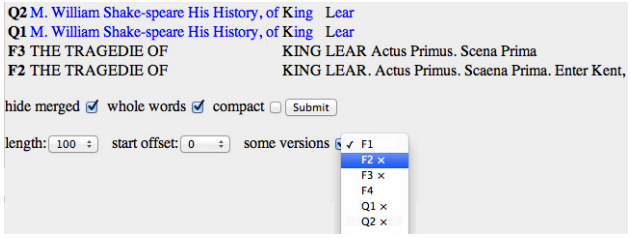


Figure 2:
Table view

Conclusion

The goal of collation on the Web is to provide the user with a variety of display options for exploring variation in a work. Collation conceived as a process for producing an apparatus or a filtered set of differences external to the text is too limited a technique to satisfy the flexible delivery options of the Web. Merging variant versions into a single digital object, on the other hand, provides a more efficient and direct way to query differences between versions, and to present the results through various views. Embedding markup into the text also creates problems for collation, and its removal allows differences between versions of text and markup to be merged as separate layers into the final

result. The medium of the Web thus offers more than just new ways to display old data. It challenges us to rethink fundamentally the way we create the modern edition.

References

Bourdaillet, J. and J. G. Ganascia (2006). MEDITE: A Unilingual Textual Aligner. In J. G. Carbonell and J. Siekmann (eds). *Lectures Notes in Artificial Intelligence*. 4139: 458-469.

Cannon, R. L. (1976). OPCOL: An Optimal text Collation Algorithm. Computers and the Humanities 10(1): 33-40.

Cannon, R. L., and R. L. Oakman (1989). Interactive Collation on a Microcomputer: The URICA! Approach. *Computers and the Humanities* 23: 469-472.

Dearing, V. A. (1962). Methods of Textual Editing, Los Angeles: William Andrews Clark Memorial Library, University of California.

Dekker, R. H., T. Andrews, B. Buitendijk, Z. Green, T.A. Griffitts, G. Middell, M. Mielnicki, L.-J. Olsson, T. Parkola, T. Vitt, and J. van Zundert (2011). CollateX. http://collatex.sourceforge.net (accessed 8 Oct, 2012).

Froger, D. J. (1968). *La critique des textes et son automatisation*. Paris: Dunod.

Gilbert, P. (1973). Automatic Collation: A Technique for Medieval Texts. *Computers and the Humanities* 7(3): 139-145.

Juxta Commons (2012). Juxta. http://www.juxtasoftware.org (accessed 8 Oct 2012).

Marín, F. (1991). Computers and Text Editing: A Review of Tools, an Introduction to UNITE and Some Observations Concerning its Application to Old Spanish Texts. *Romance Philology* 35: 102-122.

OED (2012). Oxford English Dictionary Online. Oxford: Oxford University Press.

Ott, W. (1979). A Text Processing System for the Preparation of Critical Editions. *Computers and the Humanities* 13: 29-35.

Petty, G. R., and W. M. Gibson (1970). *Project OCCULT: The Ordered Computer Collation of Unprepared Literary Texts.* New York: New York University Press.

Piez, W. (2010). 'Towards Hermeneutic Markup: An Architectural Outline'. *Digital Humanities Conference.* held July 7-10 2010 at Kings College London.

Roelens, X., R. Van den Branden, and E. Vanhoutte (2012). De trein der traagheid. http://edities.ctb.kantl.be/daisne/index.htm (accessed 8 Oct 2012).

Robinson, P. M. W. (1989). The Collation and Textual Criticism of Icelandic Manuscripts (1) Collation. *Literary and Linguistic Computing* 4(2): 99-105.

Robinson, P. M. W. (1994). *Collate 2: A User Guide*. Oxford: Oxford Computing Service.

Robinson, P. M. W. (2003). Where we are with electronic scholarly editions and where we want to be. *Computerphilologie* 5: 125-146. http://computerphilologie.tu-darmstadt.de/jg03/robinson.htm .

Robinson, P. M. W. (2009). Towards a Scholarly Editing System for the Next Decades. In: Huet, G., A. Kulkarni, and P. Scharf, (eds), Sanskrit Computational Linguistics 2007/2008. *LNCS*. 5402: 346–357.

Schmidt, D., D. Fiormonte, and N. Brocca (2008). A Multi-Version Wiki. In Opas-Hänninen, L.L., M. Jokelainen, I. Juuso, T. Seppänen (eds.), *Proceedings of Digital Humanities 2008* held June 2008 in Oulu, Finland. 187-188.

Schmidt, D. (2009). Merging Multi-Version Texts: a General Solution to the Overlap Problem, in *The Markup Conference 2009 Proceedings* held August in Montreal.

Schmidt, D. (2012a). *The Role of Markup in the Digital Humanities, Historical and Social Research/Historische Sozialforschung* 37(3): 125-146.

Schmidt, D. (2012b). http://austese.net/tests/ (accessed 8 Oct, 2012).

Schriebman, S. (2011). The Versioning Machine. http://v-machine.org (accessed 8 Oct 2012).

Shillingsburg, P. (1996). *Scholarly Editing in the Computer Age Theory and Practice.* Ann Arbor: University of Michigan Press.

Silva, G., C. Bellamy. (1968). *Some Procedures and Programs for Processing Language Data.* Clayton: Monash University.

Stringer, G., W. Vilberg. (1987). The Donne Variorum Textual Collation Program. *Computers and the Humanities.* 21(2): 83-89.

Urbina, E. (ed). (2008). Electronic variorum edition of the Quixote. http://cervantes.tamu.edu/V2/CPI/variorum/index.htm (accessed 8 Oct, 2012).

West, M. L. (1973). *Textual Criticism and Editorial Technique.* Stuttgart: B.G. Teubner.

Text to Image Linking Tool (TILT)

Schmidt, Desmond

desmond.allan.schmidt@gmail.com
University of Queensland, Australia

The digital edition has long been conceived not only as a set of transcriptions of the witnesses that underlie a work, but also as the images of the manuscript and book pages that physically represent it. Showing a readable transcription

next to the facsimile of its source document provides the reader with the same level of evidence enjoyed by the editor of a print edition.

Connecting ranges of text to areas in the image is useful wherever a manuscript has been heavily corrected or was written long ago in a writing style now hard to decipher. Even for printed works, the absence of such links forces the reader to waste time scanning the facsimile whenever the layouts differ, spelling variations have been regularised, or the text emended (Kiernan, 2006). Establishing text-image links was first explored in the electronic Beowulf edition (Dektyar et al., 2004). Other examples include the British Museum's electronic Codex Sinaticus (2009) and, using a different display technique, the 'zoom topographic' view of the Samuel Beckett digital manuscript project (Van Hulle et al., 2011). But the technical requirements of providing this form of interaction for any digital edition are considerable. There are three main problems:

1. How to display text-image links at the line and word-level over the Web
2. How to edit and automate the creation of links
3. How to store them in a reusable and efficient form

The following sections describe solutions to these problems and how they can be combined into a single system for storing, displaying and editing text to image links. This solution forms part of the AustESE (Australian electronic scholarly editing) project (Osborne, 2012), which aims to create interoperable Web-based tools for managing and creating digital editions.

1. Viewing links

As a solution to problem 1 a background image (Figure 1c) is overlaid first by a transparent pane or canvas on which drawing is done (b), and second by a set of invisible areas (a) that trigger simultaneous highlighting in the text and image as the mouse moves over them (Schmidt, 2012c). Polygonal areas are needed, since in many cases, e.g. curved or slanted lines, corrections, inserted text-blocks etc. rectangles are simply too imprecise. It is also versioned: the highlighted areas change according to the version of the text (base or corrected) chosen on the right.

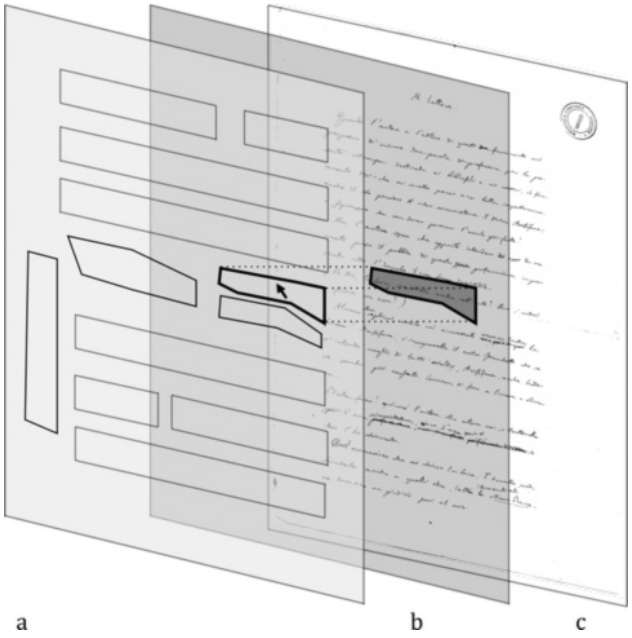


Figure 1: Web visualisation of image areas

In an alternative technique described by Cayless (2008), images of the text are converted to SVG surrogates (an XML format), and their bounding rectangles computed, to facilitate linking with the XML-encoded text. However, this increases the overall complexity of the encoding, and invites overlap between existing markup and the newly added spans. The solution described here, on the other hand, requires only simple features of HTML5, and is more amenable to automation.

2. Creating and editing links

There have been several tools already developed for creating and updating links from areas in a facsimile to spans in a transcription. This operation should be distinguished from more general kinds of image annotation, because text-image links are far more numerous, more specialised and, as Kiernan notes (2006b), likely to overwhelm any standard annotation system.

There appear to be only three tools that meet this narrow definition. The first is EPPT, created for the electronic Beowulf edition (Iacob and Kiernan, 2009) as part of the ArchWay project (Kiernan et al, 2005). This was an Eclipse plug-in (a programmer's tool) that created simple shapes without automation. The Textgrid TBLE tool is likewise an Eclipse plug-in (Al-Hajj, 2011). Although it does not yet support automation, it can draw polygons. However, having the editing tool available over the Web would facilitate updating through scholarly collaboration and crowd-sourcing (e.g. ANL, 2012). TILE (Reside et al., 2009) is so

far the only Web-based solution to the text-to-image editing problem, but it remains incomplete.

Our need was for an immediate solution for both print and manuscript texts so we could begin entering data for projects like the Charles Harpur archive (SETIS). This is a collection of large manuscript anthologies (100-350 pages) and newspaper cuttings. Thus, to be practical, the text to image linking process would have to be largely automated. Our program was designed as an applet (a Java web plug-in) because this allows for rapid and stable development, support for image analysis, and Web presentation. It is called TILT (text-image linking tool) because it is designed to deal with moderately curved or slanting text (Schmidt 2012b). This is frequently found in facsimilies due to the difficulty of laying books or manuscripts absolutely flat, and also in manuscripts where the text is written in crooked lines or at an angle. TILT uses automation at several levels:

1. Auto-recognition of pages at the word or line-level
2. Auto-outlining of individual lines in response to a click or tap
3. Auto-outlining of individual words similarly
4. Manual drawing of polygons and rectangles
5. Automatic linking of the current shape to the text

The idea is that when one technique fails another can be used with finer control, and with manual adjustment in case of error.

Word recognition

When the user clicks on a word, a small square is assessed around the click-point (Figure 2). If the darkness of this square exceeds the average document darkness it is accepted (solid outlines, left). Adjacent squares are assessed similarly. If a square fails the test it is subdivided into four. If any one of these smaller squares is accepted then adjacent full-size squares are assessed as before. Otherwise it is discarded (dotted outlines, left). This leads to a rapid determination of word-boundaries. The basic outline is then converted into a bounding box (solid outline, right), or if this is too wasteful, into a polygon (dotted outline, right).

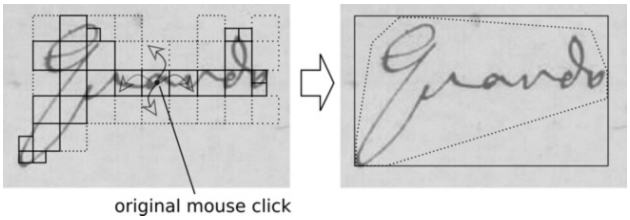


Figure 2: Detection of word boundaries

Line recognition

To facilitate line-recognition, the facsimile, which is often in colour, is first reduced to grayscale, then to pure black and white by application of a localised contrast filter, which erases differences caused by uneven lighting.

Baselines of the text can be found by summing the pixels in each horizontal row of the image. Rows with peak pixel-counts exceeding the average background darkness may be interpreted as baselines. These can be further constrained to lie within the text block, and only where there is surrounding text. The user can see the detected baselines by clicking the Lines tool. To detect entire lines the word-detection algorithm is simply applied along each baseline. To detect slanted or wavy baselines the image could be vertically divided into strips, and the lines in each strip joined up, although this is not yet implemented.

Linking word-shapes to the text

Word-shapes detected in the image are allocated approximately to corresponding word-positions in the text. Although breaking up the text into lines will make this more accurate, manually specifying positions where words and shapes do in fact match will allow this approximation to be gradually refined. Thus it is expected that, in practice, little human intervention will be needed to connect all word shapes to their correct words. For well laid-out facsimilies, e.g. of printed books, whole page recognition may be sufficient to link everything in one go.

3. Storing links

There have been several proposals for recording the data of text-image links. The model described by Audenart and Furuta (2009) focuses on the images rather than the work, whereas for our needs the latter is primary. The markup schemes described by Dekhtyar et al. (2005), and in TEI P5 (22.3) are based on embedded XML. As Kiernan (2006b) notes, text to image links, if embedded in the text, usually overlap with any markup already present. The resulting mixture of milestones and ordinary tags can be very complex.

On the other hand, simply using standoff properties (Schmidt 2012a), which are based on LMNL (Piez, 2010), to record the links, would simplify things considerably. Although saving is not yet implemented in TILT, each link, consisting of a shape and a textual span, could be saved as a separate layer. Links could then be merged with the other markup in the text, or used to generate the map which overlays the image (Figure 1a).

Conclusions and future work

The most important goal of TILT will be the testing of user requirements, which can only happen once a practical tool has been built. For example, how to deal with line-breaks in the middle of words, and what level of granularity (word or line) will work best must still be determined. TILT still lacks essential features like zooming, saving, the ability to change images as the user scrolls through the text, and export to other formats. However, based on current progress it is anticipated that these features will be completed by July 2013. There seems to be a pent-up demand for a text-image linking tool like this, given the current interest in digital facsimiles and the general exploration of more interactive interfaces for digital editions.

References

Al-Hajj, Y. A. A., and M. W. Küster (2011). The Text-Image-Link-Editor: A tool for Linking Facsimiles & Transcriptions and Image Annotations. *Proceedings of Digital Humanities* http://dh2011abstracts.stanford.edu/xtf/view?docId=tei/ab-197.xml;query=Al%20Hajj;brand=default (accessed 25 October 2012).

ANL, (2012). Trove, Australian National Library, http://trove.nla.gov.au (accessed 19 October 2012)

Audenart, N., and R. Furuta (2009). Annotated Facsimile Editions: Defining macro-level structure for image-based electronic editions. *Literary and Linguistic Computing*, 24.2 pp. 143-151.

Cayless, H. A. (2008). Linking Page Images to Transcriptions with SVG. In *Proceedings of Balisage: The Markup Conference.* held 12-15 August in Montréal, Canada. Balisage Series on Markup Technologies 1, doi:10.4242/BalisageVol1.Cayless01.

Codex Sinaticus, (2009). http://codexsinaiticus.org (accessed 25 October 2012).

Dekhtyar, A., I. E. Iacob, J. W. Jaromczyk, K. Kiernan, N. Moore, and D. C. Porter (2005). Support for XML Markup of Image-based Electronic Editions *International Journal on Digital Libraries.*

Iacob, E. and K. Kiernan (2009). Installing the EPPT-Trial. http://beowulf.engl.uky.edu/~eft/eppt-trial/EPPT-Install.htm (accessed 19 October 2012).

Kiernan, K. (2006a) Electronic Textual Editing: Digital Facsimiles in Editing. In Burnard, L., O'Brien, K., O'Keeffee and Unsworth, J. (eds), *Electronic Textual Editing*. New York: Modern Language Association of America.

Kiernan, K. (2006b). Technology. http://beowulf.engl.uky.edu/~kiernan/eBoethius/tech.htm#tech (accessed 25 October 2012).

Kiernan, K., J. W. Jaromczyk, A. Dekhtyar, D. C. Porter, K. Hawley, S. Bodapati, and I. E. Iacob (2005). The ARCHway Project: Architecture for Research in Computing for Humanities through Research, Teaching, and Learning. *Literary and Linguistic Computing* 20 (Suppl) pp. 69-88.

Osborne, R. http://austese.wordpress.com AustESE: Australian Electronic Scholarly Editing (accessed 25 October 2012).

Piez, W. (2010). Towards Hermeneutic Markup: An Architectural Outline. Digital Humanities Conference. held 7-10 July at Kings College. London.

Porter, D. (2009). The Text Image Linking Environment. Digital Humanities Summer Institute held at the University of Victoria. http://www.youtube.com?v=BiiNfDLqs6I (accessed 25 October 2012).

Reside, D., D. Lester, D. Porter, and J. Walsh (2011). tile–text image linking environment. http://mith.umd.edu/tile/ (accessed 25 October 2012).

Schmidt, D. (2012a). The Role of Markup in the Digital Humanities. *Historical and Social Research/Historische Sozialforschung* 37.3 pp. 125-146.

Schmidt, D. (2012b). TILT http://austese.net/tests/tilt (accessed 25 October 2012).

Schmidt, D. (2012c). Image. http://austese.net/tests/image (accessed 25 October 2012).

SETIS, The Sydney Electronic Text and Image Service. Harpur. http://setis.library.usyd.edu.au/ozedits/harpur/ (accessed 26 October 2012).

Guidelines for Electronic Text Encoding and Interchange. TEI P5, TEI Consortium(eds). http://www.tei-c.org/P5/ (last accesed 26 October 2012).

Van Hulle, D., M. Nixon, and V. Neyt (2011). Samuel Beckett Digital Manuscript Project. http://www.beckttarchive.org/demo/ (accessed 31 October 2012).

Fine-tuning Stylometric Tools: Investigating Authorship and Genre in French Classical Theater

Schöch, Christof
christof.schoech@uni-wuerzburg.de
University of Würzburg, Germany

This paper is concerned with stylometric classification applied to French seventeenth-century plays. It reports on ongoing investigations into parameter setting and its impact on classification of such texts by author, genre, or form, using Eder & Rybicky's stylometric scripts for R (Eder & Rybicky 2011). Based on an investigation into the Corneille-Molière controversy, several methodological issues standing in the way of reliable results have been identified. One issue concerns the degree to which such authorship classification tasks are influenced by genre (here, comedy or tragedy) and form (here, verse or prose). Investigation of this issue shows that input parameters have indeed effects on the relative influence of authorship, genre and form in the classification of plays.

Stylometry today: advances and challenges

Stylometry has made significant advances in recent years, due no doubt to the increased availability of electronic texts, of sophisticated and accessible stylometric tools, and of proposed classification methods and distance measures. Based on this range of resources, researchers in stylometry are able to use various linguistic features as input for classification tasks and may adjust a wide range of parameters.

This situation, however, also brings renewed urgency to the issue of fine-tuning input parameters and distance measures, depending on the materials under scrutiny and the type of inquiry. Arguably, this is somewhat less of an issue today for a language such as English, where a well-established stylometric tradition exists. However, despite recent advances for some languages (Van Dalen-Oskam & Van Zundert 2007, Rybicky & Eder 2011), parameter setting remains an insufficiently explored issue for languages such as German, French, Spanish, or Latin, and many more.

The Corneille-Molière controversy

This has been particularly apparent in the domain of seventeenth-century French drama, because work in this area has recently fuelled a controversy over whether or not Corneille was in fact the author of some or several plays traditionally attributed to Molière. In this controversy, traditional biographical and archival research (Boissier 2004) was complemented with results from stylometric analyses (Labbé & Labbé 2001; for a knowledgeable critique, see Brunet 2004; for a more recent approach, see Marusenko & Rodionova 2010). However, the methodological basis for stylometric analyses of this type of material seems to have been insufficiently investigated..

The conditions for reliable stylometric attribution results in this domain are challenging. The strong codification of classical literary discourse and the prevalence of stringent metrical forms mean that stylistic differences between authors are often subtle. At the same time, the available plays vary widely as to dramatic genre (e.g. comedy or tragedy) and form (i.e. verse or prose). Preliminary investigations into the Corneille-Molière corpus using Eder and Rybicky's stylometric scripts have indeed shown the fragility of the results. Depending on the composition of the text collection, on the linguistic material used as input, and on distance calculation measures, results vary widely.

Fine-tuning for author, genre and form

On the one hand, then, it can be challenging to make clear author attributions on material that is heterogeneous as to genre or form. For example, relevant research relying on the „unmasking“ technique showed unsatisfactory results for cross-genre authorship attribution (Kestemont et al. 2012). On the other hand, if only relatively homogeneous material is taken into account, the overall amount of data available for classification may be significantly reduced. What is needed is knowledge about how to limit the influence of factors other than the one of concern in any given classification task. The most relevant factors in the present case are authorship, genre (here, comedy or tragedy) and form (here, verse or prose). If the goal is to make reliable author attributions, how can the influence of genre and form be limited? The research reported on here was designed in order to explore such issues. All investigations are based on the *Théatre classique* collection (Fièvre 2007-2013), which provides XML/TEI versions of all plays. Texts were uniformely preprocessed to retain only character speeches, but no lemmatization was applied.

A first collection of plays was investigated limiting the number of relevant categories to just two, authorship and genre, and balancing the number of plays for each category. This resulted in a collection of 32 plays by Pierre Corneille and Thomas Corneille, with an equal number of comedies and tragedies by each author. The question at hand was to find out at which settings the classification would be dominated by either one of the author or genre category, and to what extent. Systematic variation was introduced as to the range of words from the frequency list taken into account: All runs relied on 100 words from the word frequency list, and each run took these from a moving onset point onwards, at an interval of 50 words.

For each run, the data was subjected to a distance measurement using Burrows' Delta (Burrows 2002), the results forming the basis of a cluster analysis. The distance

tables were saved for each run, and the proportions of the different low-level pairs for each run were extracted. The low-level pairs found can be of the following types: author-and-genre match, author-only match, genre-only match, or pairs without a match. The proportions of author-only and genre-only matches is assumed to indicate to which extent the chosen settings give precedence to textual features associated with authorship or genre, respectively. Figure 1 visualizes the results from this investigation.

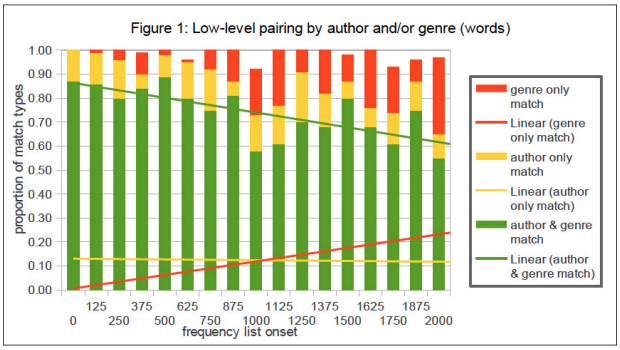


Figure 1

Author-and-genre matches decrease overall with increasing onset points, and the proportion of author-only matches remains relatively stable. However, the proportion of genre-only matches increases markedly with increasing onset points. In the range of onsets points between 0 and 1150 words (with two exceptions at 300 and 750 words), pairing is predominantly related to authorship, not genre. In the range of onset points from 1150 to 1650 words, pairing is related both to authorship and to genre, in varying proportions, while genre seems to be taking over more markedly beyond an onset point of 1650 words.

A similar investigation was run with a collection of plays with variation only as to authorship and form (verse or prose). The collection of plays consisted of 28 comedies, with an equal number of prose and verse plays by each of the following authors: Dufresny, Scudéry, Regnard, and Molière. Again, the nature and proportions of the different low-level pairs was assessed. With adjustment for the asymmetrical number of authors and forms, the graph shown here as figure 2 was constructed.

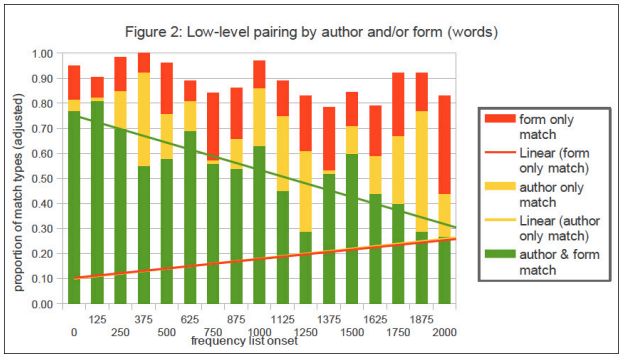


Figure 2

There is a very strong fall-off for author-and-form matches from an onset point of around 800 words. Although author-only and form-only matches increase somewhat over the range of onset points, this does not correspond to the decrease in author-and-form matches. The most important result is that over the entire frequency range, form-only matches are always present and in many though not all cases, they have a higher proportion than the author-only matches. Compared to the authorship vs. genre comparison (figure 1), there is certainly no clear cut-off point below or above which author-only matches would dominate form-only matches.

Conclusions

Despite their limitations, these preliminary results give some useful indications for authorship attribution studies in French classical verse drama, and may increase reliability of attributions. First, text collections of mixed dramatic sub-genre may be used in authorship classification tasks, provided that the wordlist used does not exceed the first 1150 most frequent words, so that influence from features related to genre remains limited. Second, form is a prevalent factor in the entire range of the frequency list, and should be controlled for when creating text collections. Applying these insights to the Molière-Corneille problem permits to enlarge the corpus of comparison texts beyond comedies, thus yielding a broader basis for classification tasks, but not beyond verse plays. While the procedure described here could be used for other languages and genre pairs, the results may be difficult to generalize: the best distinguishing parameters will likely be different from the ones found here for French classical drama.

However, more work needs to be done before the results obtained are sufficiently reliable. On the one hand, the approach taken here could be improved by enhancing the assessment of the dendrograms to take higher-level groupings into account as well. On the other hand, the fact that there are quite a few exceptions to overall trends shows

the limit of this approach. In fact, a mechanism like feature selection may be more appropriate to solve the issue. Using supervised machine learning techniques with authorship, genre or form as separate target classes, and combining this with information gain analysis for each target class, would allow generating lists of features relevant for each target category.

Acknowledgements

I would like to thank Jan Rybicky and Maciej Eder for introducing me to their tools as well as João Guerra for helping me with some Python coding.

References

Boissier, D. (2004). L'affaire Molière. La grande supercherie littéraire, Jean-Cyrille Godefroy.

Brunet, É. (2004). Où l'on mesure la distance entre les distances. *Texto!*. (4) http://www.revue-texto.net/Inedits/Brunet/Brunet_Distance.html (accessed 10 March 2013).

Burrows, J. (2002). 'Delta': a Measure of Stylistic Difference and a Guide to Likely Authorship. LLC 17 (3) 267-287. 10.1093/llc/17.3.267

Eder, M., and J. Rybicki (2011). Stylometry with R. In *DH2011: Conference Abstracts*. Stanford University, Stanford, 308-11.

Fièvre, P., (ed.) (2007-2013). *Théâtre classique*, http://www.theatre-classique.fr/ (accessed 10 March 2013).

Kestemont, M., K. Luyckx, W. Daelemans, and T. Crombez (2012). Cross-Genre Authorship Verification Using Unmasking. *English Studies*

Labbé, C., and D. Labbé (2001). Inter-textual Distance and Autorship Attribution Corneille and Molière. Journal of Quantitative Linguistics. 8(3): 213-231. 10.1076/jqul.8.3.213.4100

Marusenko, M., and E. Rodionova (2010). Mathematical Methods for Attributing Literary Works When Solving the 'Corneille-Molière' Problem. *Journal of Quantitative Linguistics* 17(1): 30-54.

Rybicki, J., and M. Eder (2011). Deeper Delta Across Genres and Languages: Do We Really Need the Most Frequent Words? *LLC*. 26(3): 315-321. 10.1093/llc/fqr031.

Van Dalen-Oskam, K., and J. van Zundert (2007). Delta for Middle Dutch. Author and Copyist Distinction in Walewein. *LLC*. 22(3): 345-362. 10.1093/llc/fqm012.

Beyond Infrastructure: Modelling Scholarly Research and Collaboration

Schreibman, Susan
susan.schreibman@gmail.com
Trinity College Dublin

Gradmann, Stefan
stefan.gradmann@ibi.hu-berlin.de
Humboldt University Berlin

Hennicke, Steffen
steffen.hennicke@staff.hu-berlin.de
Humboldt University Berlin

Blanke, Tobias
tobias.blanke@kcl.ac.uk
Kings College London

Chambers, Sally
sally.chambers@phil.uni-goettingen.de
Göttingen University

Dunning, Alastair
Alastair.Dunning@kb.nl
The European Library

Gray, Jonathan
jonathan.gray@okfn.org
Open Knowledge Foundation

Lauer, Gerhard
Gerhard.Lauer@phil.uni-goettingen.de
Göttingen University

Pichler, Alois
Alois.Pichler@fof.uib.no
University of Bergen

Renn, Jürgen
renn@mpiwg-berlin.mpg.de
Max Planck Institute for the History of Science

Morbidoni, Christian
christian.morbidoni@gmail.com
Net7

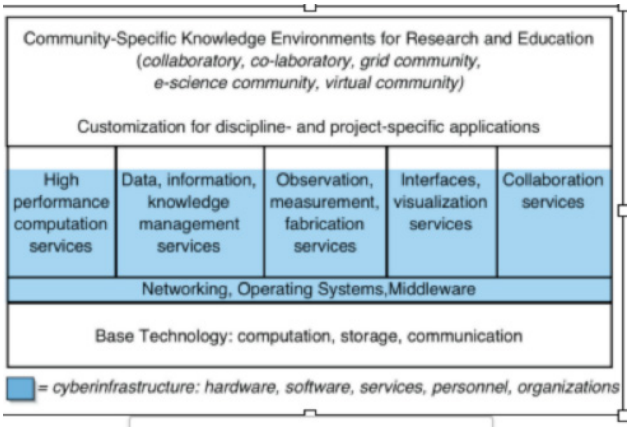
Romary, Laurent
laurent.romary@inria.fr
INRIA

Sasaki, Felix
felix.sasaki@dfki.de
Language Technology Lab

Warwick, Claire
c.warwick@ucl.ac.uk
University College London

Over the past decade considerable research has been carried out into creating infrastructure to support digital scholarship — from the "Atkins report" (Atkins et. al. 2003) commissioned by the National Science Foundation, to the more specific humanities/social science focused "Our Cultural Commonwealth" (Unsworth et. al. 2006), to the funding of large community-based infrastructure projects such as the Mellon-funded Bamboo and EU-supported DARIAH (Blanke et. al., 2011b).

The Atkins report introduced the following layered vision of the way technical research infrastructures are related to each other:



This "mother of all eScience layer cakes" introduced the hitherto canonical division between the blue area of supporting Cyberinfrastructure and the white area of discipline-specific applications. Most initiatives following the Atkins report were to focus more or less exclusively on the cyberinfrastructure layer.

The model of thought introduced by the two American-commissioned reports (Atkins and Unsworth) has been adopted in Europe, starting with the e-Science initiative in the UK that focused on the use of Grid technology (and which evolved in parallel with the NSF activity), and the German D-Grid initiative.

But it has become clear, however, that the focus on building infrastructure, while essential to support digital humanities scholarship, needs to be accompanied by a concomitant methodological emphasis. Rockwell (2010) pointed this out in the section of his contribution to "Dangers of Infrastructure"; i.e. that in building infrastructure we need to be aware of two major pitfalls:

- **Research infrastructure is not research** just as roads are not economic activity. We tend to forget when confronted by large infrastructure projects that they are not an end in themselves. [...].
- **Infrastructure projects can become ends in themselves** by developing into an industry that promotes continued investment. To sustain infrastructure there develops a class of people whose jobs are tied to infrastructure investment.

This paper will thus explore what is needed to foster an acceptance of digital practices in the humanities beyond the creation of pure infrastructure, specifically in terms of understanding and technically modelling traditional scholarly research within a digital medium while enabling new modes of scholarly work that could only be carried out within a digitally-mediated environment.

In the latter case, this means moving beyond the emulation of traditional methods of scholarship tied to the page (albeit with linking metaphors as in the first generation document-centric WWW), to new ways anchored in the web of Linked Data which might be viewed as a combination of notebook and Memex proposed by Schraefel (2007), who is, of course, channelling Bush (1949).

In order to model this, we need to better understand how scholars undertake their research now and in the past, and how their functional framework might adequately translate into a digital context in order to attract them to new working modes. Furthermore, this kind of activity needs to be an integral part of a research infrastructure, otherwise the infrastructure runs the risk of becoming a static environment rather than a dynamically evolving one that corresponds to ongoing and dynamic research needs.

John Unsworth (2000) conceptualized "scholarly primitives" as basic functions which are common to any scholarly activity in the humanities independent of discipline, theoretical orientation, or era. He suggested seven recursive and interrelated scholarly primitives — discovering, annotating, comparing, referring, sampling,

illustrating, and representing — which he saw as the basis for tool-building enterprises for the Digital Humanities. Since then, Unsworth's scholarly primitives have been often utilized and further revised.

As John Unsworth (2011) acknowledged in an interview almost a decade later, his list of scholarly primitives is not definitive. Subsequent research shows that there is no agreement on the exact definition or scope of scholarly primitives. However, the approach of using scholarly primitives or similar concepts proved to be a valuable and accepted means of structuring and conceptualizing the scholarly domain or aspects of it. Therefore we decided to use Unsworth's conceptualization of scholarly primitives as a starting point for our own Scholarly Domain Model. In our model, however, the scholarly primitives represent some of the most generic humanistic functions which are further broken down into more granular sub-functions which resemble scholarly activities.

Our research is part of web of scholarship currently being carried out within research infrastructure projects to link researchers' processes closer to the development of services and techniques (examples include Europeana Research Cloud and DARIAH's VCC2). Our contribution is a systematic investigation into how we can model primary research activities embracing the assumption that understanding what John Unsworth had originally proposed in terms of "scholarly primitives" more than a decade ago is central to any such approach at modelling the digital scholarly domain.

This paper will examine how deeper modelling of research processes in the humanities could inform the development of tools to enhance and augment scholarship. In particular we will focus on models of how students and scholars conduct research can be used to inform tool development, particularly in the area of text-based scholarship (both of primary texts and metadata), focusing primarily on transcription, translation, annotation and curation.

Furthermore, our models will be enriched by ontological models which enable scholarly functions. Therefore, the aim is not only to provide a framework for categorizing and assessing tools for the Digital Humanities but also to formalize the model into a computational model in order to capture research activity and thereby also validate our Scholarly Domain Model.

Our Scholarly Domain model will go beyond *categorizing* tools to create a formal model of interrelated research primitives and functions in order to implement operational scenarios in DM2E (see below). But, the one fundamental difference is the fact that our model is explicitly geared towards a web context, to linked data environments, as the future platforms of scholarly communication and collaboration. As a consequence it uses RDF, RDFS and OWL as "glue" in an effort to ontologically formalize the primitives and their attributes as well as the relations that can be established in such an environment.

Our research is being carried out within the EU-funded DM2E[1] project and its sister projects, which includes the development of a digital humanities collaboration environment, and the development of best of breed semantic sampling and annotation tools such as Korbo[2] and Pundit[3] (originating from the SemLib project[4]). We will also share results of the JISC-funded TEXTUS[5] project which has objectives similar to those of DM2E, but extends the semantic annotation functionality into a shared citation and referencing system. And lastly, we will include the perspective of the "Virtual and Real Architecture of Knowledge"[6] activity within the "Image, Knowledge, Gestaltung" excellency cluster funded by DFG.

Among other objectives, one of the main goals of the DM2E project is to "work with digital humanities scholars and specialized application developers to explore usage scenarios of the content provided to Europeana in a specialised environment for humanities research generating digital heuristics and making data as well as heuristics available to specialised visualisation or reasoning environments". The results of DM2E are intended to contribute to emerging distributed, interactive production and processing environments that go well beyond traditional working paradigms in the scholarly culture of the humanities.

References

Anderson, S., T. Blanke, and S. Dunn (2010). Methodological commons: arts and humanities e-Science fundamentals. In: *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*. 368. 19-25.

Atkins, D. E., et al. (2003). Revolutionizing Science and Engineering Through Cyberinfrastructure. *Report of the National Science Foundation Blue-Ribbon Advisory Panel on Cyberinfrastructure.* http://www.nsf.gov/od/oci/reports/atkins.pdf

Bamboo. (2010). Project Bamboo Scholarly Practice Report. Retrieved from http://www.projectbamboo.org/wp-content/uploads/Project-Bamboo-Scholarly-Practices-Report.pdf

Benardou, A., P. Constantopoulos, C. Dallas, and D. Gavrilis (2010). A Conceptual Model for Scholarly Research Activity. iConference 2010. Retrieved from https://www.ideals.illinois.edu/handle/2142/14945

Blanke, T., and M. Hedges (2011a). Scholarly primitives: Building institutional infrastructure for

humanities e-Science. *Future Generation Computer Systems*. doi:10.1016/j.future.2011.06.006

Blanke, T., M. Bryant, M. Hedges, A. Aschenbrenner, and M. Priddy (2011b). Preparing DARIAH. In IEEE 7th International Conference on e-Science, 2011.

Bush, V. As We May Think. *Atlantic Magazine* (July 1945). http://www.theatlantic.com/magazine/archive/1945/07/as-we-may-think/303881/ .

Gradmann, S., and J. C. Meister (2008). Digital document and interpretation: re-thinking "text" and scholarship in electronic settings. *Poiesis Praxis*. 5(2). doi:10.1007/s10202-007-0042-y

Rockwell, G. (2010). As Transparent as Infrastructure: On the research of cyberinfrastructure in the humanities. Retrieved from the Connexions Web site: http://cnx.org/content/m34315/1.2/

Schraefel, M. C. (2007). What is an Analogue for the Semantic Web and Why is Having One Important? Manchester: ACM Hypertext 2007. Retrieved from http://eprints.soton.ac.uk/264274/1/schraefelSWAnalogueHT07pre.pdf

Unsworth, J. (2000). Scholarly Primitives: what methods do humanities researchers have in common, and how might our tools reflect this? *Symposium on Humanities Computing formal methods experimental practice*. Retrieved from http://www3.isrl.illinois.edu/~unsworth/Kings.5-00/primitives.html

Unsworth, J., et al. (2006). Our Cultural Commonwealth. Report of the American Council of Learned Societies Commission on Cyberinfrastructure for the Humanities and Social Sciences. Retrieved from http://www.acls.org/cyberinfrastructure/cyber.htm

Unsworth, J., and C. Tupman (2011). Interview with John Unsworth, April 2011, carried out and transcribed by Charlotte Tupman. In: Deegan, M. und W. McCarty (Hg.): *Collaborative research in the digital humanities*. Farnham: Ashgate.

Notes

http://dm2e.eu/
http://dm2e.eu/
http://thepund.it/
http://thepund.it/
http://textusproject.org/
http://www.interdisciplinary-laboratory.hu-berlin.de/en/Virtual-and-Real-Architecture-of-Knowledge

Open Notebook Humanities: Promise and Problems

Shaw, Ryan
ryanshaw@unc.edu
University of North Carolina at Chapel Hill, United States of America

Buckland, Michael
buckland@ischool.berkeley.edu
University of California, Berkeley, United States of America

Golden, Patrick
ptgolden@berkeley.edu
University of California, Berkeley, United States of America

"When found, make a note of" (Dickens 1848, 149). In 1849, William Thoms took this rule as the motto of his new journal *Notes and Queries*, observing that following this rule for any length of time will result in "a good deal of matter in various forms, shapes and sizes . . . [in] countless boxes and drawers, and pigeon-holes of such things, which want looking over, and would well repay the trouble" (Thoms 1849, 1–2). Thoms could have been describing the offices of a contemporary documentary editing project, except that these days the "pigeon-holes" include shared hard drives. Good documentary editors follow this rule scrupulously, and as their projects stretch on into multiple decades, they amass a rich storehouse of notes, which would indeed repay the trouble of those who would look over them. Yet few have this opportunity, as only a small fraction of the content of these notes are ever made available.

Our aim in the Editors' Notes project (http://editorsnotes.org/) is much the same as that of Thoms in 1849: to provide a "medium by which much valuable information may become a sort of common property among those who can appreciate and use it" (Thoms 1849, 2). Much recent work in the digital humanities has focused on exploring ways to use networked computing to change scholarly practice. Tools like Zotero encourage open sharing of bibliographic data (Cohen 2008). Sharing of scholarly annotations has also been widely explored, and the Open Annotation Collaboration is working on developing standards and tools for making annotations interoperable across multiple tools (Hunter et al. 2010). Various projects have experimented with widening participation in humanist

scholarship through "crowdsourcing" (Causer, Tonra, and Wallace 2012). Other projects have sought to banish the stereotype of the "lone scholar" through experiments in collaborative authorship of edited volumes (Dougherty and Nawrotzki 2012). And considerable effort has been made to increase access to the products of humanist scholarship through the creation of open access journals, monographs, and scholarly editions, most of these building upon a well-established humanities practice of digital editing and publishing.

These various projects address many aspects of the humanities research process. But notes have been curiously overlooked. Reference management tools like Zotero do enable shared note taking on individual documents, but this functionality is secondary to the management and sharing of bibliographic data. Other projects primarily focus either on managing research "inputs" such as annotations and transcriptions of source documents, or on "outputs"—finished scholarly products—whether these are books, databases, or virtual environments. With Editors' Notes we are addressing the space in-between: the writing, organization, and linking of working notes, which are relevant to source documents but not necessarily tied to any specific document, and which may or may not become a formal finished product.

While not yet widely explored in the humanities, this space is one that has recently received much attention in the sciences. Both the National Institutes of Health and the National Science Foundation have instituted data sharing policies that encourage the researchers they fund to make

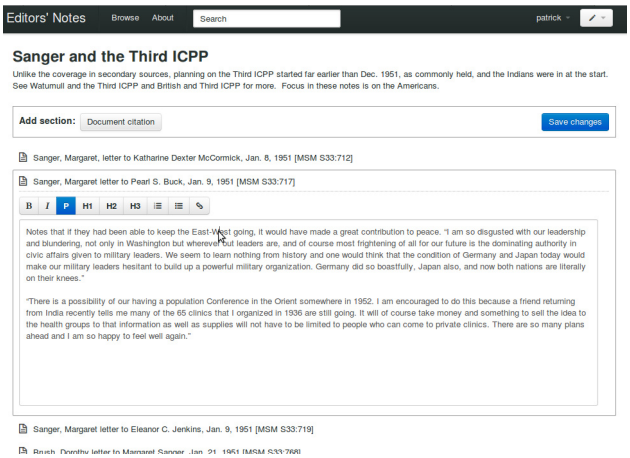


Figure 1.
Editing one section of a note on "Sanger and the Third ICPP".

available to other researchers the final research data underpinning their published work. "Open notebook science" presents a more radical vision in which not just the final research data, but all data generated during every stage of the research process is captured and made publicly available, either in "real time" or at the conclusion of a project. Advocates of this radically transparent approach to scientific practice expect it to result in more verifiable and reproducible research results, more efficient management and re-use of data on both the local and global levels, and new forms of algorithmic and "crowdsourced" research (Velden and Lagoze 2009).

The case for open notebook science rests upon the recognition that data from failed or incomplete experiments are potentially as important as those from successful ones. The problem is that current models of scientific publishing provide few incentives to publish such data, nor are there places to put it. Historians face similar problems, especially those engaged in long-running projects like documentary editions that may involve dozens of researchers working over decades. It is typical for a researcher to spend hours upon hours researching a topic only to find that she has duplicated work done years earlier and stowed away in a file cabinet or on a floppy disk.

William Thoms recognized back in 1849 that a major benefit of sharing working notes would be to induce researchers to "look over their own collections" and, by allowing others access, improve their own chances of finding past work (Thoms 1849, 2). In other words, a researcher need not be motivated by scholarly altruism to share her work. Yet Thoms believed that were sharing of notes to become cheap and frequent, then researchers would not hesitate to give help, not only to others engaged in similar lines of research, but also to "those who are going different ways, and only meet at the crossings" (Thoms 1849, 2). As this research commons grew, so would the opportunities for such crossings, and the net result would be more efficient research at the global level as well as the local.

Thoms' vision is echoed in efforts by the scientific community to create publishing models that enable both finer-grained publication units (Mons and Velterop 2009; Groth, Gibson, and Velterop 2010; Mons et al. 2011) and new attribution practices (Nature Genetics 2007; Nature Genetics 2008; Giardine et al. 2011). These efforts recognize that citing published work helps drive scientific publication. They aim to expand the universe of citable work beyond the canonical research paper to units as small as individual statements. In doing so they hope to make visible the great iceberg of scientific work of which published papers are only the tip.

Scholarly footnotes such as those produced by documentary editors can be viewed as a form of nanopublication. One of the goals of the Editors' Notes project is to give the status of individual publications to footnotes and the working notes that led to those footnotes. These notes can be complemented by machine-readable "factoids" (Bradley and Short 2005) about people, places,

organizations, and events, drawn from open-access linked datasets (Heath and Bizer 2011). Scholars can assess and improve the quality of these factoids, connecting assertions to bibliographic descriptions of evidential resources and publishing "gold standard" datasets that meet their high standards (Shaw and Buckland 2011). The scholars' notes provide context for the otherwise bare factoids, documenting why and to what extent they have chosen to accept them and the conclusions they have thus drawn.

Mons and Velterop (2009) make a distinction between "curated" and "observational" statements in scientific discourse. "Curated" statements take the form of records in trusted scientific databases. For example, a database recording known protein interactions may contain statements about these interactions along with metadata describing their context, conditions and provenance. In contrast, "observational" statements are factual statements such as "malaria is transmitted by mosquitos" that have not been formally recorded in any database but nevertheless are commonly asserted. A goal of nanopublication is to build knowledge bases that transform observational statements into curated ones.

History and the humanities mostly lack the databases of curated statements that exists in the sciences. The closest equivalents might be prosopographical or genealogical databases (Bradley and Short 2005; Church of Jesus Christ of Latter-day Saints 2012) or digital historical gazetteers (Elliott and Gillies 2011). These are exceptions that prove the rule, however, and the vast majority of factual statements in history and the humanities remain at the observational level.

To facilitate the shift from closed, personal or project-specific notes to openly shared notes, we've had to address a number of challenges. Editorial projects take varying approaches to structuring their research workflow, dividing labor among editors and student assistants, and standardizing on naming and citation practices. We have attempted to accommodate these varying work practices while creating opportunities for standardization across projects where it is desired. Our efforts to accommodate existing practices align with the broader objective of not disrupting ongoing research by integrating with research tools already in use. For example, Editors' Notes integrates with the Zotero bibliographic data management platform, allowing researchers to access their existing bibliographic databases (Shaw, Buckland, and Golden 2012).

A major challenge has been developing a data model for notes that is flexible enough to accommodate a variety of working styles (Figure 2). We have tried to support fine-grained addressing and indexing of notes, allowing researchers to search for and link to notes taken on a single source document as it relates to one narrow topic. At the same time, we have sought to develop ways that researchers can work with aggregations of these small "atoms" in ways that feel natural to them. For example, notes taken while researching "the status of birth control in India in the 1930s" might reference dozens of documents encompass several more specific topics such as the Indian birth control activist Dhanvanti Handoo Rama Rau, birth control clinics, and the Bombay Municipal Corporation. Researchers can work with these notes in the context of the broader research task, or they can pull together all the notes about Rama Rau, whether or not these were taken in the course of researching "the status of birth control in India in the 1930s."

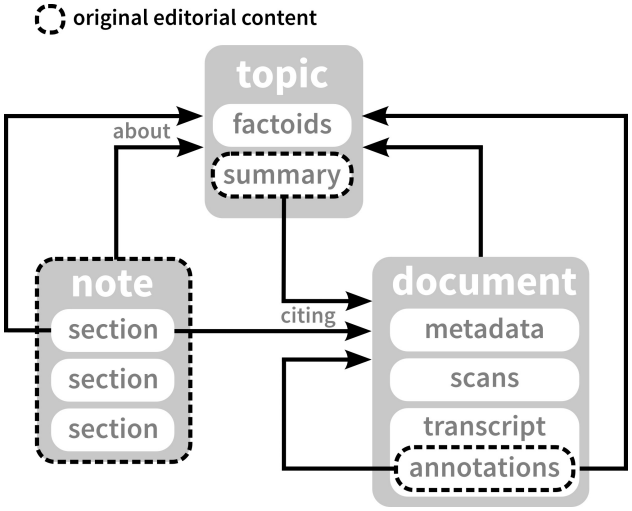


Figure 2.
Part of the Editors' Notes data model. Notes, sections of notes, and topic summaries may cite document. Notes, sections of notes, documents, and document annotations are linked to the topics to which they relate.

Another ongoing challenge has been the question of how to bridge the gap between note-taking practices in scholarly research projects and those of curators of special collections and archives. The Joseph A. Labadie special collection of radical history at the University of Michigan helped us explore this question by providing thousands of notes created by Agnes Inglis, the first curator of the collection. The subject matter of these notes overlapped with that of the editorial projects involved, but these "curator's notes" turned out to be useful less for their content per se, than for the metadata infrastructure (network of relationships among names and other topics) they produced. This realization helped catalyze our ongoing experimentation with incorporating linked data from libraries and archives.

Acknowledgements

We are grateful to the Andrew W. Mellon Foundation for funding "Editorial Practices and the Web" (http://ecai.org/

mellon2010) and for the cooperation and feedback of our colleagues at the Emma Goldman Papers, the Margaret Sanger Papers, the Elizabeth Cady Stanton and Susan B. Anthony Papers, and the Joseph A. Labadie Collection.

References

Bradley, J., and H. Short (2005). Texts into Databases: The Evolving Field of New-style Prosopography. *Literary and Linguistic Computing*. 20 (Suppl). 3–24. doi:10.1093/llc/fqi022.

Causer, T., J. Tonra and V. Wallace (2012). Transcription Maximized; Expense Minimized? Crowdsourcing and Editing. *The Collected Works of Jeremy Bentham*. Literary & Linguistic Computing 27 (2). 119–137. doi:10.1093/llc/fqs004.

Church of Jesus Christ of Latter-day Saints (2012). *FamilySearch*. https://familysearch.org/ .

Cohen, D. J. (2008). Creating Scholarly Tools and Resources for the Digital Ecosystem: Building Connections in the Zotero Project. *First Monday.* 13(8). http://firstmonday.org/htbin/cgiwrap/bin/ojs/index.php/fm/article/viewArticle/2233/2017 .

Dickens, C. (1848). *Dombey and Son*. Boston: Bradbury and Guild. http://books.google.com/books?id=3r1yo6lx3BsC.

Dougherty, J., and K. Nawrotzki (eds). (2012). *Writing History in the Digital Age*. Trinity College web-book edition. http://writinghistory.trincoll.edu/ .

Elliott, T., and S. Gillies (2011). Pleiades: an un-GIS for Ancient Geography. In *Digital Humanities*. held June 19-22 in Stanford, California. http://dh2011abstracts.stanford.edu/xtf/view?docId=tei/ab-192.xml .

Giardine, B., J. Borg, D. R. Higgs, K. R. Peterson, S. Philipsen, D. Maglott, B. K. Singleton, D. J. Anstee, A. Nazli Basak, B. Clark, F. C. Costa, P. Faustino, H. Fedosyuk, A. E. Felice, A. Francina, R. Galanello, M. V. E. Gallivan, M. Georgitsi, R. J. Gibbons, P. C. Giordano, C. L. Harteveld, J. D. Hoyer, M. Jarvis, P. Joly, E. Kanavakis, P. Kollia, S. Menzel, W. Miller, K. Moradkhani, J. Old, A. Papachatzopoulou, M. N. Papadakis, P. Papadopoulos, S. Pavlovic, L. Perseu, M. Radmilovic, C. Riemer, S. Satta, I. Schrijver, M. Stojiljkovic, S. Lay Thein, J. Traeger-Synodinos, R. Tully, T. Wada, J. S. Waye, C. Wiemann, B. Zukic, D. H. K. Chui, H. Wajcman, R. C. Hardison, and G. P. Patrinos. (2011). *Nature Genetics* 43:295–301. doi:10.1038/ng.785.

Groth, P., A. Gibson and J. Velterop (2010). The Anatomy of a Nanopublication. *Information Services and Use.* 30 (1-2). 51–56. doi:10.3233/ISU-2010-0613.

Heath, T., and C. Bizer (2011). *Linked Data: Evolving the Web into a Global Data Space*. Morgan & Claypool. doi:10.2200/S00334ED1V01Y201102WBE001.

Hunter, J., T. Cole, R. Sanderson, and H. Van de Sompel (2010). The Open Annotation Collaboration: A Data Model to Support Sharing and Interoperability of Scholarly Annotations. Paper presented at Digital Humanities, London, July 7–10. http://dh2010.cch.kcl.ac.uk/academic-programme/abstracts/papers/html/ab-860.html .

Mons, B. and J. Velterop (2009). Nano-Publication in the e-Science Era. In Clark, T., Luciano, J. S., Marshall, M. S., Prud'hommeaux, E., and Stephens, S. (eds). *Proceedings of the Workshop on Semantic Web Applications in Scientific Discourse* (SWASD 2009). http://ceur-ws.org/Vol-523/.

Mons, B., H. van Haagen, C. Chichester, P.-B. Hoen, J. T. den Dunnen, G. van Ommen, E. van Mulligen, B. Singh, R. Hooft, M. Roos, J. Hammond, B. Kiesel, B. Giardine, J. Velterop, P. Groth, and E. Schultes. (2011). The Value of Data. *Nature Genetics* 43:281–283. doi:10.1038/ng0411-281.

***Nature Genetics*.** (2007). Compete, Collaborate, Compel. 39(8):931. doi:10.1038/ng0807-931.

***Nature Genetics*.** (2008). Human Variome Microattribution Reviews. 40(1):1. doi:10.1038/ng0108-1.

Shaw, R. and M. Buckland (2011). "Editorial Control over Linked Data." *Proceedings of the American Society for Information Science and Technology* 48. doi:10.1002/meet.2011.14504801296.

Shaw, R., M. Buckland, and P. Golden (2012). Integrating Collaborative Bibliography and Research. *Proceedings of the American Society for Information Science and Technology* 49. doi:10.1002/meet.14504901245

Thoms, W. J. (1849). Notes and Queries. *Notes and Queries* s1-I(1):1–3. http://nq.oxfordjournals.org/content/s1-I/1/1.full.pdf+html .

Velden, T., and C. Lagoze (2009). Communicating Chemistry. *Nature Chemistry* 1:673–678. doi:10.1038/nchem.448.

LEXUS 3 — a collaborative environment for multimedia lexica

Shayan, Shakila
shakila.shayan@mpi.nl
Max Planck Institute for Psycholinguistics, The Netherlands

Moreira, André
andre.moreira@mpi.nl
Max Planck Institute for Psycholinguistics, The Netherlands

Windhouwer, Menzo
Menzo.Windhouwer@mpi.nl
Max Planck Institute for Psycholinguistics, The Netherlands

König, Alexander
Alexander.Koenig@mpi.nl
Max Planck Institute for Psycholinguistics, The Netherlands

Drude, Sebastian
Sebastian.Drude@mpi.nl
Max Planck Institute for Psycholinguistics, The Netherlands

LEXUS (The Language Archive 2013; Ringersma, et al. 2007) is a flexible web-based lexicon tool that was initially (between 2006 and 2010) developed at the Max Planck Institute for Psycholinguistics in Nijmegen within the program "documentation of endangered languages" (DOBES), funded by the Volkswagen Foundation (cf. Volkswagen Foundation). It is a tool specifically tailored for linguists whose research involves collecting and documenting a broad range of spoken data; data that is mainly in the form of audio and video recordings and often depicts a language that has never been documented before and is in danger of becoming extinct. Projects that are supported within the DOBES domain have specific requirements to be met by a lexicon tool. In particular the lexicon tool has to provide a high degree of flexibility with respect to lexicon structure to allow all possible structures, even those that might not have been seen in the description of other languages. Also, given the audio-visual nature of the data, there needs to be proper facilities for presenting multimedia within the lexicon. In what follows we will give a brief overview of how LEXUS is fine-tuned to meet most of the demands of this domain and how it is a useful tool for documenting endangered languages while at the same time suitable as a lexicographic tool for any other language.

LEXUS is not the only lexicographic tool with endangered languages as its key area of application. The KirrKirr software (Manning et al, 2001), the IDD (Indiana Dictionary Database, cf. De Korne, et al. 2009), the Toolbox, Lexique Pro, WeSay and FLEx (SIL 2013) programs are other examples, among which the SIL tools are the most comparable to Lexus. Toolbox is a tool for data management and analysis which includes lexical data and a parsing and glossing engine. Still, the outdated data model of toolbox seriously impedes its sustainability and interoperability with other tools. Lexique-pro is suitable for visualizing Toolbox/Shoebox data. FLEx on the other hand is a powerful lexicon tool with advanced parsing and analytical functionalities. Using other SIL tools one has sophisticated options to present and publish FLEx data. In terms of features and functionality for lexical databases by themselves, LEXUS stands somewhere in the middle with its set of features that aim at user-friendliness. Crucially, it offers online and shared access to the lexicon. It also allows for interoperability and customizable visualization for styled HTML views of data without requiring any knowledge of markup languages.

LEXUS makes use of schema structure trees for representing lexical entry elements. However, it does not enforce any pre-described schema structure for the lexicon of a given language. Instead, upon creating a new lexicon, LEXUS provides the user with a small collection of schema templates to choose from. These templates can be further developed into more complex schemas to build a particular lexicon. Some of these proposed templates have been created based on more standard frameworks such as ISO LMF (ISO 14613, 2008) and promote the usage of concept names and conventions that are proposed by the ISO data categories in ISOcat (Kemps-Snijders et al, 2009). There are even some templates that have been suggested and fine-tuned based on the experience of field linguists who have been involved in documenting endangered languages.

Having a flexible structure as a template is expected to be a helpful starting point for the researchers who are in the initial stages of the documentation process; especially when it comes to a language that has never been studied before. LEXUS' attempt to more fully comply with ISO LMF structure demands a structure that offers circularities of a graph type structure in addition to a simple linear tree structure. LEXUS 3 now makes it possible to create cross-references between corresponding elements of two different lexical entries of a lexicon. The template collection together with cross-reference linking makes LEXUS quite flexible in creating a wide range of structures. This combination makes LEXUS a suitable lexicon tool for not only under-described languages but also for the most studied and well-documented languages such as English and German.

Previous versions of LEXUS have laid the groundwork to enable projects within the DOBES program to create and view multimedia lexica with a team. LEXUS 3 stabilizes these core features and offers new functions which broaden the scope to the wider linguistic community. One such example is making use of the templates to form the basis for flexible import and export functionality as fleshed out in the RELISH project (Aristar-Dry et al, 2012). This in turn provides better support for standardized lexicon formats and makes it, for example, possible to export a LEXUS 3 lexicon to the LEGO repository (LinguistList 2013) and to import a LEGO lexicon. This supporting feature is based

on the new RELISH-LMF serialization (Windhouwer, et al. 2013; RELISH 2013), which is extensible using RELAX NG (ISO 19757-2, 2008) and ISO/TEI feature structures (ISO 24610-1 2008). LEXUS 3 is also being integrated into linguistic infrastructures like CLARIN (CLARIN 2013), which opens it up for an expanding user base.

Digital lexicography is most helpful when there is proper visualization of the content, and it gets even more worthwhile when it is integrated with multimedia. One of LEXUS' distinct features are the possibilities it offers for visualization of the lexicon, which is often enriched with multimedia. The audio-video recordings and images are critical aspects of the semantic knowledge. Data of such nature is often the only available resource to study the kind of languages that LEXUS is designed for. The written form, together with grammatical, morphological and phonological descriptions of words, completes the semantic knowledge. A given lexicon tool should facilitate an unified way of presenting the text and multimedia together to be able to put the form and meaning next to each other in one picture. With version 3, LEXUS introduces a single unified environment, where users can describe how their lexica would look like, for e.g. in an HTML view (Moreira, et al. 2013). In doing so, users are not required to have any HTML knowledge. Instead, LEXUS offers a graphical tree tool, which mimics the hierarchical structure that is behind the markup-based technologies such as HTML, and which specifies the layout and style of the lexical views. The same tool can be used to create a formatted PDF view for the full lexicon, which is extractable and available for export and print. *Figure 1* shows an example of a lexical entry, with an image and different styles for various elements of the entry. The style and the content of list of entries shown on the left side are customized with the same tool as those of the selected lexical entry shown on the right side.

***Figure 1*:**
an example of customized list view and lexical entry view using the same styling tool.

Finally and possibly most importantly, LEXUS allows shared access to a given lexicon. Owners can easily share their lexica by dragging and dropping other users from a list of all registered LEXUS users to the list of readers/writers of an individual lexicon. When a lexicon is shared, it becomes available in the target user's workspace. This feature paves the way for better collaboration among researchers and facilitates simultaneous work on a given language even from different places in the world. Being an online resource, any archived lexicon will in future be available as an addressable resource in the CLARIN infrastructure. Having an online basis LEXUS also allows for annotated multimedia sessions from the relevant language archive (e.g. sessions depicting particular cultural uses or social practices) to be linked to any part of the entry. The web-based accessibility of LEXUS was initially designed so as to allow the speech community members to get involved in the lexical documentation process. Such collaboration would allow for the continuous and faster growth of linguistic information. However, this appealing potential turned out to be more challenging in practice, mainly due to unforeseen obstacles within the social dynamics in the community (Cablitz 2011, 240).

In the future, special focus will be on an expansion of LEXUS' functionality to match the requirements of a lexicon tool for sign language documentation; a domain for which there doesn't exist a suitable tool or a unified schema of lexicography.

With its online accessibility, together with its archive-linking capacity, its multimedia visualization features and its interoperability capacities, we offer LEXUS as an advanced resource and research tool for the scientific community.

References

Aristar-Dry, H., S. Drude, J. Gippert, I. Nevskaya, and M. Windhouwer (2012). Rendering Endangered Lexicons Interoperable through Standards Harmonization: the RELISH project. In European Language Resources Association (ed), *Proceedings of the Eight International Conference on Language Resources and Evaluation,* held 23-25 May 2012 in Istanbul, Turkey.

Cablitz, G. (2011). The Making of a multimedia encyclopedic lexicon for and in endangered speech communities. In Haig, G. L. J., N. Nau, S. Schnell, and C. Wegener (eds.), *Documenting endangered languages: Achievements and perspectives.* Berlin: De Gruyter, 223-261.

CLARIN (2013). *Common Language Resource and Technology Infrastructure.* http://www.clarin.eu/ (accessed 14 March 2013).

De Korne, H., and the Burt Lake Band of Ottawa and Chippewa Indians (2009). The Pedagogical Potential of Multimedia Dictionaries — Lessons from a Community Dictionary Project. In Reyhner, J. and Lockard, L.(eds), *Indigenous Language Revitalization: Encouragement, Guidance and Lessons Learned.* Flaggstaff, AZ: Northern Arizona University, 141-153.

DOBES (2013). *Documentation of Endangered Languages*. http://www.mpi.nl/dobes/ (accessed 14 March 2013).

ISO 14613. (2008). Language resource management — Lexical markup framework (LMF). *International Organization for Standardization.* http://www.lexicalmarkupframework.org/ (accessed 14 March 2013).

ISO 19757-2. (2008). Information technology — Document Schema Definition Language (DSDL) — Part 2: Regular-grammar-based validation — RELAX NG, *International Organization for Standardization.*

ISO 24610-1.(2008). Language resource management — Feature structures — Part 1: Feature structure representation, *International Organization for Standardization.*

Kemps-Snijders, M., M. A. Windhouwer, P. Wittenburg, and S. E. Wright. (2009). ISOcat: Remodeling Metadata for Language Resources. In *Open Forum on Metadata Registries of the International Journal of Metadata, Semantics and Ontologies* (IJMSO), 4:4: 261-276.

LinguistList. (2013). *Lexicon Enhancement via the GOLD Ontology.* http://lego.linguistlist.org/ (accessed 14 March 2013).

Manning, C. D., K. Jansz, and N. Indurkhya (2001). Kirrkirr: Software for browsing and visual exploration of a structured Warlpiri dictionary. *Literary and Linguistic Computing,* 16: 123–139.

Moreira, A., M. Windhouwer, A. König, and S. Shayan (2013). LEXUS 3: Uniform Presentation Methodology for Lexica, *International Conference on Language Documentation & Conservation' (ICLDC)*, held 28 February-3 March 2013 in Hawaii.

RELISH.(2013). RELISH-LMF. http://tla.mpi.nl/relish/lmf/ (accessed 14 March 2013).

Ringersma, J., and M. Kemps-Snijders (2007). Creating multimedia dictionaries of endangered languages using LEXUS. In Van Hamme, H. and van Son, R. (eds), *Proceedings of Interspeech 2007.* Baixas, France: ISCA-Int. Speech Communication Assoc, 65-68.

SIL.(2013). http://www.sil.org/resources/software (accessed 14 March 2013).

TLA.(2013). *LEXUS — A web based lexicon tool.* http://tla.mpi.nl/tools/tla-tools/lexus/ (accessed 14 March 2013).

Volkswagen Foundation. (2013). http://www.volkswagenstiftung.de/ (accessed 14 March 2013).

Windhouwer, M., J.Petro, I. Nevskaya, S. Drude, H. Aristar-Dry, and J. Gippert (2013). Creating a serialization of LMF: the experience of the RELISH project. In G. Francopoulo (ed.), *LMF: Lexical Markup Framework, theory and practice.* iSTE/Wiley.

Meta-Methodologies and the DH Methodological Commons: Potential Contribution of Management and Entrepreneurship to DH Skill Development

Siemens, Lynne
siemensl@uvic.ca
University of Victoria

Given that the nature of the research work involves computers and a variety of skills and expertise, Digital Humanities (DH) researchers are working in teams with Humanists, Computer Scientists, other academics, undergraduate and graduate students, computer programmers and developers, librarians and others within their institutions and beyond (Williford et al., 2012). These projects' scope and scale also requires larger budgets than is typically associated with Humanities research (Siemens, 2009, Siemens et al., 2011a). As a result, Digital Humanists must developed meta-methodological skills, including teamwork and project management, to support the important and necessary methodological and technological ones in order to achieve project success. Programs such as the University of Victoria's Digital Humanities Summer Institute's Large Project Management Workshop (www.dhsi.org), MITH's Digital Humanities Winter Institute (MITH 2012), University of Virginia's Praxis Program (Scholars' Lab, 2011) and internships with libraries and DH centres (Conway et al. 2010) are contributing to the formal skill development in these meta-

methodological areas while DH teams themselves are reflecting on the importance and nature of collaboration skills developed directly through experience (Liu, et al. 2007; Ruecker, et al. 2007; Ruecker, et al. 2008; Siemens, et al. 2010).

While McCarty (2005) articulates a "intellectual and disciplinary map" for DH (118), which has become an important and agreed upon description of the discipline, no corresponding framework of these necessary meta-methodological skills exists that can guide the preparation of Digital Humanists to be as "comfortable writing code as they are managing teams and budgets" (Scholars' Lab 2011), either within a traditional academic post or an alternative academic one (The Praxis Program at the Scholars' Lab 2012). By drawing upon exemplary DH projects which exhibit individual components of this framework, this paper contributes to larger discussions by suggesting those important meta-methodological skills, knowledge and tasks, whose absences may impact a digital project's success and long-term sustainability. It ties together the many discussions about undergraduate and graduate DH training and education happening simultaneously across many various forums.

As can be seen in Figure 1, these meta-methodological skills involve more than collaboration and project management skills and include those typically associated with academic entrepreneurship. In their recent report on DH project sustainability, Maron and Loy (2011b) characterize digital resource projects as "small start-up businesses" (32) and use entrepreneurial language, such as empowered leadership, creation of a strong value proposition, cost management, revenue strategies and others to describe the ways in which case study DH projects have managed the impact of the current economic crisis on them. Further, their definition of sustainability, "the ability to generate or gain access to the resources — financial or otherwise — needed to protect and increase the value of the content or service for those who use it" draws further attention to skills, such as management, leadership, budget management, creation of alternative revenue streams, knowledge about users and their needs, and others (Maron, et al. 2011b, 10; Shane 2004). This view has been reinforced in a recent report that reviewed the Digging into Data program. Using the table metaphor, the authors likened project management expertise as the fourth leg, along side data management, domain, and analytical expertise (Williford, et al. 2012). Many digital projects are incorporating aspects of these skills, knowledge and tasks already.

For example, several projects are taking active steps to identify and understand users and their needs and then to ensure that the digital resources address these and allow the user to do more or do it faster or differently than before, while ensuring that knowledge of these resources reach the users (Warwick, et al. 2006; Siemens, et al. 2011b; Ithaka S&R, 2009; Shane 2004; Cohen, et al. 2005). To this end, the LARIAH report (2006) recommends that digital projects consult users widely, often, in a variety of forms and at the different development stages to ensure that a digital resource is in fact used consistently show "a deep understanding and respect for the value their resource contributes to those who use it" (Maron et al., 2009, pg. 7). One way to approach this is have users directly involved in the resource's development as is the case with TEI-C (nd-b) and Zotero (nd). In other cases, a digital resource can identified new users and extended its services to them, as can be seen with NINES' move beyond the 19th century to develop a supporting resource for the 18th (18thConnect; Bromley 2011). Alternatively, the digital resource may broaden offerings with new value added services, such as community websites, discussion lists, book and journal distributions and others, in response to users' requests (Iter 2011). Social media, such as twitter, youtube, and Wikipedia, can also be used to both gather information about users and promote the resource itself. For example, The Modernist Versions project combined an #yearofulysses, webpage, digital version release of Ulysses, and Ulysses Art Competition as a way to generate knowledge (and perhaps even excitement) about the initiative (The Modernist Versions Project). Alternatively, the TEI-C used a viral marketing experiment to understand the size and geographical distribution of its community of practice as it increased awareness of the guidelines (Siemens, et al. 2011b). Digital resource logos embedded in other projects can be important for generating awareness of a resource and its potential uses (TEI-C, nd-a; Zotero). At a basic level, projects must be pro-active in planning effective ways to promote themselves (Guiliano 2012).

As various granting agencies are increasingly prioritizing funding for project development rather than the maintenance and ongoing operations and resource improvements (National Endowment for the Humanities, 2010, Maron et al., 2009), digital projects must develop alternative revenue and funding models that allow them to move from grant funding to "a longer term plan for ongoing growth and development" and to even survive changes in the funding environment (Maron, et al. 2011a, 4; 2011b). These models might include cooperative advertising and click-through ads (Internet Shakespeare Editions, 2010), subscriptions (Iter 2011) and smart phone apps (iHistory Tours 2010b). In order to ensure appropriate levels of resources can be gathered, digital projects must also understand methods to budget the costs need to "cover the costs of the tasks essential to the development, support, maintenance and growth of their projects" (Maron, et al. 2009, 15; Guthrie, et al. 2008; Scholars' Lab, 2011). The latter becomes particularly important since there are expectations that digital resources will continue and be

updated and improved with changes in scholarship and technology (Kretzschmar Jr. 2009; Ithaka S&R 2009). Digital projects can then combine both revenue and cost management into long term sustainability plans, often required by funders (National Endowment for the Humanities 2010).

While financial resources are important, long term sustainability and development of digital projects also relies heavily on leadership (Maron, et al. 2011a; Maron, et al. 2009; Shane 2004; Siemens, 2009). These leaders have "a certain passion and tireless attention to setting and achieving goals" (Maron, et al. 2009, 7). This leadership is multifaceted and includes the supervision of human resources, such as paid staff, volunteers and collaborators, and cost and revenue management as well as strategic planning (Ithaka S&R 2009; Shane 2004; National Endowment for the Humanities Office of Digital Humanities 2010). As an example, Transcribe Bentham publishes regular project updates as motivators to its volunteer transcribers (Transcribe Bentham 2011). In addition, these project leaders need to continue to educate colleagues, administration and granting agencies so that these individuals understand the value of DH and the level of support needed for continued growth and development, including financial, in-kind and human resources, recognition and others (Ithaka S&R 2009; Siemens, 2010). The project leader for the 40-year old Thesaurus Linguae Graecae considers "it her job to 'educate' current and incoming administrators about her project" (Ithaka S&R 2009, 2). Finally, a strong leader understands when outside expertise and partnerships are necessary to ensure project success (Maron, et al. 2009). The Niagara 1812 iphone app development team included not only writers, researchers, and software developers, but also the nGen-Niagara Interactive media Generator and Brock Business Consulting Group, providing expertise that is not typically available in a History department (iHistory Tours, 2010a).

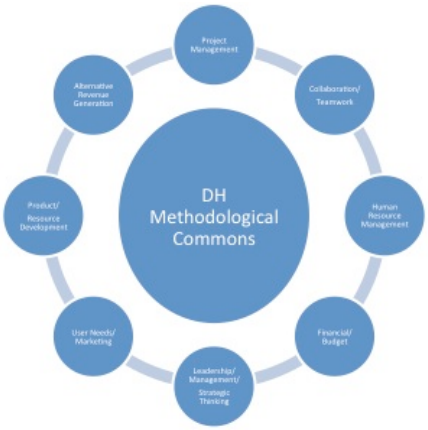


Figure 1:

Meta-methodological Commons[1]

Given the nature of digital projects, academics, particularly those from the Humanities, need to adopt these important meta-methodological skills and knowledge in addition to content and methodological ones to ensure project success (Ithaka S&R 2009; Cohen, et al. 2009; Cohen, et al. 2005). By outlining the range of meta-methodological skills, this framework has the potential to strengthen the positive work already ongoing to develop these skills and knowledge and to contribute to the discussion about the important skills that Digital Humanists need to create successful, useful and used projects (LeBlanc 2011; Leon 2011; McCarty 2011; Rovira 2011; Spiro 2011; Cohen, et al. 2005; Scholars' Lab 2011). This paper also contributes to the larger discussion of regarding undergraduate and graduate training and education in Humanities, DH, and beyond and provides a context for thinking about the additional skills needed for both faculty and alternative academic positions (Thaller, et al. 2012; Scholarly Communication Institute 2012; The Praxis Program at the Scholars' Lab 2012; Sample 2012; Pannapacker 2012; Williford, et al. 2012). Finally, it is designed to enable those who work in such teams or who work to support these individuals to recognise and develop the necessary meta-methodological skills and knowledge that lead to project success.

References

18thconnect What Is 18thconnect? http://www.18thconnect.org/18th_about/what_is.html (accessed October 13, 2011).

Bromley, A. (2011). Nines Project Enhances Tools for Digital Research in the Humanities. *UVa Today*.

Cohen, D. J., N. Fraistat, M. G. Kirschenbaum, and T. Scheinfeldt (2009). *Tools for Data-Driven Scholarship: Past, Present and Future*, Ellicott City, Maryland.

Cohen, D. J., and R. Rosenzweig (2005). Digital History: A Guide to Gathering, Preserving, and Presenting the Past on the Web http://www.chnm.gmu.edu/digitalhistory/audience/index.php (accessed October 28, 2011).

Conway, P., N. Fraistat, P. Galloway, K. Kraus, et al. (2010). Digital Humanities Internships: Creating a Model Ischool-Digital Humanities Center Partnership. *Digital Humanities 2010*. London, UK.

Guiliano, J. (2012). Neh Project Director's Meeting: Lessons for Promoting Your Project. http://mith.umd.edu/neh-project-directors-meeting-lessons-for-first-time-pis/ (accessed

Guthrie, K., R. Griffiths, and N. Maron (2008). *Sustainability and Revenue Models for Online Academic Resources: An Ithaka Report*, New York: Ithaka.

Ihistory Tours (2010a). About the Team. http://www.ihistorytours.com/about/ (accessed October 28, 2011).

Ihistory Tours (2010b). Niagara 1812. http://www.ihistorytours.com/ (accessed October 13, 2011).

Internet Shakespeare Editions (2010). Support the Internet Shakespeare Editions. http://internetshakespeare.uvic.ca/Foyer/donate.html (accessed February 2, 2011).

Iter (2011). Iter: Gateway to the Middle Ages & Renaissance. http://www.itergateway.org/ (accessed October 13, 2011).

Ithaka S&R (2009). Sustaining Digital Resources: A Briefing Paper for Leaders of Projects with Scholarly Content. http://sca.jiscinvolve.org/wp/files/2009/10/sca_bp_projects_scholarly_content_sep09_v1-02.pdf (accessed October 12, 2011).

Kretzschmar Jr., W. A. (2009). Large-Scale Humanities Computing Projects: Snakes Eating Tails, or Every End Is a New Beginning? *Digital Humanities Quarterly*, 3 (2).

Leblanc, M. (2011). Re: [Humanist] 25.382 Intro Topics and Texts. http://www.digitalhumanities.org/humanist/Archives/Current/Humanist.vol25.txt (accessed October 17, 2011).

Leon, S. M. (2011). Project Management for Humanists: Preparing Future Primary Investigators. http://mediacommons.futureofthebook.org/alt-ac/pieces/project-management-humanists (accessed June 24, 2011).

Liu, Y., and J. Smith (2007). Aligning the Agendas of Humanities and Computer Science Research: A Risk/Reward Analysis. *SDH-SEMI 2007*. Saskatoon, SK.

Maron, N., and M. Loy (2011a). *Funding for Sustainability: How Funders' Practices Influence the Future of Digital Resources*, Ithaca, NY: Ithaka S&R.

Maron, N., and M. Loy (2011b). *Revenue, Recession, Reliance: Revisiting the Sca/Ithaka S&R Case Studies in Sustainability*. Ithaca, NY: Ithaka S& R.

Maron, N., K. K. Smith, and M. Loy (2009). *Sustaining Digital Resources: An on-the-Ground View of Projects Today*, Ithaca, NY: Ithaka S&R.

Mccarty, W. (2005). *Humanities Computing*, New York: Palgrave MacMillan.

McCarty, W. (2011). [Humanist] 25.370 Intro Topics and Texts. http://www.digitalhumanities.org/humanist/Archives/Current/Humanist.vol25.txt (accessed October 17, 2011).

Mith (2012). Welcome to Dhwi. http://mith.umd.edu/dhwi/ (accessed

National Endowment for the Humanities (2010). *Digital Humanities Implementation Grants.* http://www.neh.gov/grants/guidelines/digitalhumanitiesimplementation.html (accessed October 12, 2011).

National Endowment for the Humanities Office of Digital Humanities (2010). *Summary Findings of Neh Digital Humanities Start-up Grants (2007-2010).* Washington, D.C.: National Endowment for the Humanities.

Pannapacker, W. (2012). No Dh, No Interview. http://chronicle.com/article/No-DH-No-Interview/132959/ (accessed

Rovira, J. (2011). Re: [Humanist] 25.370 Intro Topics and Text. http://www.digitalhumanities.org/humanist/Archives/Current/Humanist.vol25.txt (accessed October 17, 2011).

Ruecker, S., and M. Radzikowska (2007). The Iterative Design of a Project Charter for Interdisciplinary Research. *DIS 2008*. Cape Town, South Africa.

Ruecker, S., M. Radzikowska, and S. Sinclair (2008). Hackfests, Designfests, and Writingfests: The Role of Intense Periods of Face-to-Face Collaboration in International Research Teams. *Digital Humanities 2008*. Oulu, Finland.

Sample, M. (2012). Digital Humanities at Mla 2013. http://www.samplereality.com/2012/10/17/digital-humanities-at-mla-2013/ (accessed

Scholarly Communication Institute (2012). Landscape of Alternate-Curriculum Graduate Training Programs. https://docs.google.com/document/d/1NVPiPHeOWbvMBBpFiNXRsqsjzVZ2t2E6mtxb-p89iSQ/edit (accessed

Scholars' Lab (2011). The Praxis Program at the Scholars' Lab. http://praxis.scholarslab.org/ (accessed September 12, 2011).

Shane, S. (2004). *Academic Entrepreneurship: University Spinoffs and Wealth Creation*, Cheltenham, UK, Edward Elgar.

Siemens, L. (2009). 'It's a Team If You Use "Reply All": An Exploration of Research Teams in Digital Humanities Environments. *Literary & Linguistic Computing*. 24(2). 225-233.

Siemens, L. (2010). Developing Academic Capacity in Digital Humanities: Thoughts from the Canadian Community. *Digital Humanities 2010*. London, UK.

Siemens, L., R. Cunningham, W. Duff, and C. Warwick (2011a). A Tale of Two Cities: Implications of the Similarities and Differences in Collaborative Approaches within the Digital Libraries and Digital Humanities Communities *Literary & Linguistic Computing*, 26(3). 335-348.

Siemens, L., and Inke Research Group (2010). Understanding Long Term Collaboration: Reflections on Year 1 and Before *INKE 2010*. The Hague, Netherlands.

Siemens, L., R. G. Siemens, and H. Wen (2011b). "The Apex of Hipster Xml Geekdom: Tei-Encoded Dylan and Understanding the Scope of an Evolving Community of Practice. *Journal of the TEI Encoding Initiative*, 1(1).

Spiro, L. (2011). Knowing and Doing: Understanding the Digital Humanities Curriculum. http://digitalscholarship.files.wordpress.com/2011/06/spirodheducationpresentation2011-4.pdf (accessed October 17, 2011).

Tei-C (nd-a). Tei Badges. http://www.tei-c.org/About/Badges/ (accessed October 14, 2011).

Tei-C (nd-b). Tei Workgroups. http://www.tei-c.org/Activities/Workgroups/index.xml (accessed October 14, 2011).

Thaller, M., P. Sahle, F. Clavaud, T. Clement, et al. (2012). Digital Humanities as a University Degree: The Status Quo and Beyond. *Digital Humanities 2012*. Hamburg, Germany.

The Modernist Versions Project (nd). It's All About You(Lysses). http://web.uvic.ca/~mvp1922/you/ (accessed

The Praxis Program at the Scholars' Lab (2012). About Praxis. http://praxis.scholarslab.org/about.html (accessed

Transcribe Bentham (2011). Transcribe Bentham. http://www.ucl.ac.uk/transcribe-bentham/ (accessed October 28, 2011).

Warwick, C., M. Terras, P. Hunginton, N. Pappa, et al. (2006). *The Lairah Project: Log Analysis of Digital Resources in the Arts and Humanities, Final Report to the Arts and Humanities Research Council*, London: University College.

Williford, C., and C. Henry (2012). *One Culture: Computationally Intensive Research in the Humanities and Social Sciences: A Report on the Experiences of First Respondents to the Digging into Data Challenge*, Council on Library and Information Resources.

Zotero (nd). Get Involved with Zotero. http://www.zotero.org/getinvolved/ (accessed October 14, 2011).

Notes

1. The meta-methodological commons is centred around McCarty's (2005) articulation of the Digital Humanities Methodological Commons.

The Crowdsourcing Process: Decisions about Tasks, Expertise, Communities and Platforms

Siemens, Lynne
siemensl@uvic.ca
University of Victoria, Canada

Introduction

Business, governments, community groups and academic projects are turning to crowdsourcing, an internet-based process, to facilitate access outside expertise needed to complete various tasks (Brabham, 2008; Howe, 2006). This mechanism holds great potential for academic projects, particular for those involving large amounts of data that needs to be identified, transcribed, analyzed and/or catalogued. It can often supplement meager project budgets by sourcing the work at relatively low cost (Corney et al., 2009; Holley, 2010). Further, crowdsourcing supports growing calls for public engagement in research projects (SSHRC, 2012). Several academic projects are at the forefront of this trend with public involvement in data classification (Galaxy Zoo, 2010), error correction in texts (Holley, 2009), text transcription and/or annotation (Transcribe Bentham, 2012; Zou, 2011; Pynchonwiki, nd; Siemens, 2012), cataloging and metadata and database creation (The Bodleian Library, 2012; Picture Australia, nd), and many others.

As crowdsourcing is introduced into more activities, it becomes important to understand "how to manage the crowd in a networked society" to ensure that a project achieves its objectives (DISH, 2011). Initial studies in participation motivation have found that individuals are interested in participating in a larger cause, using their skills, earning recognition and/or being part of a community, and potentially money (Wexler, 2011; Organisciak, 2010; Raddick et al., 2010; Brabham, 2008, 2010). Other research has focused on the most appropriate ways to solicit and encourage participation among a potential community of contributors by making the activity in question fun, presenting a big challenge to be undertaken, and reporting on progress regularly (Digital Fishers, nd; Holley, 2012). However, most of the resaerch has been conducted within the private sector context and on short-term projects with

little need to manage volunteers over a longer period of time (Wexler, 2011), which may limit the applicability of results to academic projects.

While the use of crowdsourcing is increasing, little work has been done to understand ways to organize the work to ensure that this crowd's contribution is delivered within an academic's project's schedule, budget and other resources and to the required quality standard (Geiger et al., 2011; Organisciak, 2011). In particular, projects need to understand the most appropriate ways to organize work flows, technical infrastrucure, and staffing requirements to manage volunteers and confirm quality (Zou, 2011). Opportunity exists to build from reflections of several projects to understand these issues (DISH, 2011; Holley, 2012; Zou, 2011; Corney et al., 2009; Holley, 2009).

This paper will contribute to this discussion by reviewing the literature to suggest a crowdsourcing process framework explore the range of decisions that must be addressed in advance to ensure quality and successful project outcomes. In addition, it will report on interviews with several crowdsourcing projects with regards to their workflow organization. The paper will conclude with recommendations for projects contemplating this tool as a way to reach project outcomes with limited project funds.

Literature Review

While no common definition exists for crowdsourcing (Holley, 2010), every crowdsourcing project shares several components. As seen in Figure 1, there must be an organization, a particular task to be completed to meet project goals and outcomes at a specified quality level, and a community, comprised of both experts and novices, which is willing to do the work for little or no money. These interactions are facilitated through an internet-based platform (Brabham, 2012; Schenk et al., 2011; Geiger et al., 2011; Tong et al., 2012). The interested organization must make a series of decisions regarding the type of expertise, qualification and/or knowledge required, the presence of a contributors, the mechanisms by which they will participate and contribute, project remuneration, motivators to keep participants engaged, and quality control mechanisms (Geiger et al., 2011; Corney et al., 2009; Rouse, 2010; Organisciak, 2011). The range of tasks and required expertise can be seen along the continuum in Figure 2.

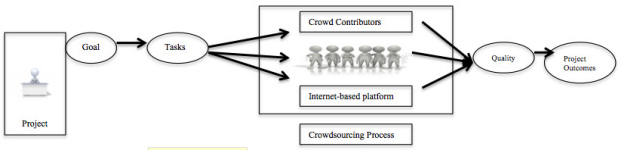


Figure 1: Crowdsourcing Approach

(Adapted from Geiger et al., 2011; Corney et al., 2009; Rouse, 2010)

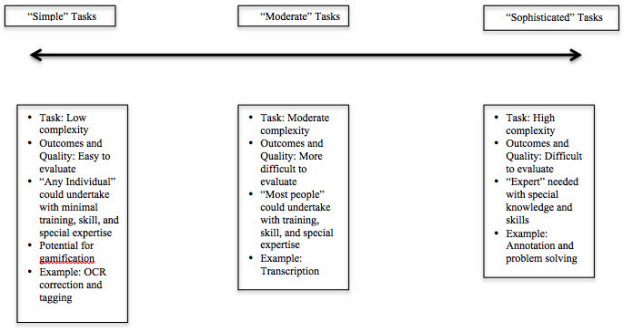


Figure 2: Crowdsourcing Task Continuum
(Adapted from Corney et al., 2009; Rouse, 2010)

Ultimately, projects then select the appropriate interface to both solicit and receive contributions from the public in ways that keep the "crowd" motivated and participating (Tong et al., 2012; Organisciak, 2010, 2011).

Methods

This project uses a qualitative research approach with in-depth interviews with members of crowdsourcing projects. The interview questions focus on the participants' use of the "crowd" to achieve tasks, ways to organize workflows, type of infrastructure in place to support the work, and challenges (Marshall et al., 1999; McCracken, 1988).

At the time of writing this proposal, final data analysis is being completed. The results will inform decision making needed to be made by projects to effectively and successfully use crowdsourcing.

This research will make several contributions to the knowledge base about ways to incorporate the public into academic projects. First, it builds on work already undertaken to community and engage the public in academic research through crowdsourcing (Causer et al., 2012; Holley, 2009; Organisciak, 2010). Second, it builds on earlier studies on participants' motivation to participate in these projects with an exploration of the organization of the tasks, process, and other components of the crowdsourcing process from the perspective of the project itself (Brabham, 2010; Organisciak, 2010; Wexler, 2011). Finally, it extends understanding about the nature of academic collaboration as projects expand relationships beyond the team to the public (Siemens, 2009).

References

Brabham, D. (2008). Crowdsourcing as a Model for Problem Solving: An Introduction and Cases. *Convergence: The International Journal of Research into New Media Technologies*, 14.1: 75-90.

Brabham, D. (2010). Moving the Crowd at Threadless: Motivations for Participation in a Crowdsourcing Application. *Information, Communication & Society*, 13.8 1122-1145.

Brabham, D. (2012). The Myth of Amateur Crowds: A Critical Disccourse Analysis of Crowdsourcing Coverage. *Information, Communication & Society*, 15.3 394-410.

Causer, T., J. Tonra, and V. Wallace (2012). Transcription Maximized; Expense Minimized? Crowdsourcing and Editing the Collected Works of Jeremy Bentham. *Literary and Linguistic Computing*.

Corney, J. R., C. Torres-Sánchez, P. Jagadeesan, and W. Regli (2009). Outsourcing Labour to the Cloud. International Journal of Innovation and Sustainable Development, 4.4. 294-313.

Digital Fishers (nd). Digital Fishers. http://digitalfishers.net/ (accessed February 21, 2012).

Dish (2011). Theme: Co-Creation and Crowdsourcing. http://www.dish2011.nl/themes/crowdsourcing-and-co-creation (accessed February 21, 2102).

Galaxy Zoo (2010). Galaxy Zoo. http://www.galaxyzoo.org/ (accessed February 22, 2012).

Geiger, D., S. Seedorf, T. Schulze, R. Nicerson, et al. (2011). Managing the Crowd: Towards a Taxonomy of Crowdsourcing Processes. *Seventeenth Americas Conference on Information Systems*. Detroit, Michigan.

Holley, R. (2009). *Many Hands Make Light Work: Public Collaborative Ocr Text Correction in Australian Historic Newspapers*, National Library of Australia.

Holley, R. (2010). Crowdsourcing: How and Why Should Libraries Do It? *D-Lib Magazine*.

Holley, R. (2012). Digital Cultural Heritage Awards for Crowdsourcing (and Thoughts on Gamification). http://rose-holley.blogspot.com/2012/02/digital-cultural-heritage-awards-for.html (accessed February 21, 2012).

Howe, J. (2006). The Rise of Crowdsourcing. *Wired*.

Marshall, C. and G. B. Rossman (1999). *Designing Qualitative Research*, Thousand Oaks, California, SAGE Publications.

Mccracken, G. (1988). *The Long Interview*. Newbury Park, CA: SAGE Publications.

Organisciak, P. (2010). Why Bother? Examining the Motivations of Users in Large-Scale Crowd-Powered Online Initiatives. *Humanities Computing — Library and Information Studies*. Edmonton: University of Alberta.

Organisciak, P. (2011). When to Ask for Help: Evaluating Projects for Crowdsourcing. *Digital Humanities 2011*. Stanford.

Picture Australia (nd). Trove: Australia in Pictures. http://www.flickr.com/groups/pictureaustralia_ppe/ (accessed

Pynchonwiki (nd). A Literary Wiki Exploring the Novels of Thomas Pynchon. http://pynchonwiki.com (accessed

Raddick, M. J., Bracey, G., Gay, P. L., Lintott, C. J., et al. (2010). Galaxy Zoo: Exploring the Motivations of Citizen Science Volunteers. *Astronomy Education Review*, 9.1.

Rouse, A. C. (2010). A Preliminary Taxonomy of Crowdsourcing. *ACIS 2010*. Brisbane, Australia.

Schenk, E. and C. Guittard (2011). Towards a Characterization of Crowdsourcing Practices. *Journal of Innovation Ecnomics*, 7.1: 93-107.

Siemens, L. (2009). 'It's a Team If You Use "Reply All": An Exploration of Research Teams in Digital Humanities Environments. *Literary & Linguistic Computing*, 24.2: 225-233.

Siemens, R. (2012). A Social Edition of the Devonshire Ms (Bl Add 17, 492). http://en.wikibooks.org/wiki/The_Devonshire_Manuscript

Sshrc (2012). Knowledge Mobilization. http://www.sshrc-crsh.gc.ca/society-societe/community-communite/index-eng.aspx - 2

The Bodleian Library (2012). Help Us to Describe the Libraries' Digitised Music Collections. http://www.whats-the-score.org

Tong, R., and K. R. Lakhani. (2012). *Public-Private Partnerships for Organizing and Executing Prize-Based Competitions*. Berkman Centre for Internet & Society.

Transcribe Bentham (2012). About Us. http://www.ucl.ac.uk/transcribe-bentham/about/ .

Wexler, M. N. (2011). Reconfiguring the Sociology of the Crowd: Exploring Crowdsourcing. *International Journal of Sociology and Social Policy*, 31.1-2: 6-20.

Zou, J. J. (2011). Civil War Project Shows Pros and Cons of Crowdsourcing. http://chronicle.com/blogs/wiredcampus/civil-war-project-shows-pros-and-cons-of-crowdsourcing/31749 (accessed February 21, 2012).

Digital Textual Studies, Social Informatics, and the Sociology of Texts: A Case Study in Early Digital Medievalism

Simpson, Grant Leyton

glsimpso@indiana.edu
Indiana University, United States of America

This paper investigates the use of digital textual studies alongside the discourse of Social Informatics (SI) for the purposes of developing rich sociotechnical analyses of information and computing technologies (ICTs) in their manifestations as both projects and objects. I intend this approach both as a way of studying various digital humanities (DH) projects but also as a way of practicing a form of DH that, while not focused on building DH objects itself, investigates the social and technical conditions under which they are built. Toward that end, I discuss the affinities between digital textual studies and, as a demonstration of the joint analytical power of textual studies and SI, I present a case study of an early project in the history of humanities computing, Bessinger and Smith's *A Concordance to Beowulf* (Bessinger & Smith, 1969).

Bibliography and textual criticism have long been characterized by an orientation toward media consciousness, one which has lately expanded to include the social aspects of artifacts. Indeed, as Matthew G. Kirschenbaum has argued (2008, p. 16), "textual studies," an umbrella term for bibliography and textual criticism, "should be recognized as among the most sophisticated branches of media studies we have evolved." As such, they are not just a set of methods for media-specific analysis; they can provide DH with valuable insights into the workings of digital media. For example, Alan Galey (2010, p. 93) has argued that "Textual scholarship stands to contribute two key ideas to the digital humanities: first, that there is more to electronic forms than what reaches the screen; and second, that the relationship of form to content is complex and sometimes beyond exhaustive modeling."

Contemporary theorists, such as D.F. McKenzie (1999) and McGann (2004) have argued for the application of textual studies to a broad range of electronic artifacts. Indeed, bibliography, according McKenzie (1999, 12), "is the discipline that studies texts as recorded forms, and the processes of their transmission, including their production and reception." (Thus, for McKenzie, textual criticism is a form of Bibliography writ large.) Toward this end, McKenzie (1999, 13) defines "texts" as

> verbal, visual, oral, and numeric data, in the form of maps, prints, and music, of archives of recorded sound, of films, videos, and any computer-stored information, everything in fact from epigraphy to the latest forms of discography. There is no evading the challenge which those new forms have created.

Many of these "new forms" overlap with the ICTs that SI and other sociotechnical discourses study. As Rob Kling, Howard Rosenbaum, and Steve Sawyer (2005, 11), explain, "The acronym 'ICT' refers to information and communication technology—artifacts and practices for recording, organizing, storing, manipulating, and communicating information." Moreover, there is also overlap in the social aspects of technology between SI and textual studies. In addition to "physical forms" and "textual versions," McKenzie (1995, 13) counts among bibliography's interests, "technical transmission, institutional control, [texts'] perceived meanings, and social effects."

Indeed, for McKenzie, bibliography is properly considered the sociology of texts. It is in that sense a sociotechnical discipline and thus shares affinities with SI. SI, as defined by Rob Kling, is the "interdisciplinary study of the design, uses, and consequences of information technologies that takes into account their interaction with institutional and cultural contexts" (Kling, 2000, 218). SI and the sociology of texts share an attention to the ways in which technology and society are mutually shaped. Texts are key nodes in what Kling, Geoff McKim, and Adam King (2003) have called socio-technical interaction networks (STINs).

As a demonstration of how textual criticism and the insights of SI might be used together to produce an analysis that is above and beyond what either discourse can provide by itself, I present in this paper the case of Jess B. Bessinger and Philip H. Smith's long term collaboration to produce a concordance of the *Anglo-Saxon Poetic Records*. Bessinger, a professor of English, and Smith, a professor of Computer Science (and former researcher with IBM) formed various alliances with colleagues, vendors, publishers, and others during the life of their project. They mobilized these alliances to acquire and produce software, have custom hardware produced for them, distribute knowledge about new techniques, solicit advice, and sell the two concordance volumes produced by the project. The story of their collaboration cannot be told without discussing such alliances, specifically the ways in which the ICTs used and made by the collaborators shaped and were shaped by disciplinary cultures and institutional contexts. SI is very well suited for discussing such aspects of the project. A preliminary depiction of a STIN for this case study, which shows various social connections between people, institutions, and technologies could look like the following:

Despite the pioneering technical aspects of the work that went into the project, the goal for Bessinger and Smith was always to produce a paper volume that would be used by Old English scholars—that is, a book that would fit into traditional paper-based scholarly workflows. Thus I propose employing, alongside SI, studying the language features of the concordances and what Jerome McGann (2004, 11) has called "the apparitions of text—its paratexts, bibliographic codes, and all visual features—[which] are as important in the text's signifying programs as the linguistic elements." By paying attention to such elements as they appear in the concordances, we can see evidence of a program of extreme automatization at work in their production. For example, whereas other computer-produced concordances of the era used computers to produce word lists that were then typeset by traditional methods, Bessinger and Smith used them not only to produce their word lists but also to typeset the camera-ready copy. This is clear, for example, in the fact that the text of the *Beowulf* concordance is struck in IBM's Artisan font, which has a distinctly 1960s computerized look.

Both the objects produced by digital humanists and the projects that produce said objects are worthy of study. By "project" I mean the nexus of formal and informal relationships between people, institutions, and technologies that come together to meet a set of goals. Social Informatics is an approach to studying this nexus. In the case of DH, project goals are often the creation of objects (both concrete and virtual). Thus, digital textual criticism provides a powerful approach to analyzing such objects. The contribution my paper makes to the critical conversation in DH is to introduce and theorize this project-object distinction. I will demonstrate how pairing SI and digital textual criticism helps us to open up projects and objects in a way that exposes relationships between people, tools, and the work they produce. This is not simply a matter of saying that we should look at an artifact in context, for that would imply a primacy of object over project where I don't intend one. In this paper, I lay out the case for why SI is a valuable discourse to be used in the study of DH. I discuss affinities between SI and digital textual criticism, and why a sociotechnical approach is important to understanding what DH practitioners do. By means of bibliography, textual criticism, and the STIN approach, I explore the Bessinger and Smith collaboration to demonstrate both the project's historical importance and the strength of this hybrid approach. I believe that this paper, given its interdisciplinary focus and willingness to bridge disciplinary boundaries, is in keeping with the theme of DH 2013, "Freedom to Explore."

A Humanist Perspective on Building Ontologies in Theory and Practice

Simpson, John Edward
john.simpson@ualberta.ca
University of Alberta, Canada

Brown, Susan
susanirenebrown@gmail.com
University of Alberta, Canada; University of Guelph, Canada

Goddard, Lisa
lgoddard@mun.ca
Memorial University, Canada

150 Word Summary: Bowker and Star (1999) warn of the shaping force of classification systems, the power of their apparent naturalization as part of infrastructures, and the consequences thereof. The rise of the semantic web and its grounding in the use of ontologies threatens to explode the scale of this issue. This paper explores the practical implications and theoretical limits of using ontologies, offering a survey of best practices and ongoing areas of active research. The survey shows that the process of ontology construction and integration in the semantic web does have the potential to restrict human thought. It is then argued that certain features of the semantic web make it possible to avoid this detrimental outcome but only if a critical mass of people—the community of humanists might be enough—actively use these features to keep the web a place of flexible ontologies and humane classification.

Introduction

Creating an ontology amounts to creating a classification system and asserting a system of logic with it. There are inherent consequences to doing this that are both "real world" and potentially serious. Bowker and Star make an assessment of these, arguing that, "it is politically and ethically crucial to recognize the vital role of infrastructure in the 'built moral environment.' Seemingly technical issues like how to name things and how to store data in fact constitute much of human interaction and much of what we come to know as natural" (Bowker and Star 1999, 326). The semantic web promises to explode this observation as the interplay of ontology proliferation and convergence

begins to create the largest classification system in human history. Insofar as using an ontology tends to establish certain actions, things, and dispositions as natural, this same tendency leads to problems of commensurability that intersect with seminal work by prominent philosophers such as Quine, Wittgenstein, Foucault, Kuhn, and the poststructural movement as a whole (cf. Cope, Kalantzis, and Magee 2011). Bowker and Star advocate the use of flexible or "living" classifications as the approach to solving these deep issues (Star 2010; Bowker and Star 2011), but this is only a prescription and it is left open how the medicine should be delivered.

In the semantic web, ontologies provide the vocabularies and the schemas that tell the machines used to crawl the web just what to do with data when it is found, a view that is supported by the number of introductions to the semantic web that treat it in a purely mechanical way. From such perspectives ontologies bring the promise of inferential reasoning on content and an open environment where information is more easily exchanged (W3C). Along with these benefits come potential problems, including the rise of near silent normalizations around how we think about the world that we live in and the very real possibility of incommensurability among ontologies. Clearly there is a large divide between benefits and concerns, one that is extended through the contrasts between practice and theory, stability and flexibility, mechanical and human(e).

This paper explores possibilities for traversing this divide and considers the extent to which it can be narrowed by considering both the practical implications and theoretical limits of using ontologies, offering a survey of best practices and ongoing areas of active research. From this basis, it argues that the process of ontology construction and integration currently taking place around the semantic web amounts to a massive act of world construction with the potential to severely impact human life through rigidly framing human thought. Certain features of the semantic web make it possible to avoid this, but only if a critical mass of people---the community of humanists might be enough---actively use these features to keep the web a place of flexible ontologies.[1]

This argument and the accompanying assessments of the state of the semantic web today emerge from work that we have done around building a web-based editor to allow scholars with limited technical expertise in either XML or RDF to contribute or enhance online scholarly materials in these data formats.[2] Enabling scholars to use these tools to connect with the open web through a system of references and annotations has meant that the various research groups we are involved with (CWRC, INKE, and the Text Mining & Visualization for Literary History Project) are grappling with ontology construction and integration. Continuing from preliminary work presented at the NeDiMAH workshop at DH2012 (Brown et. al. 2012), this paper will cover the following:

1 Review of practical approaches to ontology integration
2 Current ontology use: summary and analysis
3 Philosophical issues at the core of ontology creation, selection, integration, and use

Review of practical approaches to ontology creation and integration

The challenge of creating ontologies goes back at least to Aristotle's Categoriae. Problems surrounding such classification systems were given life in the digital world with the creation of the first databases in the 1960s and issues of structuring and integrating data have been of interest to the computer science and business communities since then. The emphasis from these communities has mostly been on the pragmatics of making the related systems work. Particular attention has been given to the challenge of combining ontologies (for extensive examples of this see Noy et. al. 2005 and Klein 2001), a particularly complicated problem that has roots reaching into problems of normalization and commensurability. In the interest of expanding the repertoire of both conceptual and practical tools available to humanists working with ontologies, we share the results of a survey of these approaches to ontology integration, highlighting those from which humanists stand to benefit most particularly. The survey covers:

- ontology conformity, as seen in projects like Pelagios and NINES;
- mediating ontologies/schemas;
- super ontologies such as SUMO (Suggested Upper Merged Ontology);
- simple ontology joins such as those allowed by owl:import;
- ε-Connections (Cuenca Grau et al. 2011);
- ontology design patterns (Presutti and Gangemi, 2008).

Summary and analysis of current ontology use

Some general statistics are available for ontology use (semanticweb.org 2012; UMBC). These reveal that the Dublin Core (DC) namespace is referenced by 82% of semantic websites, with Friend of a Friend (FOAF) a distant second at 38%. While such figures suggest convergence around particular ontologies, they are also clearly influenced by the nature of existing semantic web

sites, in which archival institutions such as a libraries have exposed large quantities of material. Fully understanding the implications of such figures requires further inquiry, such as what elements of the DC namespace are in use, what elements of the schemas these sites are leveraging from each namespace, and what semantic web elements sites have created themselves. This information is not readily obtained, so we have built a web crawler to collect it. By the time of DH2013 we aim to have a representative sample from which to report on the state of the semantic web in terms of ontology use at a more granular level.

We will combine the results of this web-crawling endeavour with the results of an analysis and assessment of the leading namespace options available. We will highlight probable points of failure when using ontologies in combination, including type propagation and the misuse of symmetric predicates such as subPropertyOf. Our focus, in terms of capturing relations, properties, and spatial-temporal locations, will be on the following ontologies:

Document & Metadata	Space & Time	Relationships	General
• Functional Requirements for Bibliographic Records (FRBR) • Metadata Object Description Schema (MODS) • Dublin Core (DC) • Eprints Type • Open Annotation (OA)	• Event Ontology • OWL Time • Place • Geonames • WGS84 Geo Positioning • FAO Country Profiles • City/Country Lists • Linking Open Descriptions of Events (LODE)	• Friend of a Friend (FOAF) • Encoded Archival Context for Corporate Bodies, Persons, and Families (EAC-CPF) • BIO • Relationship	• RDF/S • XML Schema Definition (XSD) • Web Ontology Language (OWL) • CIDOC • SKOS

Philosophical issues raised by ontology creation, selection, integration, and use

Building an ontology might start out as a simple matter of enabling certain features or capacities but doing so raises philosophical issues that humanists are experienced at recognizing and considering. The most significant of these is the social construction of reality, with the predicted increase in the prevalence of ADHD via the introduction of new criteria for diagnosis in the DSM-5 (Ghanizadeh 2012; Whitely 2010) standing as a tangible example. Closer to home for digital humanists, the DSM-5 also recommends that a new condition known as "Internet Use Disorder" be the subject of further study so that it may be determined whether to fully include in the future (Walton 2012).

The semantic web was in its fetal stages when Bowker and Star worried that, "In the past 100 years, people in all lines of work have jointly constructed an incredible, interlocking set of categories, standards, and means for interoperating infrastructural technologies. We hardly know what we have built. No one is in control of infrastructure; no one has the power centrally to change it" (Bowker and Star 1999, 319). Now that the semantic web is more established it is possible to see that there are opportunities to leverage its features to address such concerns and to hope that a spillover effect might occur. The particular features that we have in mind are the tolerance of the semantic web to ontologies that are incomplete, contradictory, or otherwise 'imperfect', the various levels of granularity that are possible using tools like RDF and OWL, and the transparency that linked online ontologies necessarily provide. We consider these features in light of the insights of the previous sections and conclude that while there is still reason for concern, mindfulness, and caution in constructing ontologies for the semantic web, it still holds promise for cultivating humane practices of classification.

References

Bowker, G. C., and S. L. Star (1999). Sorting Things Out: Classification and Its Consequences. Cambridge: MIT Press.

Brown, S., L. Goddard, and M. Paredes-Olea (2012). RDF for a Dynamic Literary Studies Collaboratory: A Pragmatic and Incrementalist Approach, Proceedings of NeDiMAH workshop on Ontology-based annotation, DH2012, Hamburg, July 2012.

Cope, B., M. Kalantzis, and L. Magee, (eds.) (2011). Towards a Semantic Web: Connecting Knowledge in Academic Research. Oxford: Chandos Publishing.

Cuenca Grau, B., B. Parsia, and E. Sirin (2011). Combining OWL Ontologies Using Econnections. *Web Semantics: Science, Services and Agents on the World Wide Web* 4(1). http://mail.websemanticsjournal.org/index.php/ps/article/view/83 (accessed March 11, 2013).

Ghanizadeh, A. (2012). Agreement Between Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition, and the Proposed DSM-V Attention Deficit Hyperactivity Disorder Diagnostic Criteria: An Exploratory Study, *Comprehensive Psychiatry*. http://linkinghub.elsevier.com/retrieve/pii/S0010440X12001137 (accessed March 11, 2013)

Klein, M. (2001). Combining and Relating Ontologies: An Analysis of Problems and Solutions, *Workshop on Ontologies and Information Sharing*, IJCAI, 1:4-5.

Liu, A. (2012). The State of the Digital Humanities A Report and a Critique, *Arts and Humanities in Higher Education*, 11(1-2): 8-41.

Noy, N. F., A. Doan, and A. Y. Halevy (2005). Semantic Integration. *AI Magazine*, 26(1). http://www.aaai.org/ojs/index.php/aimagazine/article/view/1794 (accessed March 11, 2013)

Presutti, V., and A. Gangemi (2008). Content Ontology Design Patterns as Practical Building Blocks for Web Ontologies, *Conceptual Modeling-ER 2008*: 128-141.

S. L. Star (2010). This Is Not a Boundary Object: Reflections on the Origin of a Concept, *Science, Technology & Human Values* 35(5): 601-617. semanticweb.org. "Ontology", June 13, 2012. http://semanticweb.org/wiki/Ontology .

UMBC Ebiquity Research Group (n.d.). *100 Most Common RDF Namespaces*. http://ebiquity.umbc.edu/blogger/2006/08/23/100-most-common-rdf-namespaces/ (accessed March 11, 2013).

W3C (n.d.). *W3C Semantic Web Frequently Asked Questions*. http://www.w3.org/RDF/FAQ (accessed March 11, 2013).

Walton, A. (2012). Internet Addiction: The New Mental Health Disorder?. http://www.forbes.com/sites/alicegwalton/2012/10/02/the-new-mental-health-disorderinternet- addiction/ (accessed March 15, 2013)

Whitely, M. (2010). Speed up & sit still : the controversies of ADHD diagnosis and treatment. Crawley, W.A.: UWA Publishing.

Notes

1. While this critical mass might be drawn from any discipline, humanists most regularly make a practice of attempting to understand the relationships between humans, cultures, representations and artifacts, all of which are deeply intertwined with the semantic web. This understanding is born from a practice of critical reflection that is importantly stretched between the practical and the theoretical, two poles which must be reconciled for the web to avoid the potential problems raised here.
2. The tool we are developing is the CWRC-Writer. More information about it may be found at http://www.dh2012.uni-hamburg.de/conference/programme/abstracts/cwrc-writer-an-in-browser-xml-editor/ .

Digital Humanities: Egalitarian or the New Elite?

Skallerup Bessette, Lee
lee.bessette@gmail.com
Morehead State University, United States of America

Silva-Ford, Liana
liana.marie.silva@gmail.com
University of Kansas

Risam, Roopika
rrisam@emory.edu
Emory University

Moesch, Jarah
jarah.mo@gmail.com
University of Maryland, College Park

Stalsberg Canelli, Alyssa
astalsb@emory.edu
Emory University

McMillan Cottom, Tressie
tcottom@emory.edu
Emory University

Who is represented within digital humanities and how? This question guides our paper, which focuses on dynamics of representation and exclusion within digital humanities. The field has been subject to a range of criticisms, from its very definition to its relationship to data, building, and hacking. Yet, only recently have scholars begun to raise questions about raced, gendered, and heterosexist assumptions within digital humanities. By exploring these assumptions, we take up the question of how to define digital humanities in ways that are not radically reductive.

Authors Peter Lunenfeld, Anne Burdick, Johanna Drucker, Todd Presner and Jeffrey Schnapp write the following in their new book from MIT Press, *Digital_Humanities*:

> Perhaps, then, the utopian impulse of the Digital Humanities can be characterized as a modality of radically opening discourse to participation for everyone. What if there were no conditions on participation? What if utterances were neither admitted nor denied based on gender, sex, race, ethnicity, language, location, nationality, class, or access to technology? We are not saying that these facticities do not matter or cease to matter in the digital world; instead, we are saying that the utopian element of the Digital Humanities is to at least posit, if not fully enable, a future in which participation is possible for everyone, anywhere, anytime. It would be as if it were possible to bring about a public sphere in which no one was excluded. This is a core human value of the Digital Humanities. (95)

The authors present DH as a "work around" for the issues of race, class, and gender as they have been posited in the "zero-sum game" of the culture wars that have been taking place in the humanities (and elsewhere) for more than

50 years (23-24). We contend that in order for us to even approach the Utopian future of creating a public sphere in which no one is excluded, we cannot work around these issues, but instead confront them head on. The inequities are still too great to ignore.

At the heart of this question is the very definition of "digital humanist." Ernesto Priego has outlined what he calls the new "super-humanist" who can quote literary theory and create DH interfaces from scratch. Are these super-humanists, armed with large research grants, hardware, and human capital, becoming the "face" of not just DH but the humanities in general? If this is, in fact, the presumptive definition of "digital humanist," what roles are available to academics and aspiring academics without access to the resources, support, and training that seem to be necessary to be a successful digital humanist? How are gendered, racialized, and queer bodies represented or not represented in such an articulation of DH? How can we begin to address multiple forms of privilege that proliferate in DH? Does DH challenge existing authority structures that define in-group and out-group status? Is it a tool for dismantling those structures?

With these questions in mind, this paper draws attention to the fraught relationship between DH and those who have been marginalized and silenced within traditional power structures both within and outside of academia. As illustrated by Amy Earhart, in her essay in *Debates in the Digital Humanities*, the promise of open and egalitarian access to materials has largely turned into a funding arms race prioritizing the same texts and projects long favored by academia. In the same collection, Tara McPherson raises questions about the historical separation of technology from studies of race. Accounting for these concerns, who has access and the ability to really do digital humanities? Is digital humanities egalitarian, or is it opening the door to a new elite? Cherie Ann Turpin recently investigate the National Endowment of the Humanities funding numbers to understand who was being funded and found that the vast majority of them went to large, R1 institutions and not HBCUs (Historically Black Colleges and Universities), Primarily Hispanic Institutions, or regional public college, which are often located in rural and/or economically depressed areas. Not only are the kinds of projects being funded not representative, neither are the kinds of institutions, limiting the size, scope, and reach of DH.

The questions of representation within digital humanities that we raise take many forms but center around the tension between macro and microscales of legibility and labor that emerge as scholars define digital humanities. For example, if big data repositories or significant digital archives translated into visualization platforms come to symbolize the work of digital humanities, who is counted and uncounted by absences in big data and silences in the archive? If we foreground technologies and metaphors of visualization and mapping, how do we navigate the imperialist histories that will inevitably be encoded in the structures of digital humanities? We are arguing, as have scholars like Alan Liu, that the collaboration sequence, and its implied hierarchy, needs to be questioned. To borrow and repurpose some phrases from Gayatri Spivak, the active ideology of imperialism provides the discursive field from which the Digital Humanities emerges as both a discipline and a signifier.

Through such definitions of digital humanities, we examine the likely effects on labor, particularly labor performed by women and people of color, who are already plagued with disproportionate service demands in the name of diversity. We will examine where digital humanities fits in with demands on our time made by academic institutions and how we render digital humanities work visible and legible within the broader trajectories of our careers. As put by Julie Flanders, the time and effort that staff who work in DH centers is measured in fundamentally different ways than a "normal" academic. We will discuss further whether digital humanities labor "counts" as teaching, research, or service and propose how those of us in marginal spaces within the academy might best advocate for how digital work is evaluated in our professional portfolios. As such, this paper outlines the issues faced by individuals whose identities render their work marginal or invisible within stringent definitions of digital humanities, identifying solutions for redressing these inequalities.

References

Earhart, A. (2012). Can Information be Unfettered? Race and the New Digital Humanities Canon. In Gold, M. K. (ed), *Debates in the Digital Humanities*. Minneapolis: University of Minnesota Press. 309-318.

Earhart, A. (2012). Recovering the Recovered Text: Diversity, Canon Building, and Digital Studies. *DH2012 Programme*. http://www.dh2012.uni-hamburg.de/conference/programme/abstracts/recovering-the-recovered-text-diversity-canon-building-and-digital-studies/ .

Flanders, J. (2011). You Work at Brown. What Do You Teach? *#alt-academy: A Media Commons Project*. http://mediacommons.futureofthebook.org/alt-ac/pieces/you-work-brown-what-do-you-teach .

Liu, A. (2012). The State of the Digital Humanities: A Report and a Critique. *Arts and Humanities in Higher Education* 11(1): 1-34.

Lunenfeld, P., et al. (2012). *Digital_Humanities*. Cambridge, MA: MIT Press.

McPherson, T. (2012). Why Are the Digital Humanities So White? or Thinking the Histories of Race and

Computation. In Gold, M. K. (ed.), *Debates in the Digital Humanities*. Minneapolis: University of Minnesota Press. 139-160.

Priego, E. (2012). Various Shades of Digital Literacy: The New Digital Divides. *HASTAC.org* http://hastac.org/node/105147 .

Turpin, S. A. (2013). *Digital Humanities: Access and Empowerment*. http://tinyurl.com/mlsvesg

Expanding and connecting the annotation tool ELAN

Sloetjes, Han
han.sloetjes@mpi.nl
The Language Archive — Max Planck Institute for Psycholinguistics, The Netherlands

Somasundaram, Aarthy
aarthy.somasundaram@mpi.nl
The Language Archive — Max Planck Institute for Psycholinguistics, The Netherlands

Drude, Sebastian
sebastian.drude@mpi.nl
The Language Archive — Max Planck Institute for Psycholinguistics, The Netherlands

Stehouwer, Herman
herman.stehouwer@mpi.nl
The Language Archive — Max Planck Institute for Psycholinguistics, The Netherlands

van de Looij, Kees Jan
keesjan.vandelooij@mpi.nl
The Language Archive — Max Planck Institute for Psycholinguistics, The Netherlands

Abstract

The annotation tool ELAN allows for adding time-linked textual annotations to digital audio and video recordings. It is applied in various disciplines within the humanities, with linguistics, sign language and gesture research represented most prominently in its user base. This paper highlights new developments in ELAN with an emphasis on those features that introduced new technological and methodological approaches to analysing both audio/video and derived textual data.

1. Introduction

Annotation of audio and video recordings, be it manual or (semi-)automatic, is a crucial step in many areas of research within the humanities. ELAN[1] , developed at The Language Archive (TLA)[2] /Max Planck Institute for Psycholinguistics, is a tool for manual annotation that is already available for more than a decade and that is applied in various types of projects: language documentation, sign language and gesture studies, psychological and educational behaviour studies etc. ELAN enables users to create multi-levelled, multi-participant, time-linked annotations to one or more media streams, including timeseries streams. Both qualitative and quantitative research is supported; arguably the qualitative oriented use is predominant but the quantitative application is gaining popularity. In this paper we focus on recent developments that improve the workflow of researchers by introducing task oriented modes, expand the scope of the program by implementing a framework for computational annotation creation modules and by connecting to web services that, in a similar way, apply computational techniques to create annotations. ELAN is free and open source software and runs on Windows, MacOS and Linux.

2. The Interlinearization mode and text processing modules in Lexan

One of the recent and still ongoing developments concerns the introduction of the Interlinearization mode. This mode, on the one hand, provides a user interface optimized for the task of adding linguistically relevant layers to an orthographical transcription of the media. Layers for morphological break down, part of speech tags and glossing are part of the common repertoire of documentary linguists (Bow 2003). On the other hand this mode is the hub to Lexan, the extensible framework for annotation and text processing modules. Such modules can perform a variety of tasks, from simple to complex, from word segmentation to interlinearization based on machine learning algorithms. Some modules are expected to produce multiple suggestions for new annotation layers and to improve their suggestions based on interactive user feedback accommodated by the user interface of, primarily, the interlinearization mode. The name "Lexan" indicates that this framework interconnects ELAN with the

TLA multimedia lexicon tool LEXUS[3] . This architecture allows to build and enrich a lexicon while annotating and at the same time to use information in the lexicon in the annotation suggestions process. This combination of NLP (Natural Language Processing) techniques with manual media annotation marks a new line of development in ELAN and brings together technologies that usually seem to develop apart.

For this sort of work other tools are and have been around for a long time and providing interoperability with these tools (often implemented on the level of file format conversion) is highly important for many users.

3. Interoperability with FLEx

The FieldWorks Language Explorer[4] is a prominent example of such tools, therefore import and export facilities for the FLEx format have been implemented and revised with the goal to make repeated transfer of data ("round-tripping") between the tools as seamless as possible. Importing FLEx files was possible since ELAN version 3.8 (2009) but because the FLEx format at that time did not support time alignment and speaker information, an export function was not implemented simultaneously. That has been added recently, after the introduction of the "begin-time-offset", "end-time-offset" and "speaker" attributes at the phrase level of the flextext format (2012, FLEx 7). The import has been updated such that per speaker a group of tiers is created. Additionally efforts are made to retain punctuation and font information where possible. Punctuation elements are on import linked to an ISOcatdata category so that on export these elements receive the correct attribute again. Exporting an ELAN document that is the result of a FLEx import is fairly straightforward. Exporting just any ELAN document to FLEx remains a challenge; where ELAN is very flexible and allows to have any number of tiers without predefined content designation, there is FLEx much more rigid, providing a fixed set of layers of known categories. Resolving the mapping from one to the other is not (always) possible without user intervention.

4. Connecting to web services and online resources

ELAN is a standalone desktop application that in principle works with locally available (local hard drive, local network) resources. Audio and video files are (more and more) often very large, up to several Gb. per file, and high accuracy annotation is still problematic when using media streaming, even in situations with high speed internet connections. For the vast majority of features of ELAN an internet connection is not required, but recently several options have been added that allow the user to connect to online services and resources. In 2008 association of tiers and annotations with ISOcat[5] data categories was introduced and this feature has recently been improved and made more relevant. By default tiers are generic annotation "containers", oblivious of the type of content of the annotations; there are no predefined tiers, e.g. "translation tier" or "gesture phases tier". By associating a tier with a concept registered and described in ISOcat it acquires an explicit content designation.

In the CLARIN-NL SignLin[6] project support for external controlled vocabularies was added, enabling collaborators to share vocabularies over the network (Crasborn and Sloetjes 2010). This feature improves consistency within a team and prevents team members from making (unchecked) changes to the vocabularies. In the context of several CLARIN[7] projects and of the AVATecH[8] project extensions were developed that call web services which produce segmentations and/or annotation content taking audio, video or text, or a combination thereof, as input. The WebLicht [9] tool chaining framework is a core service in CLARIN-D[10] (Hinrichs, et al. 2010) and preliminary support for calling services registered with this framework is now available to users of ELAN. Tiers can be uploaded (in the required XML format, TCF) in order to be processed by well known parsers and taggers; the results are added as new tiers and thus enrich the annotation document.

No matter how useful these web services are or will become, for many field linguists, and other researchers who are working offline a lot, these provisions will not be available. Therefore the core functions will always be independent of online services. For some services, like ISOcat, it is possible to work with a local cache; a selection of categories is stored on the local machine for use while offline.

Figure 1
Preparing a web service call

Most parsers and taggers are only available for a small number of major languages, linguists who study lesser described, let alone, endangered languages usually don't have similar mature, well tested and well trained systems at their disposal. The Lexan approach, stepwise building up "personalized" computational assistance based on the input and feedback of the user, can come to their rescue.

5. Local corpus enrichment and exploration

Though ELAN has been a multiple document application almost from the start, most functions of ELAN allow the user to interact with one, the current active, document. But in recent years more and more functions have been introduced that operate on multiple files e.g. on an entire local corpus. The urge for such functions emerged with the growing number of recordings and transcriptions research teams nowadays are working on (Johnston 2010).

A shortlist of multiple files functions contains creation of transcriptions for a set of recordings, editing the collection of tiers in transcriptions, creating annotations by applying logical operations (AND, OR, XOR) on annotations of selected tiers (Lausberg and Sloetjes 2009), extracting information by executing search queries, generating simple statistics, converting multiple files to and from specific formats etc.

For some types of research assessing the quality of the annotations and the skills of the annotators is crucial. How the inter-annotator reliability best is assessed is still under discussion (Gut 2004; Holle and Rein 2012; Lücking 2011) and the best approach can differ depending on the properties of the data and the focus of the research. A few algorithms for calculating the inter-annotator agreement have been implemented in ELAN and are available for application on multiple files. Especially concerning time-alignment (the segmentation step of the annotation process) there seems to be no generally accepted algorithm for assessing agreement. By offering several alternatives, the choice remains to the user while some of the hassle of exporting data to other tools is taken away from her/him.

6. Conclusion

In this paper we show how researchers working with digital audio/video materials across disciplines can apply new technologies as a result of connections established between ELAN and local or online modules and services. Features that allow to enrich and explore a local corpus are introduced and briefly discussed.

References

Bow, C., B. Hughes, and S. Bird (2003). Towards a general model of interlinear text. In *Proceedings of EMELD Workshop 2003: Digitizing and Annotating Texts and Field Recordings.* Lansing: MI

Crasborn, O., and H. Sloetjes (2010). Using ELAN for annotating sign language corpora in a team setting. In *Proceedings of the 4th Workshop on the Representation and Processing of Sign Languages: Corpora and Sign Language Technologies.* Paris: ELRA, 137-142.

Gut, U., and P. Bayerl (2004). Measuring the Reliability of Manual Annotations of Speech Corpora. In *Proceedings of Speech Prosody,* Nara.

Hinrichs, M., T. Zastrow, and E. Hinrichs (2010). WebLicht: Web-based LRT Services in a Distributed eScience Infrastructure. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10).* Valletta, Malta.

Holle, H., R. Rein Assessing interrater agreement of movement annotations. In Lausberg, H. (ed.), *Neuroges: The Neuropsychological Gesture Coding System.* Berlin: Peter Lang.

Johnston, T. (2010). Adding value to, and extracting of value from, a signed language corpus through secondary processing: implications for annotation schemas and corpus creation. In *Proceedings of the 4th Workshop on the Representation and Processing of Sign Languages: Corpora and Sign Language Technologies*. Paris: ELRA, 137-142.

Lausberg, H., and H. Sloetjes (2009). Coding gestural behavior with the NEUROGES-ELAN system. *Behavior Research Methods, Instruments, & Computers* 41(3):841-849.

Lücking, A., Ptock, S., and Bergmann, K. (2011). Staccato: Segmentation Agreement Calculator. *Proc. of the 9th International Gesture Workshop* held 25-27 May in Athens, Greece.

Notes

1. http://tla.mpi.nl/tools/tla-tools/elan
2. http://tla.mpi.nl
3. http://tla.mpi.nl/tools/tla-tools/lexus
4. http://fieldworks.sil.org/flex
5. http://www.isocat.org
6. http://www.ru.nl/sign-lang/@673229/pagina/
7. http://www.clarin.eu
8. http://www.mpi.nl/avatech

9. http://www.clarin-d.de/language-resources/weblicht-en.html
10. http://www.clarin-d.de

'State of the Art': Negotiating a National Standards-approved Digital Humanities Curriculum

Smithies, James
james.smithies@canterbury.ac.nz
University of Canterbury, New Zealand

Millar, Paul
paul.millar@canterbury.ac.nz
University of Canterbury, New Zealand

Bellamy, Craig
craig.bellamy@versi.edu.au
VERSI, Australia

The University of Canterbury has recently completed development of New Zealand (and Australasia's) first digital humanities degree program that is also standards-approved on a national level. The process required the development of document sets that were submitted for review by the University of Canterbury Faculty, Academic Advisory Committee, Academic Board, the New Zealand Vice-Chancellor's Committee on University Academic Programs (CUAP), the New Zealand Vice-Chancellor's Committee, and the Tertiary Education Commission. Fourteen national and international reviewers, drawn from technology education, information science, computer science, high performance computing and the digital humanities also provided their opinions. The program represents a significant baseline for future digital humanities programs, and the lessons learned during its development are of importance to the broader digital humanities community. Although New Zealand universities operate with basically the same degree of independence in course and program development as universities elsewhere in the world, the requirement to submit all new programs to a national standards body is unusual, if not unique. It may be that the University of Canterbury digital humanities program is the most closely scrutinised example the digital humanities community have seen. This has resulted in a program that is embedded within both the culture of Canterbury, and the national educational policies of NZ. It therefore comes with a higher degree of legitimacy, but also a complex set of stake holders. Moreover, because of the close policy ties between New Zealand and Australia (in education as well as other areas) the program has implications for the Australasian region as a whole.

The implications of national accreditation

Programs and curriculums have pedagogical, methodological, administrative and indeed philosophical issues embedded in them. Their final form reflects not only the 'state of the art' in the discipline in question, but the 'state of the art' as parsed through academic staff, informed (and uninformed) reviewers, institutional context (and necessity), national educational politics, and the shifting sands of methodological and critical best practice. The forces are such that it is quite possible for the final program of study to be quite different from that originally intended, although for obvious reasons the applicants tend to press on regardless, making modifications where necessary but attempting to safeguard the core pedagogical principles wherever possible. This is a process that many digital humanities teams should be expected to go through in the coming years as more institutions attempt to establish programs; it is a period in time when the digital humanities are going to begin to be influenced not only by internal pressure, but external ones such as the need to conform to national educational standards.

Teaching applied and critical DH in the context of standards-approved accreditation

The program will be delivered to fourth year students undertaking their 'Honours' year, a first year of post-graduate study often taken before embarking on more advanced Masters or Ph.D. study. The program was informed by existing programs at Kings College London, University College London, the Open University (Wilks 2011) and the University of California, but the author drew most heavily on theoretical and pedagogical perspectives raised through DH social media and publishing channels over the past five years. A balance has been struck between the 'hack' and 'yak' positions (Cecire 2011; Ramsay 2011; Koh 2012 and others), in the light of what Alexander Reid has suggested is a need for the field to equip students with

a broad "yet undefined digital literacy" (Reid 2012, 354) encompassing both technical and critical skills. The position taken is similar to that espoused by Alan Liu and Andrew Prescott, who argue that tomorrows students and scholars will need to function in a world in which computers are not only ubiquitous, but knowledge itself is a commodity (Liu, 2004; Prescott, 2012). In this sense, the program assumes an ethical imperative to prepare students for work in the post-industrial society that was envisaged by Daniel Bell in 1973, and now forms the basis of both graduate employment structures (Castells and Aoyama, 1994; Aneesh, 2001; Cohen 2010) and tertiary education systems (Donoghue 2008; Brier 2012). In keeping with the core values of the digital humanities community, emphasis has been placed on the development of technical skills that can enhance and extend humanities research activities, and promote awareness of the engineered nature of the digital world.

The program is structured around two core assessment papers: *DIGI 401: Introduction to Digital Humanities* and *DIGI 402: Humanities and New Media*. *DIGI 480: Research Essay* will also be available, to students interested in exploring a topic in detail via a 10,000 word essay. A variety of other (assessment) papers will be rolled out in future years, including Applied Digital Humanities, Digital Literary Studies, and Digital History. Masters and Ph.D. offerings are expected to follow. *DIGI 401: Introduction to Digital Humanities* is modelled on courses in historical method that are well known to History students. The course provides a broad and challenging overview of the digital humanities, organised into History, Theory and Applied modules. Topics include technological determinism, systems theory, materiality and digital forensics, the nature of digital texts, and data visualization. Introductory lectures on TEI and GIS will prepare students for further study in Digital Literary Studies and Digital History. In order to provide students with generically useful programming knowledge an applied module will concentrate on teaching TEI, GIS, Python and use of APIs. Lecturers will be drawn from University of Canterbury's Digital Humanities program, Human Interface Technology Laboratory, Computer Science, Information Systems, and Geography. The aim is to offer the students an overview of tools and methods in the digital humanities, and encourage them to think about how the digital world is engineered. *DIGI 402: Humanities and New Media* is an overt attempt to blend the 'hack' and 'yak' sides to DH as a practice. Students will be strongly encouraged to take DIGI 401 before taking 402 so they have a solid understanding of the technical side to new media culture and politics. Topics in this course include digital modernity, technocracy, cybernetics, knowledge economies, the Internet, open and closed data, open and closed ecosystems. Focus will be placed on both the engineered nature of the digital world, and the concepts required to critique it. Assessment will include traditional essay-based assessment, blog posts, forum posts, and quizzes designed to ensure students are capable of analysing the digital world as an engineered phenomenon.

Pedagogical focus will be placed on graduate outcomes across the program as a whole, and students will be offered opportunities for student exchanges, internship and work experience opportunities. The aim is for graduates to have a blend of traditional humanities-related skills and applied computing skills. They should have an understanding of the moral and ethical issues surrounding digital technologies, the ability to write clear, concise prose, and an understanding of the technical constraints and opportunities provided by digital technologies. Students should be well suited to work in all new media and digital industries, but especially ones requiring a blend of analytical and technical skills. Graduates would be suitable for work in research, relationship management, business analysis, digital archiving, project management, and the creative and cultural heritage sectors. They should be particularly suited to policy analysis positions related to technology and culture, and any position that requires communication across technical and non-technical audiences. The aim is to create a 'porous' educational environment that encourages interaction both inside and outside the university, equipping students with experiences and relationships that can translate into enhanced employment prospects. Inter-disciplinarily will be encouraged, and it is hoped that a DH Commons can be developed to integrate university service support teams in the library and digital media group into the learning experience.

The accreditation process means that, while reflecting the core aims and values of the digital humanities community, the program is also relevant to the pedagogical and strategic aims of the University of Canterbury and the wider New Zealand tertiary education sector. Although challenging, once successfully negotiated the accreditation process effectively embeds the digital humanities into the New Zealand government's long-term education strategy, providing significant pedagogical sanction, integration with the secondary education sector, and a strong platform for future growth. All New Zealand, and undoubtedly Australian, universities aiming to develop digital humanities programs will need to reference the University of Canterbury as a baseline. The implications of this for the development of the digital humanities across Australasia are significant, and (as long as the Canterbury program enshrines core DH aims and values) largely positive.

This paper will provide an overview of the program from intellectual, pedagogical and strategic perspectives in an attempt to share lessons learned with the international DH community, and redress some of the "emphasis on research over teaching" prevalent in the field (Brier 2012,

391). Specific focus will be placed on the implications of the program for Australia and the development of the digital humanities across Australasia as a whole. All program documentation will be made available online so that conference participants have full-text access to the issues being discussed.

References

Aneesh, A. (2001). Skill Saturation: Rationalization and Post-Industrial Work. Theory and Society 30(3). 363–396.

Bell, D. (1973). *The Coming of Post-industrial Society: a Venture in Social Forecasting*. New York: Basic Books.

Brier, S. (2012). Where's the Pedagogy: The Role of Teaching and Learning in the Digital Humanities. in Gold, M. K. (ed), *Debates in the Digital Humanities*. Ann Arbor: University of Michigan Press.

Castells, M., and Y. Aoyama (1994). Paths towards the informational society: Employment structure in G-7 countries, 1920-90. *International Labour Review* 133(1). 5.

Cecire, N. (2011). When DH Was in Vogue; or, THATCamp Theory. *Works Cited*, October 19 2011. http://nataliacecire.blogspot.com/2011/10/when-dh-was-in-vogue-or-thatcamp-theory.html .

Cohen, D. (2008). Three Lectures on Post-industrial Society. Cambridge, MA: MIT Press.

Smart, B.(2010). *Post Industrial Society*. Sage.

Donoghue, F. (2008). *The Last Professors: The Corporate University and the Fate of the Humanities*. 1st edn. New York: Fordham University Press.

Kent, E. F. (2012). What Are You Going to Do with a Degree in That? Arguing for the Humanities in an Era of Efficiency. *Arts and Humanities in Higher Education* 11 (3). (July 1, 2012) 273–284.

Koh, A. (2012). More Hack, Less Yack?: Modularity, Theory and Habitus in the Digital Humanities http://www.adelinekoh.org/blog/2012/05/21/more-hack-less-yack-modularity-theory-and-habitus-in-the-digital-humanities/ . (accessed May 21, 2012).

Liu, A. (2004). *The Laws of Cool: Knowledge Work and the Culture of Information*. Chicago: University of Chicago Press.

Prescott, A. (2012). An Electric Current of the Imagination: What the Digital Humanities Are and What They Might Become. *Journal of Digital Humanities*. http://journalofdigitalhumanities.org/1-2/an-electric-current-of-the-imagination-by-andrew-prescott/ . (accesssed June 26, 2012).

Ramsay, S. (2011). On Building. *Stephen Ramsay*, http://lenz.unl.edu/papers/2011/01/11/on-building.html. (accessed January 11, 2011).

Reid, A. (2012). Graduate Education and the Ethics of the Digital Humanities. *Debates in the Digital Humanities*. Ed. Matthew Gold. Ann Arbor: University of Michigan Press.

Thaller, M. (2012). Controversies Around the Digital Humanities: An Agenda. *Historical Social Research* 37(3). 7–23.

Wilks, L. (2011). *Developing the Digital Humanities at the Open University*. Open University, June 2011.

VizOR: Visualizing Only Revolutions, Visualizing Textual Analysis

Solomon, Dana Ryan
danasolomon@umail.ucsb.edu
UC Santa Barbara, United States of America

Thomas, Lindsay
lindsaythomas@umail.ucsb.edu
UC Santa Barbara, United States of America

Introduction

This paper describes VizOR, a new digital resource currently under development, that visualizes Mark Z. Danielewski's 2006 novel *Only Revolutions*. VizOR is built on top of a MySQL database comprised of the complete text of *Only Revolutions* and is programmed in Python to produce a dynamic, database-driven visualization of the novel. In this paper, we discuss the methods and procedures we are currently developing to create this visualization and highlight the implications of these methods and procedures for theoretical concerns in both digital humanities and media studies. This project, we propose, is both a re-reading of and an exploration of the process of reading Danielewski's novel: as such, it joins the novel in examining the interrelation of human and machine "reading" and "authorship," pointing to a procedural understanding of reading and writing and suggesting that both activities occur across a wide variety of actors and platforms.

Context

The form of Danielewski's novel is unique: the narrative consists of two parallel yet interrelated narratives, one by Hailey and one by Sam, and which one the reader reads

depends on how the reader holds the book. If the reader is reading Hailey's narrative, for instance, she must flip the book upside down to read Sam's (and vice versa). Apart from the main narrative, each page also contains what we call a chronology section — a date with a list of historical people and events associated with that time or date, which Danielewski crowdsourced from his fans while writing the novel. Thus, each page is divided into four sections: Sam's narrative, Sam's chronology (which runs from November 22, 1863 to November 22, 1963), Hailey's narrative, and Hailey's chronology (which runs from November 22, 1963 to June 19, 2063).

In addition to its formal innovations, Only Revolutions is interesting for this project because it proliferates numbers, playing with the boundaries between "data" and "narrative." For example, the numbers 360 and 8 and their factors and multiples are particularly important. Each narrative is 360 pages long, and each narrative and chronology section on each page contains 90 words (90 X 4 = 360). Furthermore, the narrative is divided up into sections of eight pages each, and the number of lines in each narrative section decreases, at regular intervals, from 22 to 14 as the reader progresses through these sections (22 - 14 = 8). "H," for Hailey, is the eighth letter in the alphabet, and "S," for Sam, is the eighth letter from the end of the alphabet; "Mark Z. Danielewski" has 16 letters; and Sam and Hailey are described as being "allways sixteen." There are many, many more examples we could cite here. In this way, Only Revolutions encourages readers — human and machine alike — to count, and count again, the many different numbers that emerge from its pages.

In *How We Think: Digital Media and Contemporary Technogenesis*, N. Katherine Hayles includes a coda featuring visualizations of Only Revolutions produced with Google Maps. Collaboratively designed with Allen Beye Riddell, these three visualizations trace the geographical "place-names" of Sam and Hailey as they travel throughout the story, and then layer the two to create a composite map of the characters' movement through the text (243-244). The resulting visualization shows that Hailey and Sam take similar paths across the map and that their overall directionality is nothing if not inconsistent. They move on a whim, together, wherever their overwhelming affection for one another points them. On one hand, VizOR is a response to Hayles's map-based visualization and is concerned with the assumptions underpinning her choice of visualization platform. On the other hand, the project is addressing a larger trend within the digital humanities: the increased prevalence of data visualization as a mode of literary interpretation.

Due in large part to its often powerful and aesthetically pleasing visual impact, relatively quick learning curve, and overall "cool," the practice of visualizing textual data has been widely adopted by the digital humanities. This prevalence is evidenced by, for instance, the high frequency of the term "data visualization" in the 2011-2012 Digital Humanities conference abstracts as well as the 2011-2013 Modern Language Association digital humanities panel descriptions. If the first wave of large-scale database projects in the digital humanities is exemplified by the practices of digitizing texts, constructing archives, and determining best practices for digital preservation, then the practice of data visualization is emblematic of the second wave of projects devoted to mining and interpreting this newly available data.

VizOR is influenced by the thinking of scholars and practitioners like Franco Moretti, Matthew Jockers, Jeremy Douglass, and Lev Manovich, as well as by visualization projects like UC San Diego's "Cultural Analytics" initiative and Stanford's "Gephi" visualization engine. Like other critical DH projects, VizOR is not only interested in engaging with literature via data visualization, but also in performing this engagement to ask what is at stake in this new mode of interpretation, both in terms of the individual scholar and the digital humanities as a field.

Project Description

The database is designed to be highly flexible, and this architecture allows us to enact the same linguistic layering that occurs while reading the text. We can query a specific word of a particular character's narrative or chronology, the text from a specific line of a character's narrative or chronology on a specific page, or the narrative or chronology text from a whole page for a specific character. Querying the database in this way allows us to instantly compare the words and phrases used by one character with those used by the other character on the same page and in the same position. For example, on Sam's page seven, line five, we see "Gold Eyes with flecks of Green," while the words in the same position on Hailey's page seven, line five read,"Green Eyes with flecks of Gold" (Danielewski). VizOR surfaces this kind of reading through its very design, allowing readers to see the similarities and, perhaps more importantly, the differences between the narratives.

The visualization is currently still under development using the Python programming language. Python was selected as the programming language for VizOR due to its general flexibility and its ability to produce dynamic, database-driven visualizations. However, static mock-ups have been designed using Adobe Illustrator and are included at the end of this document.

The visualization reproduces the image of the page numbers in the novel, the two small rotating circles enclosed within the larger circle. In the visualization, however, each of the two smaller circles represents the narratives of

Hailey or Sam. All three circles are comprised of absent centers with the text literally expanding outward. The larger circle is comprised of the chronological headings and entries. Clicking on one of the text strings in either narrative will query the database for the corresponding narrative's string. This correspondence, as well as the inclusion of the Boolean double pipe (symbol for "or") as link, is shown in the mock-ups at the end of the document. This, in effect, mirrors the layering that occurs during the act of reading. The visualization, though, has the ability to position the corresponding lines on the same plane at the same time, a phenomenon in the novel that is always already delayed or fading away. Rather than attempting to flip the book fast enough to see both sides of the coin, so to speak, VizOR attempts to freeze the text at each moment of mirroring. This moment of pause opens up the possibility of interpretation without disrupting the line's relationship to other lines or removing it from its context.

Users can navigate the visualization in a number of ways, hovering over any of the terms to magnify chosen words or lines. Once users click on a given line, the circles will rotate to realign the corresponding terms. Further, rotating the encircling chronology forward or backward in time will result in the rotation of the two inner narratives, mirroring the motion of the flip-book page numbers of the print text.Users can navigate the visualization in a number of ways, hovering over any of the terms to magnify chosen words or lines. Once users click on a given line, the circles will rotate to realign the corresponding terms. Further, rotating the encircling chronology forward or backward in time will result in the rotation of the two inner narratives, mirroring the motion of the flip-book page numbers of the print text.

The finished form of the visualization will be searchable and will also contain external hyperlinks. In keeping with the novel's data-driven construction, produced in some ways by a crowd-sourced or "collaborative" author, the visualization produces a similarly data-driven reading experience. The goal here is to mirror the outward push of the novel, its awareness and incorporation of external databases like online encyclopedias, and the uniquely distributed reading experience of excitedly setting the book down to search for one of its vague historical entries.

Conclusion

In the name of this speed and efficiency, however, these technologies often strip data of its context and idiosyncracies, creating what Tara McPherson has called "a system of interchangeable equivalencies" (35). VizOR, through its emphasis on the distributed processes of reading and writing across different technologies and media, pushes against this seamless homogenization. By highlighting particular, idiosyncratic moments of reading, we hope to activate the messy, "seamy" place where data meets narrative.

References

Barthes, R. (1977). Rhetoric of the Image. *Image, Music, Text.* New York: Hill and Wang.

Borner, K. (2003). Visualizing Knowledge Domains. *Annual Review of Information Science & Technology.* 37. Medford, NJ: Information Today, Inc./American Society for Information Science and Technology. 179-255.

Cartwright, L. (2009). *Practices of Looking: An Introduction to Visual Culture.* Oxford: Oxford University Press.

Danielewski, M. (2006). *Only Revolutions.* New York: Pantheon.

Drucker, J. (2008). The Virtual Codex from Page Space to E-space. In Schreibman, S., and R. Siemens, (eds). *A Companion to Digital Literary Studies.* Oxford: Blackwell. http://www.digitalhumanities.org/companionDLS/ (accessed September 21, 2012.

Hayles, N. K. (2012). *How We Think: Digital Media and Contemporary Technogenesis.* Chicago: The University of Chicago Press.

Lima, M. (2011). *Visual Complexity: Mapping Patterns of Information.* New York: Princeton Architectural Press.

Manovich, L. and J. Douglass (2007). Cultural Analytics. UC San Diego Software Studies Initiative. http://lab.softwarestudies.com/2008/09/cultural-analytics.html (accessed: September 21, 2012.)

McPherson, T. (2012). U.S. Operating Systems at Mid-Century: The Intertwining of Race and UNIX. In Nakamura, L. and Chow-White, P. A., (eds). *Race and the Internet.* New York: Routledge.

Terras, M. (2008). *Digital Images for the Information Professional.* London: Ashgate.

Tufte, E. (1983). *The Visual Display of Quantitative Information.* Cheshire, CT: Graphics Press.

Vesna, V. (2007). *Database Aesthetics: Art in the Age of Information Overflow.* Minneapolis: University of Minnesota Press.

Yau, N. (2011). *Visualize This: The FlowingData Guide to Design, Visualization, and Statistics.* Hoboken: Wiley Press.

Theorizing Data Visualization: A Comparative Case-Study Approach

Solomon, Dana Ryan
danasolomon@umail.ucsb.edu
English Dept. UC Santa Barbara, United States of America

Data visualization is a digital technique for organizing, representing, and interpreting information visually. Etymologically, data visualization is the process wherein "that which is given" is "made visible to the eye," translating a particular collection of statistics, values, or words into a visual diagram, design, or architecture. Due to its efficient (computational) and scalable workflow, data visualization lends itself particularly well to working with large amounts of raw or unprocessed information, which is typically collected and organized into datasets and stored in electronic collections called databases. This project takes as its critical object the application of data visualization as a means of textual analysis in the digital humanities, examining how a technology with a long-running history in corporate business contexts and the social and STEM sciences might affect the practice of literary scholarship, particularly in terms of reading and interpretation.

Due in large part to its often powerful and aesthetically pleasing visual impact, relatively quick learning curve, and overall "cool" the practice of visualizing textual data has been widely adopted by the digital humanities. This prevalence is evidenced by, for instance, the high frequency of the term "data visualization" in the 2011-2012 Digital Humanities conference abstracts as well as the 2011-2013 Modern Language Association panels related to the digital humanities. If the first wave of large-scale database projects in the digital humanities is exemplified by the practices of digitizing texts, constructing archives, and determining best practices for digital preservation, then the practice of data visualization is emblematic of the second wave of projects devoted to mining this new data. The NEH-funded "Digging into Data" granting program, a yearly challenge that asks how the notion of scale affects humanities research, has specifically supported this practice of engaging with huge databases and archives.

This paper directly engages with the (oft-overlooked) notion that data visualization is in fact an argumentative, non-neutral process and asks: what is and is not visualized, how are visualizations produced, how do aesthetics factor into this discussion, why and how are digital humanists using this technology, and how can data visualization be contextualized historically, materially, and politically. The paper does so by offering focused case studies of two specific data visualization environments:

1.) IBM's *Many Eyes* — This case study involves close-reading mission statements, web content, user-instructions, and sample and showcase datasets and visualizations. This case-study helps to construct and illuminate the project's corporate identity. IBM's history, after all, is deeply intertwined with the invention of the punch-time clock and other aspects of contemporary (Taylorist) business culture and embodies the corporate values of efficiency and time management. Further, due to its work with Swiss-style graphic designer Paul Rand on the company's logo and minimalist aesthetic identity, IBM is emblematic of the intersection of early computing and mid-century graphic design. The company is unique in that it managed to synthesize cutting edge technology, profit-driven corporate culture, and "cool" graphic design, an accomplishment that I argue contributes to the prevalence of *Many Eyes* data visualizations in the digital humanities, especially in projects produced by individuals who are making their first forays into the field.

2.) Alan Liu's NEH-funded *Research-oriented Social Environment (RoSE)* — This section contrasts *Many Eyes* with RoSE, a data visualization platform with a different material and cultural history. I use my insight here as a humanities developer on the project team to explain how RoSE privileges critical approaches to design, collaborative development, and data transparency. In fact, the data itself is produced collaboratively through cooperation with individual users and large-scale institutional partners, including Project Gutenberg and the SNAC team at the University of Virginia. Finally, unlike *Many Eyes*, RoSE hosts a supplementary website that provides details about the project's goals, developers, and potential limitations.

The project complements other recent and historical moves to examine data visualization, particularly as it is embedded within a longer engagement with graphic design, information aesthetics, and visual rhetoric. The work of theorists of information design and large-scale data analysis, including Edward Tufte, Andrea Lau, and Lev Manovich, has been crucial in formulating an informational aesthetic. This aesthetic framework extends from the mid-twentieth century introduction of the so-called Swiss graphic design, known for its emphasis on minimalism and reliance on the "grid," to new media art practice. Whereas much of the research to date has dealt with database and visualization design, my intervention is to engage with the importation and application of data visualization in the digital humanities as such. Previous research on data visualization in the field has been concerned primarily with the development of visualization software and the

technology's exciting new capabilities, rather than how such development might in fact transform literary scholarship. This project takes a critical step back from these discussions in order to consider how visualization techniques, tools, and technologies might in fact transform literary scholarship, what is at stake in their instrumentalized use, and how humanistic modes of critical engagement might be applied to them.

The focus here is on differentiating the cultural genealogies of each platform to shed light on their unique ideological foundations. The paper examines the corporate lineage of *Many Eyes*, while constructing a parallel but alternative genealogy for RoSE, one rooted more in scholarship, new media art practice, and collaborative knowledge production than profit-driven commercial activity. The goal of the paper is not to critique the corporate affiliation of *Many Eyes* as such, but to ask what might be at stake in the implementation of a technology that has its roots outside of the humanities, specifically a technology that has existed for half a century in other contexts and contains its own embedded goals, methodologies, and ideological underpinnings.

References

Borner, K. (2003). Visualizing Knowledge Domains. *Annual Review of Information Science & Technology,* 37, Medford, NJ: Information Today, Inc./American Society for Information Science and Technology. 5. 179-255.

Fitzpatrick, K. (2011). *Planned Obsolescence: Publishing, Technology, and the Future of the Academy*. New York: New York University Press.

Gold, M. (2012). *Debates in the Digital Humanities*. Minneapolis: University of Minnesota Press.

IBM Research (2012). *Many Eyes Data Visualization Platform*. 2007-present. URL: http://www-958.ibm.com/software/data/cognos/manyeyes/ (accessed 15 September 2012).

Lau, A. and A. Vande Moere (2004). Towards a Model of Information Aesthetics in Information Visualization. http://web.arch.usyd.edu.au/~andrew/publications/iv07.pdf (accessed 15 September 2012).

Lima, M. (2011). *Visual Complexity: Mapping Patterns of Information*. New York: Princeton Architectural Press.

Liu, A. (2004). *The Laws of Cool: Knowledge Work and the Culture of Information*. Chicago: University of Chicago Press.

Manovich, L., and J. Douglass (2007). Cultural Analytics. UC San Diego Software Studies Initiative. 2007-present. URL: http://lab.softwarestudies.com/2008/09/cultural-analytics.html (accessed 18 September, 2012).

Mirzoeff, N. (2011). *The Right to Look: A Counter History of Visuality*. Durham: Duke University Press.

Moretti, F. (2007). *Graphs, Maps, Trees: Abstract Models for Literary History*. London: Verso.

Nowviskie, B., and J. Unsworth (1999). *Is humanities computing an academic discipline? An interdisciplinary seminar*. University of Virginia.

Ramsay, S. (2011). *Reading Machines: Toward an Algorithmic Criticism*. Chicago: University of Illinois Press.

Terras, M. (2008). *Digital Images for the Information Professional*. London: Ashgate.

Tufte, E. (1983). *The Visual Display of Quantitative Information*. Cheshire, CT: Graphics Press.

Vesna, V. (2007). *Database Aesthetics: Art in the Age of Information Overflow*. Minneapolis: University of Minnesota Press.

Yau, N. (2011). *Visualize This: The FlowingData Guide to Design, Visualization, and Statistics*. Hoboken: Wiley Press.

XQuery databases for language resources in the IAIA and UyLVs Projects

Sperberg-McQueen, Michael
cmsmcq@acm.org
Black Mesa Technologies LLC, United States of America

Dwyer, Arienne M.
anthlinguist@ku.edu
Institute for Digital Research in the Humanities, University of Kansas

This paper reports on work done in the Interactive Inner Asia (IAIA) and Uyghur Light Verbs (UyLVs) projects[1] on two problems arising in the construction of XML query interfaces for linguistically annotated corpora in Central Asian languages: (1) providing both simple search interfaces for casual users and open-ended interfaces for specialists, without security vulnerabilities; and (2) using XQuery for search patterns involving precedence relations (on sequences of siblings) and not only dominance (containment) relations.

XML query interfaces are not the focus of either project, just means to an end. The immediate purpose of both projects is the study of language contact, language areas, and the development of particular linguistic structures; a

broader aim is the creation of annotated corpora useful both for comparative and historical linguistic research and for ethnological, folkloric, literary, historical and other study.

1 Current corpus format and user interfaces

Although the goals of the two projects differ, the IAIA and UyLVs projects share a common corpus format. Transcriptions of audiovisual recordings and transliterations of manuscripts are presented in multiple tiers: one tier holds the text in a practical orthography, one in the International Phonetic Alphabet (IPA); other tiers contain a morphological segmentation, part-of-speech labels, an interlinear gloss, free translations in English, Russian, German, Chinese, and/or Tibetan, and comments (not all tiers are present for all utterances).

In the current version of our XML corpus format (Sperberg-McQueen 2012), a sample Uyghur sentence takes the following form:

```
<s ref="152">
  <orth>un öjigɛ čaŋ bolup ketti</orth>
  <seg>un öj-i-gɛ čaŋ bol-up#ket-t-i.</seg>
  <pos>N N-POSS3-DAT N Vi-CNV#LVV-PST.dir-3s</pos>
  <ilg>flour room-POSS3-DAT dust become-CNV#KET-PST.dir-3s</ilg>
  <gloss lang="eng">The flour scattered in the room
    like a cloud of dust. </gloss>
</s>
```

Future versions will replace the implicit alignment of the SEG, POS, and ILG tiers with an explicit alignment to simplify validation and searching (Dwyer / Sperberg-McQueen 2013).

2 Sample queries

Simple searches (e.g. for light verbs, coded “LV” in the pos tier) are useful for the project. But linguistically motivated queries easily become complex. Like those of many corpora, our POS tier does not encode grammatical relations (including Subject), so to study the co-occurrence of light and main verbs with the subjects of their sentences, we search for types of noun (tagged N, PN, etc.) in certain positions and with certain predicted inflections, in order to identify subjects. In Uyghur, a subject or agent is likely to be the leftmost uninflected noun (e.g. Bu in the sentence “Bu [Alimdin kelgen xet] PN [Npr-ABL Vi-PST.REL N] 'This is the letter from Alim',”) even when the subject is a relative clause (e.g. xet, in: “[Alimdin kelgen xet] uzun iken. [Npr-ABL V-PST.REL N] AJ COP.indir '[The letter from Alim] seems to be long.')”.

To exclude adjuncts (like after lunch in “After lunch, Mahire took a nap”), which are potentially the leftmost N, we search for the leftmost N (excluding the string "NUM"), and excluding N-*, N(*) POST and then sort the results by the surface form of the matrix verb or light verb.

A third query searches for patterns across different languages in the IAIA project. Both SE Monguor and Baonan have calqued a postposed indefinite article nege~nige ‘one’ from Tibetan zeg ‘one’. We've seen instances of postposed yi-ge (yi ‘one’ (Chinese) + ge ‘classifier’ (Chinese) in the same postnominal position, at least in mjg-se. We might hypothesize that that yige is the Mandarin calque equivalent of the Mongolic nege calque, and that both originate in Tibetan zeg. The presence of yige is thus a good indicator of a strong Mandarin influence on the language. To test the hypothesis, we will want to search the segmentation layer for the morphemes yi and -ge in adjacent positions, or alternatively for adjacent pairs of the POS tags NU-CL.

3 Supporting open-ended queries

The queries above illustrate a common pattern. The first is trivial to express in XPath or XQuery:

```
//pos[contains(.,'LV')/..
```

or alternatively

```
//s[contains(pos,'LV')]
```

Similar queries will be necessary for other POS tags; the query “Find all sentences containing a morpheme tagged X” (for any POS-tag X) will be useful both to the project team and to casual outside users of the collection.

Not all users will know enough XPath to formulate such queries on the fly, however, so it would be helpful to provide query interfaces for this and similar queries which do not require the use of a query language. The usual solution to this common problem is to make some pre-defined searches available in some fill-in-the-blank or point-and-click interface where users can select a pre-defined query and select its parameters from a list. In our full paper, we will show the search forms in use in the IAIA and UyLVs projects for this kind of canned search.

For intensive use of the collection, however, such fixed-parameter searches will not suffice. As the other queries above illustrate, intensive use of any collection will require new ad-hoc searches, not predefined ones, and arbitrary

Boolean combinations of basic searches. How can projects support such open-ended queries?

One easy solution is to allow users to formulate queries in the underlying query language (for these projects, XQuery; in the case of relational data, SQL). Feeding raw user data to a live XQuery interpreter, however, represents a sizable security hole and cannot be advised.[2] We discuss other ways to provide open-ended searches:

- Define an acceptable ('safe') subset of XQuery. If the user input is in that subset, hand it to the XQuery engine for evaluation; otherwise, hand it back to the user for correction (or conclude that the user is an attacker and shut down).
- The safe subset can be checked by:
  - a full context-free parser for the subset, or
  - regular expressions which capture 'regular approximations' of the subset; we will generate these automatically from the subset definition and execute them with standard regular-expression evaluators.
- Define a new open-ended query language for advanced users, using an ad-hoc character-based syntax, then write (a) a parser for the language, and (b) a translator into XQuery, for communication with the underlying XQuery engine.
- Implement some existing query language (or a subset); candidates include Annis 2 (Zeldes 2012); Arras (Smith 1985); CSS Selectors (); the DynaText query language (EBT 1996), (Silicon Graphics 1995); the ‘extended pointer notation’ of TEI P3 and P4 (ACH/ACL/ALLC 1994), (ACH/ACL/ALLC 2001-2004); the query language of Sara / Xaira (Burnard, n.d.); and subsets of XPath 1.0 or 2.0 (W3C 1999), (W3C 2010).
- Define a query language using an XML syntax, with a user interface defined in XForms to allow the user to formulate complex queries (roughly analogous to the 'query builder' interfaces familiar from search tools like Xaira [Burnard, n.d.] and EXMARaLDA [Schmidt 2010]).

4 Searching on siblings in XQuery

A second class of challenge is illustrated by the third query described above. In building our corpora we have, like many others before us, focused our attention initially on segment-by-segment annotation of the data. We have not built parse trees for our sentences; we are building corpora, not tree banks. Automatic parse-tree generators are widely available for ‘major’ languages, but for the non-standardized, unwritten, and ‘non-major’ languages at the center of our work, such tools will become possible only as a result of projects like ours. In the absence of XML representations of full parse trees, subjects, agents, etc., our users will need to search for grammatical relations by using surrogates. As illustrated above, a search for a subject will need to be reformulated as a search for a particular pattern in the sequence of anotated morphemes in a sentence.

Sequence searching is a particular challenge for linguistic resources. XQuery engines are typically very good at building indexes for Boolean combinations of context-sensitive searches (find X within Y where Z and (W or V) ...), but sequence searches are much trickier to support. By a 'sequence search' we mean (for example) a search for morphemes M1, M2, M3 in that order, with M1 matching this pattern, M2 matching that pattern, and M3 the other pattern, and M1 adjacent to M2 and M3 anywhere later in the same utterance. Some of our work will focus on whether we can construct indexes manually to help make such queries faster or more convenient. Sequence searches are likely to be of interest for corpus linguists in any area, as well as for people interested in tight control of full-text phrase searches.

In the full paper, we survey some existing work in this area (e.g. in the Prague Markup Language [Pajas 2010] and in corpus search tools like EXAKT [Schmidt 2010] or Sara and Xaira [Burnard, n.d.]), and describe our work on supporting sequence pattern searches in XQuery. In so doing, we also illustrate the balance between the inclusion of syntactic annotation and ease of querying.

References

Association for Computers and the Humanities, Association for Computational Linguistics, and Association for Literary and Linguistic Computing. (1994). Extended Pointers, section 14.2 of *Guidelines for Electronic Text Encoding and Interchange (TEI P3)*. Sperberg-McQueen, C. M. and Burnard, L. (eds), Chicago: Oxford: Text Encoding Initiative. http://quod.lib.umich.edu/cgi/t/tei/tei-idx?type=pointer&value=SAXR

Association for Computers and the Humanities, Association for Computational Linguistics, and Association for Literary and Linguistic Computing. (2001-2004). Extended Pointers’, section 14.2 of *Guidelines for Electronic Text Encoding and Interchange (TEI P4)*, Sperberg-McQueen, C. M. and Burnard, L. (eds.) XML conversion by Bauman, S., L. Burnard, S. DeRose, and S. Rahtz: Text Encoding Initiative. http://www.tei-c.org/Vault/P4/doc/html/SA.html#SAXR .

Burnard, L. All about Xaira. Oxford: OUCS, n.d. http://projects.oucs.ox.ac.uk/xaira/

Dwyer, A. M., and C. M. Sperberg-McQueen (2013). The Uyghur Light Verbs and Interactive Inner Asia Corpora.

Lawrence, KS: Interactive Inner Asia Project: Uyghur Light Verbs Project. http://uyghur.ittc.ku.edu/2013/tsd/UyLVs-IAIA-corpora.xml .

EBT (Electronic Book Technologies). (1996). 'Searching' In *DynaText Publishing: Document Preparation*. Providence: EBT.

Pajas, P. (2006-2010). The Prague Markup Language (Version 1.1). Prague: Institute of Formal and Applied Linguistics, Faculty of Mathematics and Physics, Charles University. http://ufal.mff.cuni.cz/jazz/PML/doc/ .

Schmidt, T. (2010). EXMARaLDA EXAKT Manual Version 1.0. Hamburg. http://www1.uni-hamburg.de/exmaralda/files/EXAKT_Manual.pdf

Silicon Graphics, Inc. (1995). IRIS InSight DynaWeb User's Guide. Silicon Graphics. http://techpubs.sgi.com/library/dynaweb_docs/0620/SGI_EndUser/books/IIDWeb_UG/sgi_html/index.html

Smith, J. B. (1985). Arras User's Manual. *TR* 85-036. Chapel Hill: University of North Carolina Department of Computer Science. http://www.cs.unc.edu/techreports/85-036.pdf

Sperberg-McQueen, C. M. (2012). PixCor 1.1 Tag-set documentation for Project-internal corpus vocabulary The Uyghur Light Verbs and Interactive Inner Asia Corpora. Lawrence, KS: Interactive Inner Asia Project: Uyghur Light Verbs Project. http://uyghur.ittc.ku.edu/2012/tsd/pixcor.v1.1.tsd.xml

van der Vlist, E. (2011) XQuery injection: Easy to exploit, easy to prevent in *Proceedings of Balisage: The Markup Conference 2011*, 'The Markup Conference' held 2011 at Prague and Montreal. In Balisage Series on Markup Technologies, vol. 7 (2011). doi:10.4242/BalisageVol7.Vlist02. http://www.balisage.net/Proceedings/vol7/html/Vlist02/BalisageVol7-Vlist02.html ; slides from a presentation of an earlier version of the material XML Prague 2011 at http://archive.xmlprague.cz/2011/presentations/eric-vdv-xquery-injection/xquery-injection.xhtml

World Wide Web Consortium (W3C). (1999). XML Path Language (XPath) Version 1.0. In Clark, J. and DeRose, S. (eds), *W3C Recommendation* 16 World Wide Web Consortium. http://www.w3.org/TR/xpath .

World Wide Web Consortium (W3C). (2010). XML Path Language (XPath) 2.0 2nd edn. In Berglund, A. (ed), *W3C Recommendation.* World Wide Web Consortium. http://www.w3.org/TR/xpath20 .

World Wide Web Consortium (W3C). Selectors Level 3, In ed. Çelik, T. et al. *W3C Recommendation.* World Wide Web Consortium. http://www.w3.org/TR/css3-selectors/ , (accessed 29 September 2011).

Zeldes, A. ANNIS: User Guide - Version 2.2.1. Potsdam: SFB 632 Information Structure / D1 Linguistic Database. Potsdam: Humboldt Universität zu Berlin und Universität Potsdam https://launchpadlibrarian.net/102488706/ANNIS_User_Guide_2.2.1.pdf (accessed 18 April 2012).

Notes

1. The IAIA and UyLVs projects are supported by the U.S. National Science Foundation (NSF-BCS 1065524, 1053152). Much of the data for IAIA stems from earlier work sponsored by a Fulbright-Hays and a Volkswagen Foundation DOBES grant.
2. The problem is that SQL injection is widely known as an attack vector for machines on the open net; XQuery injection is also possible and has been documented (van der Vlist 2011).

Extraction and Analysis of Character Interaction Networks From Plays and Movies

Suen, Caroline
cysuen@stanford.edu
Stanford University, United States of America

Kuenzel, Laney
laney.kuenzel@gmail.com
Stanford University, United States of America

Gil, Sebastian
sgil@stanford.edu
Stanford University, United States of America

1. Introduction

Due to recent efforts to digitize literary works, researchers have been able to perform meaningful large-scale analyses of millions of texts and reach meaningful conclusions about literature, language, and culture using statistical analysis. This approach is powerful, but frequently ignores subtleties in literary works, reducing complex texts to bags of words. Literary theorists take a different approach, performing in-depth qualitative studies examining plot intricacies and character interactions. Unfortunately, such deep analysis does not scale well due to human time constraints.

In our project we combine these two approaches to literary analysis, allowing us to benefit from the advantages of both. More specifically, we develop and apply methods for automatically extracting character interaction networks from works of entertainment and use properties of the resulting networks to draw conclusions about these works.

There are three main components:

(1) Extracting character interaction networks as weighted graphs, with characters as nodes and interaction scores as edges
(2) Computing informative properties (e.g., clustering coefficient) of the resulting networks
(3) Using those properties to answer broad questions about the works (e.g., whether different media types are characterized by distinctive interaction networks) by constructing machine learning classifiers.

2. Related work

As mentioned earlier, most computational literary analysis has been at the word level. There are, however, several exceptions. Most notably, Elson et al. (Elson et al. 2010) effectively utilized dialogue interactions in sixty 19th century literary works to form social networks and make interesting discoveries about a particular genre. Other researchers used network theory to analyze small groups of texts, such as Hamlet (Moretti 2011), Greek tragedies (Rydberg-Cox 2011), Shakespeare (Stiller and Hudson 2011), and Marvel comics (Alberich et al. 2002). These studies were all relatively narrow in focus, leading to valuable discoveries about a small number of texts. More recently, C.-Y. Weng et al. (Weng et al. 2009) proposed a network extraction method for movies and T.V. shows based on co-occurrence, successfully identifying lead roles and other attributes for several movies.

Overall, previous work primarily focused on using character interaction networks to improve understanding of individual texts or movies. We feel humans already do a very good job—better than computers—of analyzing small collections of works; our main limitation is insufficient brainpower to simultaneously analyze and compare hundreds or thousands of works. As such, we are interested in conducting a large-scale study of character interaction networks for diverse works of entertainment. Our goal is not to examine literature from a specific time period or a particular film's plot, but rather to discover sweeping trends in literature and movies across genres and over time.

3. Methodology

3.1 Building Networks

We focused on play and movie scripts because their structured format is well suited for systematically detecting interactions between characters. We obtained scripts and relevant metadata from a variety of sources (Internet Movie Script Database (2011); Project Gutenberg (2011); The Complete Works of Shakespeare (2011); EOneill.com EText Archive (1999); Read Plays Online-Read Print (2011); The EServer Drama Collection (2011); Rotten Tomatoes (2011); Robnik-Sikonja and Kononenko (1997), automating the process with Python scripts. For consistency, we then converted all data into a standardized intermediate format using more regular expressions, and a blacklist of non-verbal action commands (e.g. "fade in"). In total, we extracted 173 plays and 580 movie scripts.

We experimented with four extraction algorithms for constructing character interaction networks. Our first approach, used by Weng et al. (Weng et al. 2009), defined the interaction score for two characters as the number of scenes in which both appear. Our second algorithm extended this concept, incorporating the number of lines spoken in each scene. Unfortunately, many scripts had long scenes, resulting in falsely high interaction scores between two characters in different parts of the same scene.

We then used what we call the *Closeness* approach to consider an interaction to have occurred between two characters only when they have spoken nearby lines in the same scene, increasing their scores by an amount linearly decreasing with increased distance. Our fourth and final algorithm weights interactions by the total number of words exchanged.

3.2 Property Calculation

For each character interaction network, we computed the following network properties, which represent different concepts in literary works:

- **Average clustering coefficient**: how much groups of characters tend to cluster together
- **Single character and relationship centrality**: how much the work focuses on a single character above all others
- **Single relationship centrality**: how much the work focuses on a single relationship between characters above all others
- **Top character weight variance**: whether the group has a large group of similarly prominent characters or a few main characters and many less important roles

- **Top relationship strength variance**: whether relationships are emphasized roughly equally, or if there is an emphasis on a select few
- **Entropy of node degrees and edge weights**: an alternate approach to quantifying the spread in the distribution of character and relationship importance
- **Mean and variance of top character relationship strengths**: whether the work has one or several main storylines
- **Percentage of existing edges**: an alternate approach to determining number of storylines
- **Betweenness centrality — maximum, difference, and entropy**: another alternate method of determining the relative importance of main characters
- **Number of characters**: used as a final feature in our classifiers

3.3 Classification

We used our network properties as features in binary classifiers for various media aspects:

Media type: plays or movies
Date of movie: before or after 2000
Date of play: before or after 1800
MPAA rating
Audience and critic ratings
Single genre (e.g. romance or not)
Between genres (e.g. romance or horror)
Author (e.g. Shakespeare or George Bernard Shaw)

We experimented with logistic regression classifiers and decision trees, because these classifier types easily allowed us to understand how features were being used to arrive at predictions. We used the Orange library for Python, normalized our features, used k-fold cross validation to test our classifiers, and used the Relief algorithm [14] for top feature selection.

Because two classification classes did not always have the same number of examples, classification accuracies were sometimes misleadingly high even for poor classifiers. Thus we used area under the curve (AUC) as our primary performance metric.

4. Results

We found logistic regression to have higher AUC's for 26 of our 35 classification tasks. Of the remaining 9 tasks, 8 performed relatively poorly on both classifiers (AUC < 0.65). Decision trees had consistently high AUC's (0.8-0.9) on training data, suggesting overfitting despite our parameter selection efforts. The logistic regression classifiers did not suffer from this problem, so we focused on logistic regression results and used decision trees as means of gaining intuition for the role of certain features in the classification step.

Task	AUC
Movie vs. Play	0.892
Play: pre-1800 vs. post-1800	0.776
Movie: pre-2000 vs. post-2000	0.479
Movie: G/PG vs. PG-13/R	0.594
Movie: G/PG/PG-13 vs. R	0.538
Movie: Audience good vs. bad rating	0.449
Movie: Critic good vs. bad rating	0.468
Play: Shakespeare vs. Shaw	1.000
Play: Shakespeare vs. Galsworthy	0.929
Play: Shaw vs. Galsworthy	0.750

Table 1: Logistic regression classifier AUCs for various classification tasks

	Comedy	Romance	Drama	Action	Horror	Thriller	Crime
Comedy	0.690	0.320	0.632	0.773	0.825	0.650	0.573
Romance	0.320	0.565	—	0.561	0.682	0.614	0.646
Drama	0.632	—	0.576	0.721	0.667	0.587	0.692
Action	0.773	0.561	0.721	0.662	0.643	0.640	0.563
Horror	0.825	0.682	0.667	0.643	0.660	—	0.721
Thriller	0.650	0.614	0.587	0.640	—	0.527	0.622
Crime	0.573	0.646	0.692	0.563	0.721	0.622	0.454

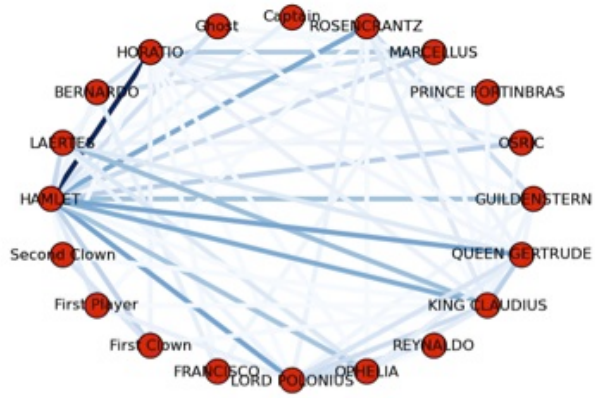
Table 2: Logistic regression classifier AUCs for genre-related classification tasks

Our results are shown in the above tables. Dashes indicate insufficient data for proper classification.

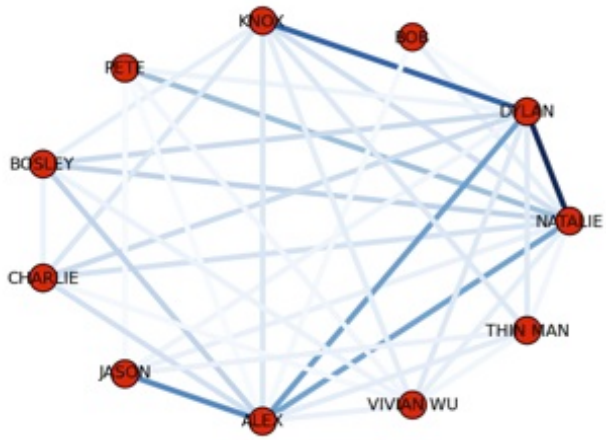
5. Analysis

5.1 Media type classifier

We were very successful in classifying plays versus movies. We found that plays are characterized by high top character relationships, high single character centrality, and low top character weight variance relative to movies, suggesting that plays tend to have a clear-cut main character with several important supporting characters that interact primarily with the main character. A classic example is *Hamlet*, as can be observed by its interaction graph:



Results for movies suggest they tend to have several main characters, as in Charlie's Angels:



5.2 Play date classifier

Important features from our pre or post 1800 play classifier, which also performed well, suggest older plays had more disjoint groups of characters and more distinct plotlines than newer ones. Misclassifications such as Shakespeare's *The Tempest* (set on an island where most characters interact with each other), which was misclassified as new, corroborated our hypothesis.

5.3 Movie date classifier

Our movie date classifiers performed poorly. We think this may be due to insufficient data, or no marked difference in interaction patterns between old and new films.

5.4 MPAA and rating classifiers

These classifiers performed poorly, aligning with our expectations because there is a great diversity in the types of movies (and their interaction networks) that are enjoyed by audiences, praised by critics, or given a certain MPAA rating.

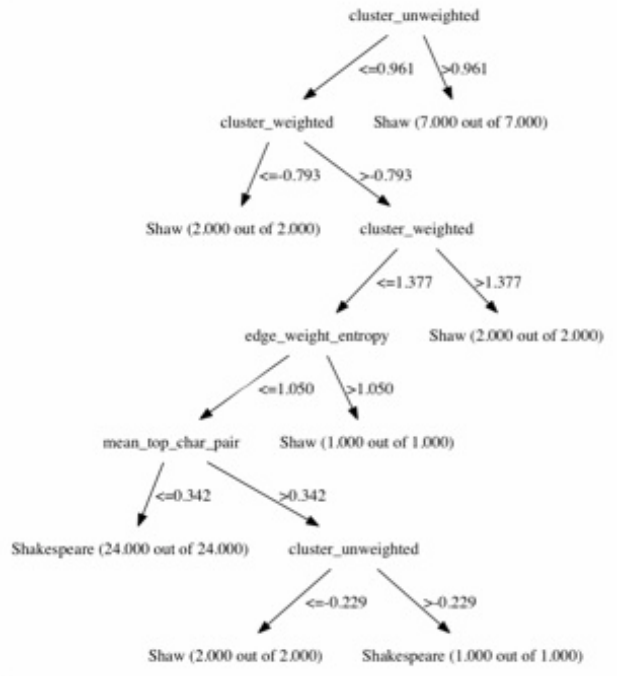
5.5 Genre classifiers

Overall, our classifier analysis confirms several common assumptions about genre stereotypes and assumptions. For example, "horror" classifiers performed particularly well, and were often characterized by high average top character relationship strength. This implies that most horror movies have one simple storyline, which is the stereotype.

As another example, romance and comedy proved far too similar to be successfully classified. Upon further reflection, character interaction networks for romances and comedies would be similar; comedies such as *Harold and Kumar* feature a dynamic duo that interacts much as love interests in a romance would.

5.6 Play author classifiers

Our classifiers achieved rather high AUC's, and an analysis of the decision trees shows that one of Shakespeare's defining characteristics is a large spread in the importance of main characters:



6. Conclusion

In this project, we developed a network extraction and classification strategy that sheds light on characteristics that define movies and plays. We automated a literary scholar's general approach to extracting meaning from movies and plays, leading us to valuable insights about large numbers of works. It is our hope that scriptwriters will be able to use these insights to increase the breadth and diversity of character interactions and counter our generalizations with unique works of entertainment!

References

Elson, D., N. Dames, and K. McKeown (2010). Extracting Social Networks from Literary Fiction. In *Proc. 48th Annual Meeting for the Association for Computational Linguistics*, 138-147.

Moretti, F. (2011). Network Theory, Plot Analysis. *New Left Review* 68.

Rydberg-Cox, J. (2011). Social Networks and the Language of Greek Tragedy. *Journal of the Chicago Colloquium on Digital Humanities and Computer Science* 1.

Stiller, J., and M. Hudson (2005). Weak Links and Scene Cliques within the Small World of Shakespeare. *Journal of Cultural and Evolutionary Psychology* 3.

Alberich, R., J. Miro-Julia, and F. Rossello (2002). Marvel Universe looks almost like a real social network. *e-print arXiv:cond-mat/0202174*

Weng, C.-Y., W.-T. Chu, and J.-L. Wu (2009). RoleNet: Movie Analysis from the Perspective of Social Networks. *IEEE Transactions on Multimedia* 11.

The Internet Movie Script Database (2011). *IMSDb*. http://www.imsdb.com/ .

Project Gutenberg (2011). *Project Gutenberg*. http://www.gutenberg.org/ .

The Complete Works of Shakespeare (2011). MIT. http://shakespeare.mit.edu/ .

EOneill.com EText Archive (1999). *EOneill*. http://www.eoneill.com/texts/index.htm .

Read Plays Online-Read Print *Read Print Library* (2011). http://www.readprint.com/ .

The EServer Drama Collection *EServer* (2011). http://drama.eserver.org/plays/ .

Rotten Tomatoes *Flixster, Inc.* (2011). http://www.rottentomatoes.com/ .

Robnik-Sikonja, M., and I. Kononenko (1997). An adaptation of relief for attribute estimation in regression. In *Proc. 14th ICML*. 296-304.

Citation studies in the humanities

Sula, Chris Alen
csula@pratt.edu
Pratt Institute, United States of America

Miller, Matt
mmille18@pratt.edu
Pratt Institute, United States of America

Abstract

This paper examines prospects and limitations of citation studies in the humanities. We begin by presenting an overview of bibliometric analysis, noting several barriers to applying this method in the humanities. Following that, we present a novel online tool for extracting and classifying citations in the humanities. This tool uses both document layout recognition and natural language processing techniques to classify citations in three ways: frequency, location-in-document, and polarity.

Background

Since the 1970s, bibliometrics has been an important method of analysis in studies of scholarly communication and the structure of academic networks that emerge from it. Bibliometricians typically focus on formal citation behavior in the printed scholarly record, occasionally supplemented with additional information. In the humanities, bibliometrics may also hold promise for tracing intellectual influence, especially when supplemented with social data (Sula 2012).

Bibliometric studies have typically focused on scientific and technical corpora, despite the fact that much intellectual history is located in the humanities (Hérubel and Buchanan 1994; Lamont 2000). This lack of attention may be explained by several factors. First, citation data in the humanities has been less available than in the sciences (Linmans, 2010), especially for monographs (Hammarfelt, 2011), which still form the backbone of humanities scholarship (Larivière, et. al. 2006), and for older sources, which humanists cite with greater frequency (Heinzkill 1980). As humanities citation data becomes more prevalent, digital humanists are likely to engage more fully with bibliometrics, and Smith's recent article on citation in classical studies is a notable example of this crossover (2009).

A second and more persistent barrier to applying bibliometrics to the humanities involves special features of humanities discourse. Studies show that humanists do engage in patterns of cocitation (Leydesdorff, Hammarfelt, and Salah 2011), but they credit each other less frequently than scientists credit each other (Heinzkill 1980; Swales 1990; Hellqvist 2010), and they rarely publish multi-authored articles (Price 1966; Pao 1981, 1982; Sievert and Sievert 1989; Wiberly 1989). Linmans (2010), for example, reports that journal publications in the humanities between 1980 and 2007 averaged a flat 1.06 authors per article. More importantly, the mere fact that one humanist references another says little about the *type* or *significance* of the relationship between the two. Several studies have shown that humanists are more likely than scientists to use integral references, which tend to associate their own views with those they references (Swales 1990; Hyland 1999; Harwood 2008), as well as negative references, which object to other authors' claims (Meadows 1974; Brooks 1985; Cano 1989). Even studies that disambiguate acknowledgments into different types, such as conceptual, editorial, financial, instrumental/technical, etc. (Cronin, Shaw, and Le Barre 2003), fail to capture qualitative elements of author ties, such as agreement, disagreement, intellectual indebtedness, and so on. These different reference contexts cannot be ignored, since intellectual disputes are the bread and butter of humanists.

Given how these nuances affect intellectual history and scholarly influence, reference contexts must be given greater attention in bibliometric studies of the humanities. Several classification schemes for references have been offered (see Table 1). Though there is some convergence in terms of positive, negative, and neutral/mixed contexts, few schemes are based on empirical research and even fewer lend themselves to practical application; many require subject domain expertise and human classification —a task that is far beyond current resources. In addition, several schemes also attempt to use citation context to sort references according to their importance within a work their prominence within the field as a whole (e.g., "historical," "classic," "homage"). In our view, this unnecessarily complicates classification schema. We argue that overall prominence is best estimated by extracontextual measures (e.g., pure or normalized citation counts), and we follow Maričić, et. al. (1998) in taking the frequency of each citation and its location-in-document to be important clues to a citation's role. For example, a reference cited throughout a work is quite different from one cited frequently at the beginning of that work, which usually helps to establish field background.

Table 1. Proposed Reference Classification Schemes

	Positive	Neutral/Mixed	Negative
Garfield (1965)	homage -pioneers -peers methodology substantiating claims authenticating data/facts	background alerting forthcoming work original publications -discussion -eponymic concepts priority claims	correction -self -other criticism negative homage -ideas
Chubin & Moitra (1975)	affirmative -essential --basic (1) --subsidiary (2)	-supplementary --add info (3) --perfunctory (4)	negational -partial (5) -total (6)
Moravcsik & Murugesan (1975)	evolutionary confirmative	conceptual– operational organic– perfunctory	negational negational
Frost (1979)	(primary texts) -support opinion/ fact --about specific author(s)/work(s) discussed --outside of central topic (secondary texts) -approval --support opinion/fact --take a step further --acknowledge indebtedness	(secondary texts) -independent --meaning of term -- acknowledgement --state of the field (neither primary nor secondary text) -further reading -bibliographic information about an edition	(secondary texts) -disapproval --disagree with fact/opinion --express mixed opinion
Smith (1981)	organic-positive	perfunctory- positive perfunctory- negative	organic-negative
Small (1982)	applied (used) supported by citing work (substantive)	noted only (perfunctory) reviewed (compared)	refuted (negative)
Peritz (1983)		setting the stage background methodological -design -method of analysis comparative argumental/ speculative/ hypothetical documentary historical casual	
Cullars (1990)	positive	mixed	negative

		neutral -springboard for discussion -establish background -support interpretation -supplementary readings	
Cullars (1992)	positive	value-free -historical background -cultural background -recommended readings -biographical data -support interpretation -scientific background	negative
Shadish, et. al. (1995)	supportive	personal creative classical social	negative
Camacho-Miñal & Múñez-Nickel (2009)	evolutionary confirmative	conceptual operational organic perfunctory other	negational juxtapositional (?)

Method

Based on this background literature, we propose an online citation extraction tool that examines PDF documents for citations and reports the frequency of a given reference as well as its location-in-document (relative to the length of the document). In addition, we propose a sentiment classifier that assigns a "polarity" value to each citation on a positive–negative scale (e.g., "I associate with…" and "I disassociate with…"). Sentiment classifiers attempt to determine whether particular sentences or documents express positive or negative opinions about a given topic (Jurafsky and Martin, 2008). This classification is found in nearly all of the other systems surveyed in Table 1, and we hypothesize that it is especially important in the humanities, since it predicts patterns of agreement and disagreement among scholars.

The sample set for this study is 159 articles from four humanities journals (see Table 2). Our choice of journals follows Knievel and Kellsey's (2005) use of eight humanities journals from 2002 and allows for comparison with their citation frequency results. These journals also reflect a range of citation layout formats, which helps to ensure the usefulness of the tool in other contexts.

Table 2. Journals used in this study

Journal	Year	Discipline	Format	Number of articles
Art Bulletin	2009[1]	art history	endnote	37
Journal of Philosophy	2011	philosophy	footnote	30
Language	2011	linguistics	inline, bibliography	18
PMLA	2011	language and literature	endnote, bibliography	74

The tool uses document layout patterns to extract each citation and the context of its occurrence—usually a sentence or two. Our website presents this text to users and allows them to select *n*-gram phrases from context that demonstrate positive or negative polarity. These phrases are compiled into a naive Bayes classifier training set which can predict polarity in novel contexts. The classifier reports polarity scores as probability assignments on two separate scales (positive and negative) each ranging from 0 to 1. Thus, a perfectly positive context would have a score of 1 on the positive scale and 0 on the negative scale. These scores may be combined into a single scale with -1 being purely negative, 0 being neutral, and 1 being purely positive.

Results

Aggregate results for frequency, location-in-document, and polarity in the sample set are reported in Table 3. Raw figures are visualized in three scatterplots comparing frequency and location-in-document (Fig. 1), location-in-document and polarity (Fig. 2), and frequency and polarity (Fig. 3).

Table 3. Aggregate results of citation extraction tool on sample set

Journal	Citations detected	Frequency per citation (avg. ± stdev)	Relative location-in-document (avg. ± stdev)	Polarity (avg. ± stdev)
Art Bulletin	1681	2.38 ± 2.36	43.81% ± 27.69	0.68 ± 0.37
Journal of Philosophy	713	1.78 ± 1.38	37.74% ± 29.09	0.52 ± 0.49
Language	2374	4.24 ± 4.58	38.77% ± 30.44	0.75 ± 0.31
PMLA[2]	604	1.83 ± 1.28	44.95% ± 27.96	0.57 ± 0.33

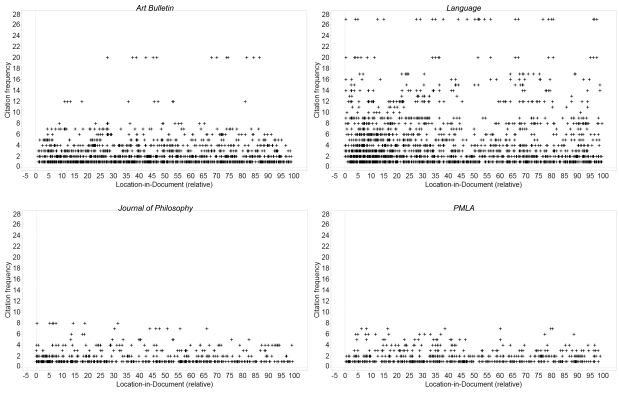


Figure 1.
Citation frequency and location-in-document

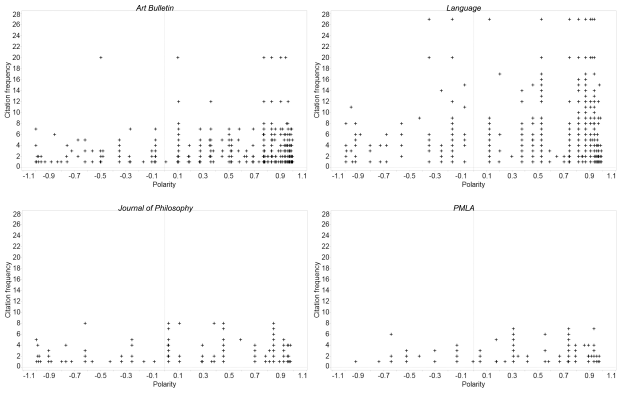


Figure 2.
Citation frequency and polarity

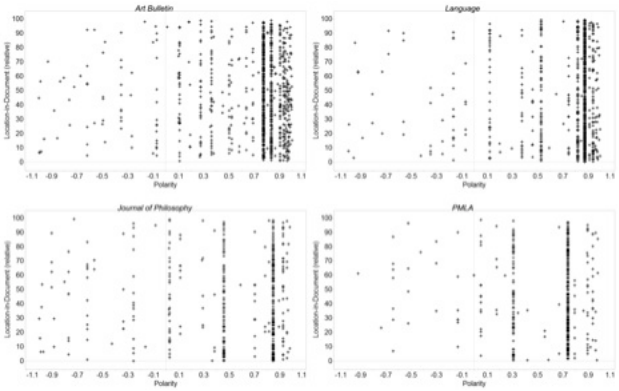


Figure 3.
Citation location-in-document and polarity

Disciplinary differences in frequency and polarity are especially evident, as is clustering near the beginning of articles.

Future directions

Though automated results were checked informally in the context of manual polarity classification, each article should be manually inspected to determine the reliability of the extraction tool. Patterns of error here may help to improve the citation extraction techniques. In addition, further training of the sentiment classifier would help to clarify the resolution of polarity scores, especially at the positive end. In particular, we are interested in examining the power of crowdsourced classifications for improving the results of classifier and for providing new document layouts that will increase the flexibility of the tool.

References

Bavelas, J. B. (1978). The social psychology of citations. *Canadian Psychological Review* 19(3): 158–163.

Brooks, T. A. (1985). Private acts and public objects: An investigation of citer motivations. *Journal of the American Society for Information Science and Technology*, 36(4): 223–229.

Camacho-Miñano, M. and M. Núñez-Nickel. (2009). The multilayered nature of reference selection. *Journal of the American Society for Information Science* 60(4): 754–777.

Cano, V. (1989). Citation behavior: Classification, utility, and locating. *Journal of the American Society for Information Science and Technology*, 40(4): 284–290.

Chubin, D. E., and S.D. Moitra. (1975) Content analysis of references: Adjunct or alternative to citation counting? *Social Studies of Science* 5(4): 423–441.

Cronin, B., D. Shaw, and K. La Barre. (2003). A cast of thousands: Coauthorship and subauthorship collaboration in the 20th century as manifested in the scholarly journal literature of psychology and philosophy. *Journal of the American Society for Information Science and Technology*, 54(9): 855–871.

Cullars, J. (1990). Citation characteristics of Italian and Spanish literary monographs. *The Library Quarterly* 60(4): 337–356.

Cullars, J. (1992). Citation characteristics of monographs in the fine arts. *The Library Quarterly* 62(3): 325–342.

Frost, C. O. (1979). The use of citations in literary research: A preliminary classification of citation functions. *The Library Quarterly* 49(4): 399–414.

Garfield, E. (1965). Can citation indexing be automated? In Stevens, M., et al. (eds.) *Statistical Association Methods for Mechanized Documentation. Symposium Proceedings, Washington, 1964*. (National Bureau of Standards Miscellaneous. Publication. 269, 189–192).

Hammarfelt, B. (2011). Interdisciplinarity and the intellectual base of literature studies: Citation analysis of highly cited monographs. *Scientometrics*, 86, 705–725.

Harwood, N. (2008). Citers' use of citees' names: Findings from a qualitative interview-based study. Journal of the American Society for Information Science and Technology, 59(6): 1007–1011.

Heinzkill, R. (1980). Characteristics of references selected in scholarly English literary journals. *Library Quarterly* 50, 352–364.

Hellqvist, B. (2010). Referencing in the humanities and its implications for citation analysis. *Journal of the American Society for Information Science and Technology*, 61(2): 310–318.

Hérubel, J.-P., and A. L. Buchanan. (1994). Citation studies in the humanities and social sciences: A selective and annotated bibliography. *Collection Management*, 18(3/4): 89–136.

Hyland, K. (1999). Academic attribution: Citation and the construction of disciplinary knowledge. *Applied Linguistics*, 20(3): 341–367.

Jurafsky, K., and J. Martin.(2008). *Speech and Language Processing*, 2nd Edition. Pearson Prentice Hall.

Knievel, J., and C. Kellsey. (2005). Citation analysis for collection development: a comparative study of eight humanities fields. *The Library Quarterly* 75(2): 142–168.

Lamont, M. (2000). Meaning-making in cultural sociology: Broadening our agenda. *Contemporary Sociology*, 29(4): 602–607.

Larivière, V., É. Archambault, Y. Gingras, and É Vignola-Gagné. (2006). The place of serials in referencing practices: Comparing natural sciences and engineering with social sciences and humanities. *Journal of the American Society for Information Science and Technology*, 57(8): 997–1004.

Leydesdorff, L., B. Hammarfelt, and A. Salah. (2011). The structure of the Arts & Humanities Citation Index: A mapping on the basis of aggregated citations among 1,157 journals. *Journal of the American Society for Information Science and Technology* 62(12): 2414–2426.

Linmans, A. J. M. (2010). Why with bibliometrics the Humanities does not need to be the weakest link: Indicators for research evaluation based on citations, library holdings, and productivity measures. *Scientometrics*, 83, 337–354.

Lisée, C., V. Larivière, and Archambault, É. (2008). Conference proceedings as a source of scientific information: A bibliometric analysis. *Journal of the American Society for Information Science and Technology* 59(11): 1776–1784.

Maričić, S., J. Spaventi, L. Pavičić, and G. Pifat-Mrzljak (1998). Citation context versus frequency counts of citation histories. *Journal of the American Society for Information Science* 49(6): 530–540.

Meadows, A. J. (1974). *Communication in science*. London: Butterworths.

Moravcsik, M. J., and P. Murugesan (1975). Some results on the function and quality of citations. *Social Studies of Science* 5(1): 86–92.

Pao, M. L. (1981). Co-authorship as communication measure. *Library Research*, 2, 327–338.

Pao, M. L. (1982). Collaboration in computational musicology. *Journal of the American Society for Information Science*, 33, 38–43.

Peritz, B. C. (1983). A classification of citation roles for the social sciences and related fields. *Scientometrics* 5: 303–312.

Price, D. J., and D. B. Beaver (1966). Collaboration in an invisible college. *American Psychologist*, 21: 1011–1018.

Shelton, R. D., and L. Leydesdorff (2011). Publish or patent: Bibliometric evidence for empirical trade-offs in national funding strategies. *Journal of the American Society for Information Science and Technology*.

Sievert, D. E., and M. E. Sievert (1989). Philosophical research: Report from the field. In *Humanists at work: Disciplinary perspectives and personal reflections*. Chicago: University Library, University of Illinois.

Shadish, W. R., D. Tolliver, M. Gray, and S. K. Sen Gupta. (1995). Author judgements about works they cite: Three studies from psychology journals. *Social Studies of Science* 25(3): 477–498.

Small, H. G. (1982). Cited documents as concept symbols. *Social Studies of Science* 8(3): 327–340.

Smith, A. G. (2004). Web links as analogues of citations. *Information Research*, 9(4) paper 188 [Available at http://InformationR.net/ir/9-4/paper188.html].

Smith, L. C. (1981). Citation analysis. *Library Trends* 30(1): 83–106.

Smith, N. (2009). Citation in classical studies. *Digital Humanities Quarterly* 3(1).

Sula, C. A. (2012). Visualizing social connections in the humanities: Beyond bibliometrics. *Bulletin of the American Society for Information Science & Technology* 38(4): 31–35.

Swales, J. M. (1990). *Genre analysis: English in academic and research settings*. Cambridge, UK: Cambridge University Press.

Wiberly, S. E., and W. G. Jones (1989). Patterns of information seeking in the humanities. *College and Research Libraries* 50, 638–645.

Notes

1. The 2009 issues of Art Bulletin were used because of their high OCR quality as compared to available 2011 editions.
2. Citation detection in PLMA exhibited several weaknesses at the time these proceedings were due. We believe these errors are due to OCR abnormalities, particularly errant spaces within words, as well as a varied mixture of parenthetical and in-text citation styles by authors. We continue work on refining the tool for these documents.

Identifying the author of the Noh play by considering a rhythmic structure — Validating the application of multivariate analysis

Takahashi, Mito
takahashi.mito@gmail.com
Doshisha University, Japan

Tezuka, Kana
bik02480248@gmail.com
Doshisha University, Japan

Yano, Tamaki
kundaikan@gmail.com
Doshisha University, Japan

I. Introduction

Noh is a classical Japanese stage art involving singing and dancing accompanied by music. Kan'ami (?-1384) and Zeami (1363–1443) are responsible for the present format of Noh. The lyrics of the song have been documented and are known as Noh texts, which are also called as "Utai bon."

The current repertoire consists of 250 plays. Noh plays are divided into five categories, numbered in the order of the frequency of performance in the past and referred by the following numbers. Japanese words Sho, Ni, San, Yon, and Go correspond to 1, 2, 3, 4, and 5, respectively , and "banme" is the ordering suffix. For each category, we list the role that the "Shite" (main actor) plays.

1st Sho-banme: Kami mono. Deity. Mythic story.
2nd Ni-banme: Shura mono. Warrior's ghost. Battle.
3rd San-banme: Katsura mono. Female. Songs and Dances.
4th Yon-banme: Kyoran mono. Madness or vengeful ghost.
5th Go-banme: Oni mono. Monsters, goblins, or demons.

Each category has a typical construction. However, it is natural that the author's preference is considerably revealed in the play, both in the texts and songs. It is very important to judge what kind of Noh play can be claimed to be Zeami's work.

The type of rhythm in Noh is called "Nori type." The standard "Nori" is "Hira-nori," in which two beats are arranged for three characters. One phrase corresponds to a verse in the seven-and-five syllable meter; thus, 7 + 5 = 12 characters (pronunciation units) will correspond to eight beats or 16 half-beats. It is shown as follows: 1 2 3 4 5 6 7 8 9 10 11 12th character (0.5, 1, 0.5, 0.5, 1, 0.5, 0.5; 1, 0.5, 0.5, 0.5, 0.5) beats.

Compared with the constant rhythmic sentence ("Teiritsu" in Japanese), the sentence is broken rhythmically ("Haritsu") if a phrase has extra syllables (more than seven for the upper phrase and more than five for the lower phrase) or has an insufficient number of syllables (less than seven or less than five) many times.

In particular, the "Kuse" part, which if exists, is the most important part in a Noh play and should be composed of Haritsu sentences. As an example, we list some Nori parts of the Noh play "Atsumori" in Table 1. The first column is "Ma" for the phrase, which usually represents the start of singing (written on the left side in Utai-bon), whereas "tori" (written on the right side) means that the lower phrase is vacant. The 2nd and 3rd columns are upper ("Kamino-ku") and lower phrases ("Shimono-ku"), respectively, and the 4th and 5th columns contain the number of characters. In the 6th column, we mention the symbolic expression of

the total phrase, whose meaning is easily recognized. In this table, we observe that

1) There are four phrases, whose upper phrase has seven characters.
2) There are three phrases, whose lower phrase has five characters.
3) There are only two phrases whose total phrase has 7 + 5 characters = 12 characters (in the symbol “h75”).

This is a typical example ofin a “Haritsu” sentence.

	ma	upper	lower	u	l	symb
1	toru	Shi-ka-ru-ni He-i-ke	-	7	0	h70
2	yaa	Yo-wo to-t-te	ni-ju-u-yo-ne-n	5	6	h56
3	yawo	Ma-ko-to-ni	hi-to-mu-ka-shi-no	4	6	h46
4		Su-gu-ru-ha yu-me-no	u-chi-na-re-ya	7	5	h75
5	yawo	Ju-e-i-no	a-ki-no ha-no	4	5	h45
6		Yo-mo-no a-ra-shi-ni	sa-so-wa-re	7	4	h74
7		chi-ri-ji-ri-ni na-ru	i-chi-yo-o-no	7	5	h75
8	yaa	Fu-ne-ni u-ki	na-mi-ni fu-shi-te	5	6	h56
9		Yu-me-ni-da-ni-mo	ka-e-ra-zu	6	4	h64

An example of phrases in “Kuse” for a play “Atsumori”

The first author expected that the authorship should be revealed in the Kuse part. The second author prepared more than 3000 phrases of the Kuse data such as that shown in Table 1 from over 70 Noh plays. The third author analyzed the data using multivariate analysis. Here we explain the partial result concerning 45 Noh plays.

We mainly studied the distribution of Zeami’s Noh plays among other plays during almost the same period. Our result shows that Zeami's work has the specific feature.

II. Previous Studies

Two previous studies were identified.

M. Yokomichi & A. Omote (1963), “Yokoku-shu. Ge."

On p.8, Prof. M. Yokomichi mentioned the proportion of unusual phrases in the Kuse part for 10 of Zeami’s Noh plays (plays with asterisk in Table 2). On p. 12 and 13, he quoted the three partial Kuse parts of Zeami’s plays in comparison with those of Kanze Nobumitsu. According to the first author, who was one of Yokomichi’s graduate students, Yokomichi seemed to conduct more studies on Haritsu, but did not publish anything on the subject.

Yoshimi Iwata (2012) The tendency of the authors of the Noh play with respect to the basic rhythm. (presentation, not published)

Iwata used Kuse and Kiri (the last song and dance part) data in 20 Noh plays, mainly in the 2nd category (Shura mono). The result shows that with the exception of Sanemori, Zeami’s plays focus on a narrow area. This study was conducted by the third author following the suggestion of the first author two years prior. At this instant, Iwata proposed the parallel usage of the Kiri part and showed that the Kiri part is helpful for the analysis.

III. Analysis

We used texts from two books “Yokyoku-shu. Jyo & Ge” (Yokomichi and Omote 1963) and extracted the Kuse parts. Then, we checked “Ma” in “Kanzeryu yokyoku hyakuban & zoku hyakuban.” We listed only 45 plays in Table 2. Zeami’s father is Kan’ami, and Motomasa (1394?-1432) is his son. Moreover , we added three plays that are attributed to either Zeami or Motomasa.

We prepared a cross table for all Kuse data by using the names of plays and phrase symbols (such as h75). Then, we obtained the 45 × 44 table, in which the rows represent “names of plays,” and the columns represent “symbols.”

First, we analyzed that table using “principal-component analysis”(PCA). Then, we selected variables using “random forests”(RF) and again analyzed them using PCA. Finally, we made a synthetic consideration.

author	cat.	plays			
Kan-ami	3	Matsukaze	Eguchi		
<K>	(3),4	Yoshino-shizuka			
	4	Jinenkoji			
Old	2	Michimori			
<O>	4	Funabashi	Unrin-in	Kashiwazaki	Hyakuman
		Ukifune	Akoya-no-matsu	Tango-monogurui	
	5	Ama	Shokun		
Zeami	1	Takasago*	Oimatsu*	Yumiyawata*	
<Z>	2	Sanemori*	Kiyotsune*	Tadanori	Yorimasa*
		Atsumori	Yashima		
	3	Izutsu*	Higaki*		
	(3),4	Saigyou-zakura			
	4	Shun-ei	Ashikari*	Hanagatami	Hanjo
		Aridoushi	Kinuta		
	5	Nue*			
Zeami?	2	Tomoakira	Tamura	Tomonaga**	
Motomasa	2	Morihisa			
<Z?><M>	4	Utaura	Yoroboshi		
others	2	Tsunemasa	Kanehira	Ikuta-atsumori	Tomoe
<o>		Ebira	Shunzei-tadamori		
<> is an abbreveation.		*: M.Yokomichi mentioned those ten plays. **: This play might be made by Motomasa.			

Table 2. Forty-five plays by Zeami and others

IV. Results

The cross table is transformed using the relative frequencies of symbols for each play. We used the 12 variables of the top frequencies to perform PCA. We formed a scatterplot for the 1st and 2nd principal components

(Fig. 1) and depicted convex hulls on the plot according to the categories of Zeami's plays (shown by big numerals). Labels are added in the abbreviated form (cf. Table 2.). Factor loadings are shown in Fig. 2.

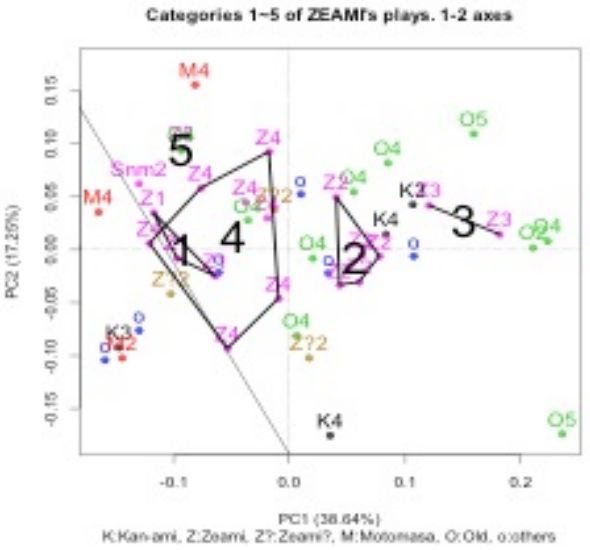


Figure 1.
Scatterplot of PCA scores for 45 plays and convex hulls

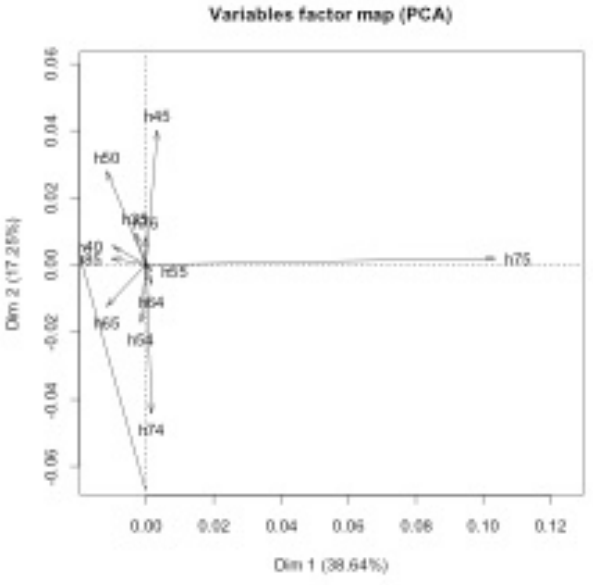


Figure 2.
PCA loadings for 12 variables.

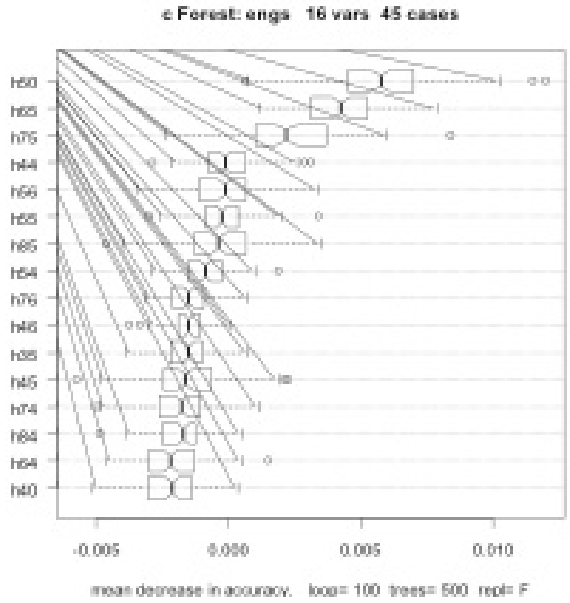


Figure 3:
Distribution of mean decrease in accuracy

	code	h50	h65	h75	h44	h56	h55	h85
Ashikari	ZEAMI	0.0789	0	0.2895	0	0	0.0789	0.0789
Izutsu	ZEAMI	0.1034	0.069	0.4483	0	0.0345	0.069	0
Utaura	Motomasa	0.1325	0.0361	0.1566	0.0482	0.0241	0.0964	0.012
Hanagatami	ZEAMI	0.0222	0.1111	0.2667	0.0444	0	0.1111	0
Ama	Old	0	0	0.4615	0	0	0	0
Aridoushi	ZEAMI	0.1	0	0.3	0.0333	0.0333	0	0.0333

Table 3.
Part of data with selected 7 variables

The result shows a greater concentration of Zeami's plays compared with those shown in Figure 1. At the far right on top, there is one "Z?2" identified with "Tomonaga." The play "Tomonaga" is attributed to Zeami or Motomasa. From this diagram, we can say that "Tomonaga" has a very different character from those of Zeami and Motomasa (M2, M4). On the other hand, "Tamura" is very near to Zeami's 2nd category.

Finally, we performed an "evaluation analysis," which was developed by the third author. The result (Fig. 5) shows that with few exceptions, for n=4–8 & m in 4–6, {hnm}'s form 5 nearby groups under the framework of selected seven variables.

V. Conclusion

The combination of PCA and RF was successfully applied for the identification of the authors of the Noh play. This methodology can be applied to other analyses related to authors will attempt to apply variable selection by

regression and compare the result with those obtained by other methods.

For Noh plays, it was revealed that the rhythm is important for discriminating characteristic features. In the near future, we will develop an appropriate decision method by using a combination of lyrics and the rhythm of Noh plays and also for the art of other kinds.

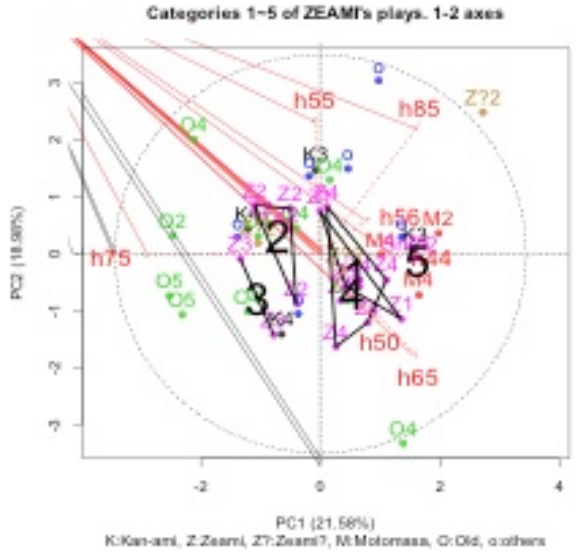


Figure 4.
Scatterplot of PCA scores and loadings

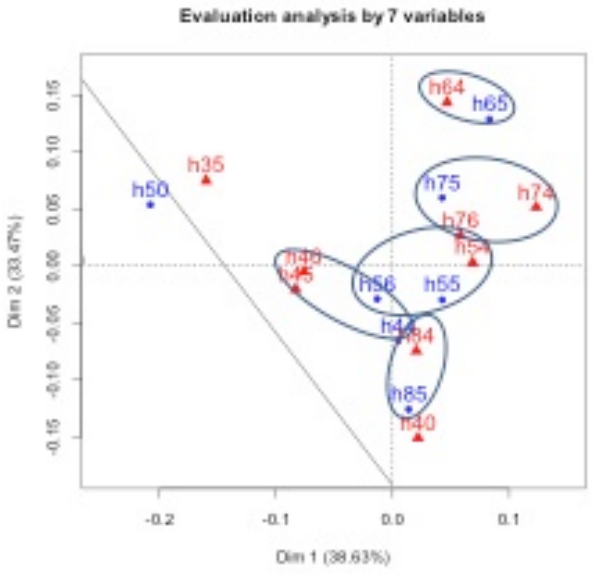


Figure 5.
Scatterplot of Evaluation Analysis

References

Yokomichi, M., and A. Omote (1963). "Yokyoku-shu, Jyo & Ge" (Anthology of Noh lyrics, volume 1 & 2), Iwanami Shoten.

Strobl, C., A. L. Boulesteix, A. Zeileis, and T. Hothorn (2007). Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics*, 8:25. http://www.biomedcentral.com/1471–2105/8/25

An Environment to Support User-Structured Digital Humanities Sources

Teehan, Aja
aja.teehan@nuim.ie
An Foras Feasa, Ireland

Keating, John
john.keating@nuim.ie
An Foras Feasa, Ireland

When building digital environments that provide suitable tools for scholarly research, there is a design tension between being generalisable and configurable: purpose-tailored tools only work with specific subjects and objects, while generalised tools can be difficult to use for a specific subject and object. However, there are meta-Use Cases that traverse many disciplines, and it is possible to build tools to support these within in a single hosting environment. In this paper we describe a hosting environment, CRADLE, that is configurable and generalisable.

At An Foras Feasa we repeatedly faced the same meta-Use Cases; researchers wished to create digital versions of their *source-system* (which might either be a single source represented by components within the system e.g. a manuscript of page-components, or multiple sources connected together, e.g. an archive of diverse materials[1]) that were reflective of their personal theoretical perspective, provided tools for investigation, and permitted personal and community re-use. This extended to researchers who wished to use the digital source-system to support their teaching remit and to others; allowing them to reinterpret and repurpose hosted digital source-systems to create a

digital version consistent with their own theoretical position. This implied that any solution would need to be able to handle multiple, configurable, versions of a digital source-system, and multiple metadata descriptions of any single component within the source-system.

CRADLE (collaborate, research, archive, discuss, learn, engage) was designed as a general software solution to support the hosting of multi-modal Humanities sources, discourse and learning resources. CRADLE has been designed to provide support for a variety of theories and interpretations of source collections. In particular, the general, but sophisticated, creator-user can configure multiple structures for their collections, and also describe multiple versions of each source-object within the collection. Furthermore, it is possible for other users to repurpose and reconfigure those source-objects to suit their own perspective upon them.

In the following fictionalised scenarios, we provide an example of how CRADLE supports this multiplicity of representation. A post-doctoral fellow with the Institute, Jane, is a professional editor, researcher and lecturer. She wanted to create a digital critical edition of an old Irish tale, originally extant in three manuscripts, so that she and her students could undertake research upon them. The digital surrogate object was composed of high-quality images of the manuscripts, their diplomatic transcriptions in XML documents, her edited best-version critical edition in XML, and an english translation of the critical edition. CRADLE provides a comparison panel to show any two of the three versions at a time(figure 1). When examining text, the XML versions are rendered, and equivalent passages across image or text versions can be located. Other tools are available to her and are briefly described further on within this paper.

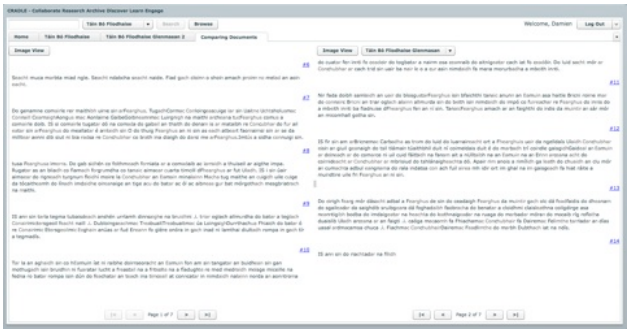


Figure 1.
The CRADLE comparison panel for digital critical edition components.

Let us now examine the multiplicity of the structures and descriptions that underlie Jane's version of her Irish Tale. Importantly, Jane self-consciously considers herself to be building a Digital Critical Edition, not a representation of the real work, within CRADLE. She holds that the XML diplomatic transcriptions are inherited from the extant manuscript, not an abstract *work*. As such, they are less *authentic* than the manuscripts. Furthermore, she does not consider the images of the manuscripts to be surrogates for the original manuscripts because they have different materiality. Each of the objects is described in metadata, but the images have two descriptions: one for their manuscript-self (in TEI), and one for the digital-image-self(in VRA). Thus, she would describe and arrange all of these objects in a hierarchy as shown below.

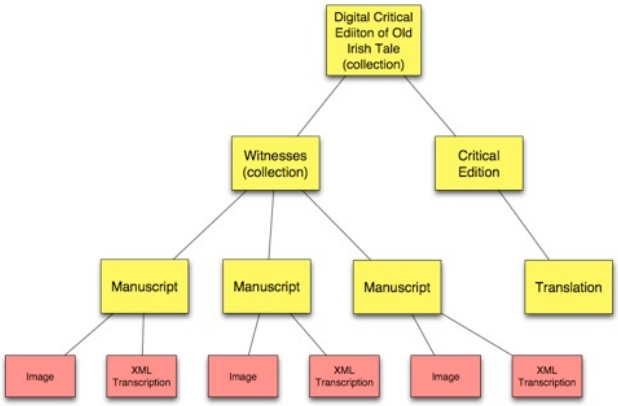


Figure 2.
Jane's personal view of the structure of her source-system, a digital critical edition.

Another researcher and lecturer wants to (re-)use the source-system. Contrary to Jane, Steven believes that all extant versions, including the XML and the translation, have the same scholarly merit and that no single one is more *authentic* than another. For him, all of the versions, including the XML documents, are children of a single, parent, abstract *work*. Steven wants to use CRADLE to describe this actual *work*, not the digital version of it. Therefore, he wants the manuscripts to be fully represented by their digital images. Steven wants to be able to explain this to his students, and then to examine the tale from his own scholarly perspective. In CRADLE it is possible dynamically redefine the spring-graph-displayed relationships between the components of the digital critical edition in a personalised user space ("My Collection"). He can also redefine the metadata description associated with the images — while Jane had associated two metadata descriptions with the images (VRA for the image and TEI for the original manuscript page), he only wants to describe the manuscript page because he is using the system as if it really were the *work*. The structure of this system follows:

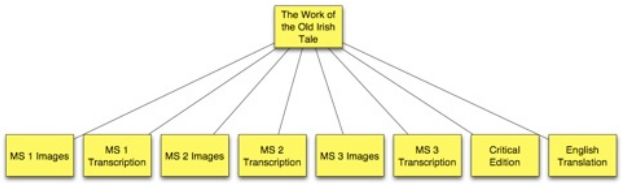


Figure 3.
Steven's personal view of the structure of Jane's original source-system.

This configuration and reconfiguration of editions, in particular, along previously undefined lines and based on personal interpretation is a direct response to calls from Eggert (2009) and Burnard (1999), among others, for what Burnard terms "an uncritical edition", and while operating at a higher level of abstraction than that to which he referred, it nonetheless allows for user-driven and defined editions. In this way, it extends the conceptual work already started in projects such as Wittgenstein's Nachlass: The Bergen Electronic Edition (BEE), which allows users to manipulate the edition within the online environment, but through predefined and encoded versions. Gabler's recent call for a means to navigate the "pyre of humanities objects" using the tools of digital humanities to "allow us to relate, or to model, their relationships — to inscribe the relations and correlations into the materials we digitise" (Gabler, 2012) is also answered with this type of approach. It is possible to use it as a direct model of the source or system, where the reader-user suspends their disbelief and operates within the collection as if it really were the real-world collection, or as an indirect model, where the difference between the real-world collection and digital collection is made obvious and described in detail.

Existing repository solution software, such as FEDORA or Hydra, support the hosting of sources and the many configurations such a system would demand, but they require significant expertise to operate. Other digital library solutions could support multiple representations of sources, but do not supply source-specific tools for detailed investigation, such as the comparison panel; nor do they support dynamic source structural re-configuration in a way that is open to general Users to manage. As far as the authors are aware, no existing systems allow for the direct association of discussion forums or learning resources with sources within a single environment. CRADLE was therefore designed to sit on top of existing repository software(currently FEDORA), provide an intuitive front-end to users wishing to create complex digital versions of their source systems, and provide tools for the research-driven interrogation of those, along with teaching and learning support.

Below it is possible to see the configurable relationships within the test collection used during development of CRADLE; the 1950s newsreel Amharc Eireann Archive, held in the Irish Film Institute.

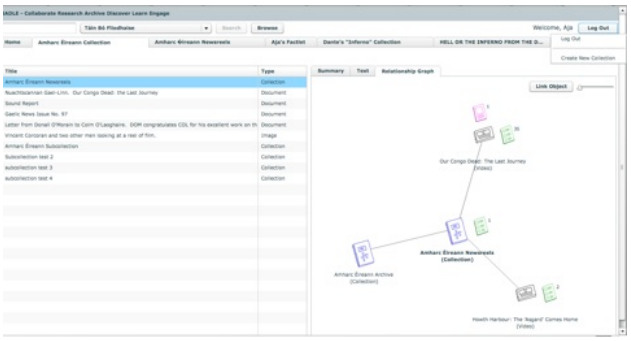


Figure 4.
Examining the relationships between the newsreel sources within an archive collection.

Metadata describing structure, such as TEI or VRA, can be interpreted by CRADLE and used to generate the relationship graphs. Here, two sample newsreels have been configured as a collection, and that collection is a child of the parent digital Amharc Eireann Archive object. On the left, the source is depicted. On the right, its relationships are viewed in a spring-graph of objects, showing their associated learning resources (a single annotation) and discourse (35 discussions). Once the new user's personal collection has been created (using the 'Create New Collection' option, in the top right of the image), it is the relationships within these graphs that can be clicked and edited to redefine the relationships. The creator-user can then control permission on the new collection to share it with colleagues, students or teachers.

Multiple perspectives on the nature and structure of humanities sources can be accommodated within CRADLE, even for the same real-world objects. Using CRADLE to create models that not only describe our sources, but represent their structure in a manner that is manipulable, reconfigurable and shareable allows researchers to define their own understanding of their material, while also making that material usable to a much wider community. Along with the many source-specific tools, and the provision of a hosting environment for learning resources and discourse associated directly with the humanities sources, this provides a valuable solution to those active in the area of digital humanities research.

References

Burnard, L. (1999). *Is Humanities Computing an Academic Discipline? or, Why Humanities Computing*

Matters. IATH Seminar. http://www.iath.virginia.edu/hcs/burnard.html. (Accessed 31 October 2012)

Eggert, P. (2009). Text-encoding, Theories of the Text, and the "Work-Site". *Literary and Linguistic Computing* 20(4). 25. Oxford: Oxford University Press.

Gabler, H. (2012). *The Importance of Digital Humanities.* Panel 1. Digital Repository Ireland Conference: Realising the Opportunities of Digital Humanities. held 23-25 October 2012 in Dublin, Ireland.

Notes

1. These distinctions are loosely based on the VRA view of collections, objects and images.

4Humanities: Designing Digital Advocacy

Thomas, Lindsay
lindsaythomas@umail.ucsb.edu
University of California, Santa Barbara, United States of America

Liu, Alan
ayliu@english.ucsb.edu
University of California, Santa Barbara, United States of America

Rockwell, Geoffrey
geoffrey.rockwell@ualberta.ca
University of Alberta

Sinclair, Stéfan
stefan.sinclair@mcgill.ca
McGill University

Terras, Melissa
m.terras@ucl.ac.uk
University College London

Bielby, Jared
bielby@ualberta.ca
University of Alberta

Smith, Victoria
victoriassmith@gmail.com
University of Alberta

Turcato, Mark
mark.turcato@mail.mcgill.ca
McGill University

Henseler, Christine
henselec@union.edu
Union College

Introduction: Cuts, Crisis and Criticisms of the Humanities

The Great Recession beginning in 2008 has resulted in a series of budgetary cuts to many universities, education programs, and cultural institutions that reflect indifference about, and even hostility toward, the humanities and arts in favor of scientific, engineering, business, and other applied fields. Even scientists are now concerned about the perceived legitimacy of their "basic research" when it lacks evident short-term application. However, the sciences have an established tradition of public advocacy and media communication that is quite effective in putting their discoveries before the public. The humanities have no such consistent, planned tradition of advocacy, and in many ways are starting from scratch.

This paper argues for and demonstrates planned humanities advocacy using the special affordances of the digital humanities. In particular, it discusses how the 4Humanities initiative is leveraging DH for next-generation advocacy. In this paper, we will:

- Show how 4Humanities uses DH to help analyze public discourse, both pro and con, about the humanities (including text analysis of such discourse as well as crowd-sourced generation of arguments for the humanities).
- Show how statistics and other evidence about the contribution of the humanities to society can be analyzed and visualized in support of effective arguments.
- Discuss the role of 4Humanities and, more generally, how DH can provide special tools for humanities advocacy, while humanities advocacy can in return incentivize the creation of next-generation DH research and teaching tools with a built-in public engagement dimension.

Analysis of Arguments Against the Humanities

One way to bolster advocacy for the humanities is first to look closely at the arguments made in public against them. We have compiled a small corpus of recent articles (especially from news sources accessed by a broad public) representative of criticisms of the humanities (Auslin, 2012; Bauerlein, 2011; Cohan, 2012a; Cohan, 2012b; Ellouk, 2011; Fendrich, 2009; Fish, 2007; Fish 2008; Fund, 2012; Knapp, 2011; Murdoch, 2011; Pidgeon, 2007; Riley, 2012; Sini, 2011; Stephens, 2012; Wente, 2012; Wood, 2012).

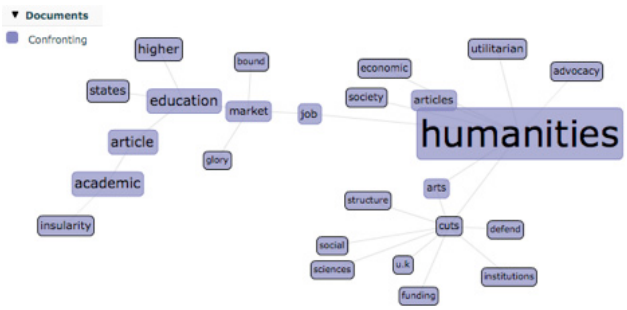


Fig 1:
Voyant Collocate Cluster Visualization (Sinclair and Rockwell)

We find that arguments critical of the humanities cluster around certain ideas. Principally, detractors accuse the humanities of lacking cultural and/or economic relevance. The most nuanced critiques mix or shade the charges of cultural and economic irrelevance. But other arguments refute the social usefulness of the humanities entirely, simply taking for granted their complete economic and social irrelevance. (Interestingly, however, some of these articles also defend irrelevance as a virtue, as in the case of arguments from friends of the humanities who feel that irrelevance is the basic nature of the humanities and is perfectly acceptable.)

A related critique of the professoriate in the humanities is that our work is no longer accessible to a larger educated public. The argument is that we are our own worst enemies because we have descended into theoretical turf battles that no one cares about. For these commentators such cultural irrelevance goes hand in hand with the supposed lack of respect that academics show for the public's values, as epitomized in Marxism, feminism, post-colonialism, and other approaches that attack iconic ideas. The antidote sometimes proposed is a return to a nebulous concept of the traditional humanities — e.g., to the venerable search for the "beautiful."

This paper will summarize our reading of the articles as well as present results from some text analysis of other rhetoric about the humanities.

Analysis of Arguments For the Humanities

As important as it is to know the arguments critical of the humanities, it is also important to gather good arguments for the humanities. 4Humanities has taken two approaches to this. The first is to blog good arguments as we come across them with summaries for those who are looking for essays to help them in their advocacy. We have also summarized these in a digestible form for people to review (Bielby, 2012). Finally we ran an All Our Ideas vote on the value of the humanities (http://allourideas.org/4humanities). At the time of writing this proposal there were over 1600 votes and 31 user submitted possible answers (as opposed to 12 seeds that we provided.) The top choices at the time of writing included:

Proposed Answer	Wins	Losses	Score	User
The humanities teach us to deal critically and logically with subjective, complex information.	22	4	82.1429	TRUE
Critical reasoning from imperfect information; empathy; understanding that the current situation is not inevitable or necessarily desirable	39	13	74.0741	TRUE
The humanities encourage us to think creatively and critically. They teach us to reason about being human and to ask questions about our world.	62	25	70.7865	FALSE
Through exploration of the humanities we learn how to think creatively and critically, to reason, and to ask questions.	11	4	70.5882	TRUE
The humanities develop informed and critical citizens. Without them democracy doesn't flourish.	53	23	69.2308	FALSE
The humanities teach us to weigh evidence skeptically, and consider more than one side of every question.	50	22	68.9189	TRUE
Humanities studies build skills in writing, critical reading. They expose us to broader cultures & create well-rounded and rational people	49	22	68.4931	TRUE
The humanities are about understanding others in the world through their languages, histories, and cultures.	172	82	67.5781	FALSE
The value of the humanities is more often in the questions posed than the answers found; humanistic study is not a formulaic equation.	23	11	66.6667	TRUE
The humanities teaches detailed, analytical thinking, as well as showing the importance of contextualized thinking.	24	12	65.7895	TRUE

There is obviously overlap in the top choices, but these indicate what the digital humanities community considers

important. In the paper we will provide a fuller analysis of the data along with our list of the best arguments.

Looking Closely at the Statistics

> "Liberal arts graduates frequently catch or surpass graduates with career-oriented majors in both job quality and compensation." (Koc, 2011)

In addition to defining and communicating the cultural value of the humanities, the 4Humanities initiative is also committed to gathering data on the economic value of the humanities. One of the most pernicious arguments against the humanities has been the poor job prospects of graduating humanists. For this reason, we review the statistical arguments carefully, especially to highlight the fact that the data on compensation of humanists at mid-career paints a different story. A humanist may find it harder to get a first job with a degree, but she/he will probably rise faster than many with professional certification.

When making arguments for the humanities' contribution to society, of course, it can be difficult to produce statistics, given that there has been no comprehensive study that gathers together available facts and figures in a usable manner. As part of the 4Humanities project we have been compiling and listing all statistics we can find in the published literature about the benefit of the humanities to society. This was done by collating the literature – including newspaper articles, reports, websites, and op-ed pieces (listed on the 4Humanities website) – and locating numeric references to the humanities. We have compiled such quantitative evidence, and at DH2013 we will present an infographic that sums up the statistical argument that the humanities are relevant to economic and intellectual development.

In addition, we have recently begun our "Infographics Friday" series (http://humanistica.ualberta.ca/category/for-the-public/humanities-infographics/), highlighting a particular statistic or graphical representation on the 4Humanities website once a week, to demonstrate the range of evidence. A core remit of 4Humanities is to gather, analyse, and disseminate this disparate information to provide a knowledge base upon which others can build their opinions and bolster their understanding of the humanities.

Conclusion: 4Humanities and a Digital Humanities response to the Cuts, Crisis, and Criticism

We argue that the mission of public engagement and advocacy in the humanities as embodied by the 4Humanities initiative provides a unique way to consolidate leading technological and methodological directions in DH with outreach to society. The humanities today have an advantage that was not available earlier: the analytical and communication methods of the digital humanities. Not only do the digital humanities provide a strong argument for the relevance of humanities learning in a digital age; they also provide unique, fresh ways of studying the contributions of the humanities to society and then getting the message out. DH research and humanities advocacy can be one, where DH helps advance advocacy, and, reciprocally, the advocacy mission helps drive research in DH. The hunt is now on to develop and extend new generations of digital humanities platforms and tools that can integrate the core research and teaching work of humanists with public visibility and engagement. Such platforms and tools (for publishing, editing, research, pedagogy, etc.) can be designed from the ground up, both to serve the needs of academics and to engage with today's networked public. This paper is a step in that direction.

References

Auslin, M. (2012). Knowledge is Good. *National Review Online*. 15 March. http://www.aei.org/article/education/higher-education/knowledge-is-good/ (accessed 13 March 2013).

Bauerlein, M. (2011). Oh, the Humanities! *The Weekly Standard*. 16 May. http://www.weeklystandard.com/articles/oh-humanities_559340.html (accessed 13 March 2013).

Bielby, J. (2012). Arguments for the Humanities. *CIRCA Wiki*, October. http://circa.cs.ualberta.ca/index.php/CIRCA:Arguments_FOR_the_Humanities (accessed 13 March 2013).

Cohan, P. (2012a). To Boost Post-College Prospects, Cut Humanities Departments. *Forbes*, 29 May. http://www.forbes.com/sites/petercohan/2012/05/29/to-boost-post-college-prospects-cut-humanities-departments/ (accessed 13 March 2013).

Cohan, P. (2012b). The 13 Most Useless Majors, From Philosophy to Journalism. *The Daily Beast*, 23 April. http://www.thedailybeast.com/galleries/2012/04/23/ the-13-most-useless-majors-from-philosophy-to-journalism.html (accessed 13 March 2013).

Davidson, C. N. (2011). Strangers on a Train. *Academe*. 97 (5). http://www.aaup.org/AAUP/pubsres/academe/2011/SO/Feat/davi.htm (accessed 13 March 2013).

Davidson, C. N. (2012). Humanities 2.0: Promise, Perils, Predictions. In Gold, M. (ed), *Debates in the Digital Humanities*. Minneapolis: University of Minnesota Press. 476-489.

Davidson, C. N., and D. T. Goldberg (2004). A Manifesto for the Humanities in a Technological Age. *The Chronicle of Higher Education: The Chronicle Review*. 13 Feb.

Delblanco, A. (2011). *College: What It Was, Is, and Should Be*. Princeton, NJ: Princeton University Press.

Ellouk, B. (2011). Do We Still Need the Humanities? *The Daily of the University of Washington*. 26 July. http://dailyuw.com/news/2011/jul/26/do-we-still-need-humanities/ (accessed 13 March 2013).

Fendrich, L. (2009). The Humanities Have No Purpose. *The Chronicle of Higher Education*, 20 March. http://chronicle.com/blogs/brainstorm/the-humanities-have-no-purpose/6738 (accessed 13 March 2013).

Fish, S. (2007). Bound For Academic Glory? *The New York Times*, 23 December. http://opinionator.blogs.nytimes.com/2007/12/23/bound-for-academic-glory/ (accessed 13 March 2013).

Fish, S. (2008). Will the Humanities Save Us? *The New York Times*, 6 January. http://opinionator.blogs.nytimes.com/2008/01/06/will-the-humanities-save-us/ (accessed 13 March 2013).

Fish, S. (2010). The Crisis of the Humanities Officially Arrives. *The New York Times*, 11 October. http://opinionator.blogs.nytimes.com/2010/10/11/the-crisis-of-the-humanities-officially-arrives/ (accessed 13 March 2013).

Fund, J. (2012). Censoring Naomi Riley. *The National Review Online*, 12 May. http://www.nationalreview.com/articles/299765/censoring-naomi-riley-john-fund (accessed 13 May 2013).

Kirschenbaum, M. (2012). Digital Humanities As/Is a Tactical Term. In Gold, M. (ed), *Debates in the Digital Humanities*. Minneapolis: University of Minnesota Press, 415-428.

Knapp, S. (2011). The Enduring Dilemma of the Humanities. *The Phi Beta Kappa Society*, 29 March. http://www.pbk.org/home/FocusNews.aspx?id=741 (accessed 13 March 2013).

Koc, E. W. (2011). Just Wait 10 Years. *New York Times*, 21 March. http://www.nytimes.com/roomfordebate/2011/03/20/career-counselor-bill-gates-or-steve-jobs/your-college-major-matter-less-over-time . (accessed 13 March 2013).

Lakoff, G. (2004). *Don't Think of an Elephant!: Know Your Values and Frame the Debate*. White River, VT: Chelsea Green Publishing Company.

Liu, A. (2012). Where Is Cultural Criticism in the Digital Humanities? In Gold, M. (ed), *Debates in the Digital Humanities*. Minneapolis, MN: University of Minnesota Press, 490-509.

Murdoch, R. (2011). The Steve Jobs Model for Education Reform. *The Wall Street Journal*, 15 October. http://online.wsj.com/article/SB10001424052970203914304576631100415237430.html (accessed 13 March 2013).

Pidgeon, S. (2007). Commented on Fish, S. Bound For Academic Glory? *The New York Times*, 23 December. http://opinionator.blogs.nytimes.com/2007/12/23/bound-for-academic-glory/#comment-100883 (accessed 13 March 2013).

Riley, N. S. (2012). The Academic Mob Rules. *The Wall Street Journal*, 8 May. http://online.wsj.com/article/SB10001424052702304363104577391842133259230.html (accessed 13 May 2013).

Sinclair, S., and G. Rockwell Collocates Cluster, from *Voyant* Tools. http://docs.voyant-tools.org/tools/links/ (accessed 13 March 2013).

Sini, M. (2011). Oh the Humanities! (OR: A Critique of Crisis). *OverLand*, 22 February. http://overland.org.au/blogs/loudspeaker/2011/02/%E2%80%98oh-the-humanities%E2%80%99-or-a-critique-of-crisis/ (accessed 13 March 2013).

Stephens, B. (2012). To the Class of 2012. *The Wall Street Journal*, 9 May. http://online.wsj.com/article/SB10001424052702304451104577389750993890854.html (accessed 13 March 2013).

Wente, M. (2012). Quebec's University Students are in for a Shock. *The Globe and Mail*, 1 May. http://www.theglobeandmail.com/commentary/quebecs-university-students-are-in-for-a-shock/article4104304/ (accessed 13 March 2013).

Wood, P. (2012). Rick Santorum is Right. *The Chronicle of Higher Education*, 29 February. http://chronicle.com/blogs/innovations/rick-santorum-is-right/31769 (accessed 13 March 2013).

Research to clarify the interrelationships between family members through the analysis of family photographs

Togiya, Norio
t_kawa@valdes.titech.ac.jp
Tokyo Institute of Technology

1. Background

In Japan, historical studies have focused on communities, such as families, have usually taken the form of using written documents. However, written documents such as diaries, records, and letters contain many descriptions of events that occurred in the past and of the actions of the head of the family etc., and they are as such not appropriate for bringing light to the relationships between the family members in the carrying out of this research. And these written documents are often dependent on the subjective view of the person who wrote them.

On the other hand, it is possible to perform an objective and holistic analysis of human relations by performing a numerical analysis of the people displayed together in photographic materials that were taken during the family's official events. Especially the aristocratic family had the habit of taking pictures at their regular rituals, meaning they left behind many official photographs. The reason that digital tools were used is that network analysis of all of the people depicted on over 100 photographs requires not only functions that assign the information of person to images but also computing facilities that perform network analysis on relationships between tagged persons, and in order to implement this, digital technology is indispensable. Thus,we constructed a digital cultural heritage system to analyze relationship in a family using photograph.

2. An Overview of Iconographic Analysis using Authoritative Information.

In this research, as shown in Fig. 1, an authoritative information database of personal names was constructed, and that information was analyzed through API using an iconographic material subject analysis system.

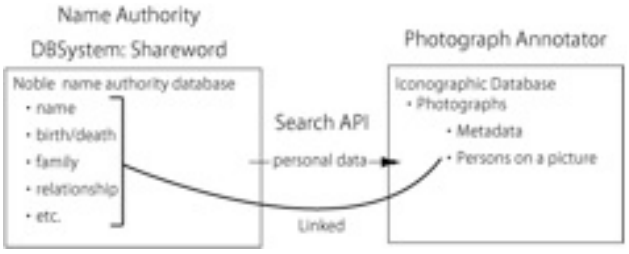


Fig 1:
System Overview

For authoritative personal name information, focusing on the pre-war imperial family, the nobility and photographers, roughly 3000 names were entered into a Shareword DB system. This authoritative personal name information was mainly made from standardized reference materials used in museums, libraries and archives, and created from element sets which allowed the sharing of information in this research, and API was added to the database to enable external searching (Togiya 2010; Kawashima and Togiya 2010). In addition, as an external system a picture annotator was created using this API.

3. Photograph Annotator

The photograph annotator was based on analysis of annotations of individuals present in group photos of the nobility. For the prototype, a stand-alone application using Java was implemented. Functions were largely divided into three: search, annotation, and analysis.

(1) Search: The search function can search through annotated photos. For each picture, metadata of title, photographer, creation date, publication, related organization, location, time and copyright is assigned, and search can be performed using this as a basis. Furthermore, for photographs that are annotated with personal names, specified characters can be searched.
(2) Annotation: Fig. 2 shows the annotation screen. Annotations are enclosed in a circle around a person's face, and are assigned the name of that individual. Using an API the name is linked to the individual in the authoritative information source, and is identified. At the present, individuals that are not included in the authoritative reference source are saved only on the local database, and are not specified. This is to prevent data from easily being added to authoritative information sources which are necessary to control, from here on the introduction of an API for adding new data is being considered.
(3) Analysis: As a further developmental feature, an analysis function was included. The analysis function creates graphs of the personal network of those assigned annotations. (Fig. 3) This graph affixes a node (point) to each individual and, by drawing edges (lines) between each person, allows the visualization of the personal relationships between individuals appearing in the same photograph.

In addition, a merit of entering blood relation data in the authoritative data DB through an API is that kinship ties between two specified individuals can be searched. By combining these two functions, it is possible to determine the kinship ties between two individuals who always appear together in photographs.

Fig 2:
Annotation Display

4. Family network analysis results

For the analysis of family members, photographs of the former ducal Iwakura family from the 1860s to the 1920s were analyzed. Divided into generations by the heads of the family, generations can be divided into the 1st generation (from the 1860s to the 1900s), the 2nd and 3rd generations (from the 1890s to the 1900s), and the 4th and 5th generations (from the 1910s to 1945). For analysis of the family network, personal name information provided from authoritative information was attached to the appropriate family heads in digitalized photographs. Then, we interpreted that "the higher level of frequency with which certain members appear together in group photographs, the stronger the interrelationship between the members", and analyzed the network relationship of the ties between each member. The basic information of analysis was summarized in Table1. The total number of photographs was 326 and the total number of family number in photographs was 29.And co-appearance pattern was 28 and the average number of occurrence of same pattern was 1.7.

As for analysis results, on a whole, since there is severance between the 1st generation and the 2nd and 3rd generations and on, and because daughters in particular bring the families into which they have married to be photographed with the main families, new groups become added to the family network in the photographs. Also, since people in the families of women who had married into the Iwakura family were invited and there were opportunities to be photographed with the family married into, it is possible to verify the formation of new family networks from the Iwakura family photographs.

Additionally, for the 1st generation, the family head was photographed alone, men and women were photographed separately, and the head, child siblings, women, and retainers were all photographed separately, network severance was observed. On the other hand, for the 2nd and 3rd generations, there were more photos which portrayed the family head together with the family members, and with the 4th and 5th generations in particular, there was a great volume of family photographs, with more photographs with relatives. Also, for the 1st generation the head husband and wife acted as a sort of "hub" which connected the family members and each of the 2nd and 3rd generations, but for the 3rd and 4th generations, the family head acted as a "hub" connecting their own generation with the previous generation.

On top of the results of this photograph analysis, we also examined written records and spoke to descendants of the family, and we learned that in actuality, in the 1st generation the family head played a feudalistic role, and family members were not able to simply interact with them. Also, the 2nd and 3rd generations had the same sort of situation, and it was the head husband and wife of the 1st generation who brought the family together and around whom the family was centered. Additionally, on the other hand, for the 4th and 5th generations, there were many cases in which the entire family was photographed together. Since the 4th generation retired from activity early on, the 4th generation head husband and wife act as the hub. Also, since the head of the 5th generation became head at the age of six, he is shown as the center of the family from early childhood, and is shown together with other family members in many photographs. Since he was the head of the family for over 60 years, it can be clearly seen from the analysis of the photographs that he acted as the "hub" which brings the family together.

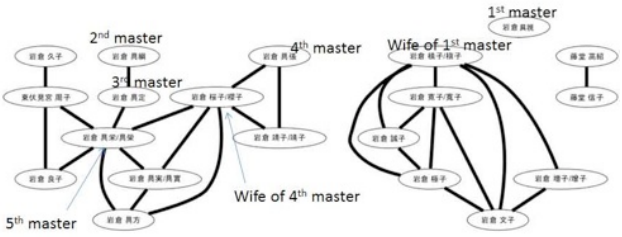


Fig 3:
Diagram of Individual Appearance Network (Chinese character Version)

Title		Data
Total number of a photographs		326
Total number of family member in photographs		29
Total number of co-appearance pattern		28
Number of occurrence of same pattern	Max	3
	Min	1
	Ave	1.71

Table 1:
The summary of the analyses

5. Usage and Future Issues

By studying photographs and other iconography using these tools, perhaps the actions of the members, families or the individuals in the targeted picture can, to a degree, be comprehended, or possibly, further investigation is necessary to objectively identify the contents of the remaining materials as well as the matter of to which of these materials these contents belong. Furthermore, in order to improve the accuracy of the examination contents, comparison and verification with other materials is necessary. For this reason, it is necessary to promote the coordination with other databases through the sharing of API.

Additionally, as explained in the previous part, by analyzing photographs added to group photos, effective analytical data regarding the manner in which human relationship networks were formed among groups such as families can be provided. By using this data, it is possible to analyze the ways in which the interrelationships between members change throughout the years and generations. Also, by using these tools, photographs which have been used as verification tools for written records can be further utilized as historical materials which can be used for the acquisition of further subjective data.

References

Togiya, N. (2010). Building an authority file for personal names that leverages societal networks: centered on personal names of the 'Shashin-shi' of the pre-War period. *The Bulletin of Japan Art Documentation Society* 17: 31-52

Kawashima, T., and N. Togiya (2010). Cooperative Database Editor for expanding Japan MARC/A *Japan Society of Information and Knowledge*. 20(1): 24-29

A Text-Mining Approach to the Authorship Attribution Problem of Dream of the Red Chamber

Tu, Hsieh-Chang
tu@turing.csie.ntu.edu.tw
National Taiwan University, CSIE Department, Taiwan, Republic of China

Hsiang, Jieh
jhsiang@ntu.edu.tw
National Taiwan University, CSIE Department, Taiwan, Republic of China

Abstract

Dream of the Red Chamber (DRC), written in the 18th century, is among the greatest Chinese classic novels. Indeed, so many studies have been devoted to this work that the term *Redology* was created to designate this field of research (Pan 1974). In 1921, Hu Shi (胡適) provided solid evidence to show that, of the 120 chapters, the first 80 were written by Cao Xueqin (曹雪芹) based on his life. He also attributed the remaining 40 chapters to Gao E (高鶚) (Hu 1921). While the first conclusion is commonly accepted, the second is not settled.

Researchers have also used statistical methods to study this problem. The most common approaches use pre-defined linguistic features, usually *function words*, to check frequencies of words. Interestingly, however, people came to different conclusions when choosing different features.

We propose a text-mining approach to the DRC author attribution problem. We define a mining function to find terms that clearly show discrepancies between the two corpuses. Some of the terms are semantic in nature, thus avoiding the pitfalls with the more syntactic function words approach. In addition to supporting the claim that the first 80 chapters and the last 40 were written by different authors, a somewhat surprising side result is the evidences that show Chapters 64 and 67, two chapters missing from the oldest existing edition, may also have been written by someone else.

1. Introduction

Authorship attribution is a well-researched subject. Brinegar (1963) used *word length* as the text feature to conclude that the 10 Quintus Curtius Snodgrass letters were not written by Mark Twain. Recent approaches, that usually assume the contextual independence of the texts being compared, make use of most frequent words and clustering analysis to identify the most likely author (Peng and Hengartner 2001, Burrows 2002, Hoover 2004, Malyutov 2006, Stamatatos 2009, Jockers and Witten 2010).

A well-known author attribution problem in Chinese literature is the author of the last 40 chapters of the novel *Dream of the Red Chamber*. Past stylistic studies lead to contradictory claims due to different feature selections and experiment designs. Karlgren (1952), Chan (1986), and He (2002) concluded that the entire DRC was written by the same person, while Zhao and Chen (1975), Yu (1998), and Yang (2003) observed significant differences between the first 80 and the last 40 chapters.

Most of these works started from choosing certain linguistic features (usually *function words*). A hypothesis testing method is then deployed to check whether the frequency distributions of features in the first 80 chapters are significantly different from those in the last 40. Yang (2003) used a different approach. They first partitioned DRC into 12 documents, each with 10 chapters. Instead of using pre-defined words, they designed a simple function that used the frequencies of unigrams to associate similarities between each pair of the 12 documents. They found strong similarities in the first 2 documents (containing Chapters 1-20), the next 6 documents (Chapters 21-80), and the final 4 (Chapters 81-120), and thus concluded that the final 40 chapters were written by a different author. However, following the same reasoning, one should also conclude that the first 20 chapters and the middle 60 were written by different authors.

2. Our text-mining approach

We propose a text-mining approach to the DRC author attribution problem. Instead of choosing pre-defined words, we design a mining function to generate candidate words. In addition to term frequencies, we also consider the number of chapters in which a term appears.

2.1 The edition question

The first question is to choose a proper edition. The earliest of DRC (1754) contains merely 16 chapters, and the second, *gengchen* edition (庚辰本) of 1760, has 78 chapters (1-80 except 64 and 67). The earliest existing version with 120 chapters, edited by Cheng Weiyuan and Gao E, appeared in 1791. The full text we chose was

	t	f(t)	$\|A_t\|$	$\|B_t\|$
1	嬤	22.3	34	0
2	裡	22.3	34	0
3	嗎	22.2	1	28
4	展	17.3	26	0
5	疆	14.8	0	11

	t	f(t)	$\|A_t\|$	$\|B_t\|$
1	豈知	31.0	0	24
2	知端	27.9	43	0
3	未知	24.5	1	31
4	一語	22.9	35	0
5	嬤嬤	22.3	34	0

	t	f(t)	$\|A_t\|$	$\|B_t\|$
6	當下	22.3	34	0
7	皆是	21.0	32	0
8	語未	20.4	31	0
9	取笑	19.8	30	0
10	惦記	18.5	0	14

Table 1.
The top 5 high-scored unigrams and top 10 bigrams computed by the mining function f(t).

the one provided by YuanZe University[1] , which is the closest to the earliest editions.

2.2 The text-mining function

Regarding each chapter as a document, we use A and B to denote the corpuses of the first 80 chapters and the last 40 respectively. Thus $|A|=80$ and $|B|=40$. We use $t \in d$ if the term t occurs in document d. Let $Dt=\{d: t \in d, d \in D\}$ be the subset of D which contains term t. We call $|Dt|$ the *document frequency* of t in D. We define the *average document frequency* of t, a term, in D, a document set, to be $pt(D)= |Dt|/|D|$. $pt(D)$ indicates the average probability for any document in D to contain t.

We define the text-mining function to be

$$f(t) = \frac{\max(p_t(A), p_t(B)) + k}{\min(p_t(A), p_t(B)) + k},$$

where a constant k is added to avoid the case $f(t)=\infty$ when $pt(A)$ or $pt(B)$ equals 0. We set $k=0.02$ in our experiments. We assume $pt(A) \geq pt(B)$ and use $k=0$ to illustrate how the function works. Then $f(t)= pt(A)/ pt(B)$ and a big $f(t)$ means a high ratio of $pt(A)$ to $pt(B)$. Thus a high-scored $f(t)$ means that the average document frequency of term t in A is significantly different from that in B.

The top 5 unigrams and top 10 bigrams obtained through $f(t)$ are given in Table 1. We have studied the top 30 unigrams and bigrams, which all showed similar behavior.

2.3 Some findings

We now briefly discuss some of our findings. The top-scored unigram *ma* (嬤) occurs only in the form of *mama*

(嬤嬤) which we shall discuss later. The second unigram *li* (裡) is interchangeable with another *li* (裏), thus could have been replaced during transcribing and should not be considered. However, we remark that among the 109 appearances of *li* (裡) in the first 80 chapters, 54 of which are in Chapter 67 alone! This strongly suggests that the current Chapter 67 (missing in the gengchen edition) was later added by another person.

The bigrams reveal even more insight. The top scored *qizhi* (豈知), which occurs in 24 chapters in corpus *B* but none in corpus *A*, does not have a clear semantics in itself. The third bigram *weizhi* (未知) appears in 31 chapters of the last 40, and the *only* chapter of the first 80 in which it appears is Chapter 64, another chapter missing in the *gengchen* edition. This provides another evident that both chapters (64 and 67) were added later by someone else. The third example, the fifth bigram *mama* (嬤嬤, a respectful title given to an elder wet nurse) occurs in 34 of the first 80 chapters but none in the last 40. There are many *mama*'s in DRC. They all conspicuously disappeared after Chapter 80.

2.4 The three-author question?

Recall that Yang (2003) also did not chose function words *a priori* and found strong discrepancies between Chapter 1-20, 21-80, and 81-120. Thus, if one is to conclude from their studies that the last 40 chapters were written by a different author, one may also need to declare that the first 80 chapters were also written by two different authors.

To make sure that our method does not pose similar problems, we ran the same experiment between the texts of Chapters 1-20 and Chapters 21-80. Not surprisingly, we found some unigrams and bigrams that appear in one corpus but not in the other. A careful analysis, however, shows that they are mostly event-dependent, involving persons or places that appeared later in the story or died. Considering that there are more than 400 characters in DRC, such event-dependent differences are expected.

3. Discussions

Our studies support the thesis that the last 40 chapters of DRC were written by someone other than Cao Xueqin. It also shows that Chapters 64 and 67 may also have been written by another person. Furthermore, the text-mining method we used offers a different approach to the author attribution problem.

A common textual analysis approach is to use function words to detect discrepancies in different texts. For instance, in Chinese *ma* (嗎) as a function word has the equivalent *me* (麼). Suppose one uses *ma* as a proof that an article is not written by a certain person, can the verdict be overturned if one uniformly replaces all the function word occurrences of *ma* by *me*?

The text-mining approach proposed here is different. Although it is also based on differences in word style, the words are generated by the method itself. Take the term *mama* as an example. *Mama* appeared in 34 of the first 80 chapters, and was used to address quite a few minor characters in the book (last appearance in Chapter 80). However, not only was the term completely missing from the last 40 chapters, so did the concept and the characters! Such "semantic" differences seem to provide more solid evidence than purely syntactic ones.

References

Brinegar, C. S. (1963). Mark Twain and the Quintus Curtius Snodgrass Letters: A Statistical Test of Authorship, *Journal of the American Statistical Association,* 58(301): 85-96.

Burrows, J. (2002). 'Delta': a Measure of Stylistic Difference and a Guide to Likely Authorship, *Literary and Linguistic Computing*, 17(3).

Chan, B.-C. (1986). *The authorship of the Dream of the red chamber based on a computerized statistical study of its vocabulary*, Joint Publishing Co Ltd., Hong Kong.

He, G.-G. (2002). From Chinese function words to the characteristics of authors – also the author attribution problem of Dream of the Red Chamber, *Traditional Chinese Literature e-Journal*, Hualian.

Hoover, D. L. (2004). Testing Burrows's Delta, *Literary and Linguistic Computing*,19(4).

Hu, Shi(1921). *Textual Research on the Dream of the Red Chamber*, reprinted by Yuandong Publishing, 1985.

Jockers, M. L., and D. M. Witten (2010). A comparative study of machine learning methods for authorship attribution, *Literary and Lingistic Computing*, 5(2).

Karlgren, B. (1952). New Excursions in Chinese Grammar, *Bulletin of the Museum of Far Eastern Antiquities* (Stockholm), 24: 51-80.

Malyutov, M. B.(2006). Authorship attribution of texts: a review. *General Theory of Information Transfer and Combinatorics*, Springer-Verlag, 362-380.

Pan, C.-G.(1974). *Sixty years of Redology*. 2,226 pages. Taipei: Wen-shi-je Publishing.

Peng, R., and N. Hengartner (2001). *Quantitative Analysis of Literacy Styles*. Department of Statistics Papers, Department of Statistics, UCLA.

Stamatatos, E. (2009). A Survey of Modern Authorship Attribution Methods, *Journal of the American Society for*

Information Science and Technology, 60(3): 538-556, March 2009.

Yang, A. C.-C., C.-K Peng, H.-W. Yien, and A. L. Goldberger (2003). Information categorization approach to literary authorship disputes, *Physica A* 329: 473-483.

Yu, Q.-X. (1998). Applications of Statistical methods to Dream of the Red Chamber, *Journal of National Cheng-Chi University*, 76: 303-327.

Zhao, G., and Z. Chen (1975). «紅樓夢研究新編» *A New Compilation on the research of The Dream of The Red Chamber*, Taipei: Linking Publishing.

Contemporary solutions to retrieve and publish information in ancient documents using RDF and Islandora

Tupman, Charlotte
charlotte.tupman@kcl.ac.uk
King's College London, United Kingdom

Jordanous, Anna
anna.jordanous@kcl.ac.uk
King's College London, United Kingdom

Stanley, Alan
astanley@upei.ca
University of Prince Edward Island, Canada

Introduction

This paper considers what can be gained from enhancing TEIencoded texts with RDF and OAC annotations and transforming to other representations, and how to facilitate their production, editing and storage. Our case study is the Sharing Ancient Wisdoms (SAWS)[1] project, which analyses the tradition of wisdom literatures. Scholarly interest is focused on semantic links within and between specific sections of these texts. SAWS produces TEIbased digital editions with semantic annotations in RDF to allow investigation of these links as Linked Data. This approach has the potential to be used widely to link and describe related sections of a variety of texts. Links can be extracted and transformed for manipulation and searching using alternative methods, illustrated by our TEItoRDF XSLT. In producing, storing and annotating such documents, the TEIediting process may present barriers to information enrichment for nontechnical users. The Islandora repository management software assists in creating and managing collections of documents through more intuitive, GUIdriven interactions with Fedora repositories. Within Islandora, the Digital Humanities Solution Pack provides a WYSIWIG online interface to help create, edit and annotate TEI documents, and simplifies the addition of semantic links. We demonstrate how TEI documents can be developed in diverse directions using Linked Data and RDF, and show how production of LinkedDataenhanced TEI documents can be facilitated using the Digital Humanities Solution Pack within Islandora.[2] RDF triples generated through the Islandora interface are exposed as standalone relationships which may be applicable in other contexts.

The Sharing Ancient Wisdoms (SAWS) use case

SAWS[3] [4] [5] is a key use case for this work, requiring an approach encapsulating various types of information including structural markup and semantic annotation. SAWS enables linking and comparisons within and between anthologies, their source texts, and their recipient texts, acting as a framework through which others can link their own materials via the Semantic Web.

SAWS focuses on *gnomologia* [6], collections of sayings that transmitted moral or philosophical ideas.[7] These sayings were selected from earlier manuscripts, reorganised or reordered, and often modified or reattributed. The texts crossed linguistic barriers in the mediaeval period, and in later centuries were translated into western European languages. They form a complex network of interrelated texts, which when analysed can reveal much about the dynamics of the cultures that created and used them.

SAWS enables investigation of the relationships between specific sayings, tracing the links through different textual variants and languages. This has been achieved by enhancing our TEI with RDF: each saying can be linked to other relevant sections of text via a subject-predicate-object relationship defined as part of an ontology. The `<relation>` element, which has recently been updated with new attributes, allows us to enter RDF directly into the TEI document and combine this with information about scholarly responsibility.[8]

Combining TEI and RDF

TEI allows for extremely granular expression within a context; RDF is often meaningful in the absence of context. The strength of RDF lies in its apparent simplicity and its interoperability: its data is discoverable and reusable. Combining subject-predicate-object assertions can convey considerable metadata and tell complex stories. RDF can also be expressed as OAC annotations, which may have any number of targets of differing types. A target may indicate a section which overlaps another (via spatial or indexing coordinates) without breaking XML validation. SAWS implements the CITE/CTS citation scheme,[9] allowing overlapping sections to be described fully and referenced using anchor points in the TEI structure.

SAWS accommodates TEI and RDFcompatible markup within the same document and workflow, using established RDF syntax for marking up information of semantic interest. For SAWS, it is preferable to keep structural, syntactic and semantic markup in the same documents where possible, and to access the semantic information using standard tools such as XSLT.[10] [11]

Previous approaches to the recording of semantic links within TEI documents have had limitations. The EARMARK ontology[13] provides an RDF model for XML information, but only for structure, so structural information is separated from text, and we cannot add semantic information while editing. RDFTEF[14] [15] requires documents to be edited in a separate environment within which standard XML tools cannot be used.[16] Approaches to incorporating RDF within XML documents do not transfer easily to a TEI representation: RDFa encodes RDF directly within specific XML attributes, but key attributes for RDFa[17] are not included in standard TEI schemas.[18]

While it would not technically be difficult to use RDFa by extending the TEI schema, this would introduce extra work which may not be necessary, and it would mean ignoring suitable alternatives proposed and accepted by the TEI community (discussed below) which require no extra schema work; if considered suitable, adopting such an alternative would enable SAWS to contribute towards establishing conventions within the TEI community for working with RDF within TEI.

Recognising the importance of combining RDF and TEI, a TEI Special Interest Group (SIG) in the use of ontologies[19] is developing XSLTs to transform TEI documents into RDF, using the CIDOC-CRM[20] as a basis.[21] [22] [23] [24] The SIG maps only a subset of elements to CIDOC-CRM, focusing on those that represent very particular entities.[25] SAWS would therefore not be able to retrieve many triples of scholarly interest such as manuscript structure and metadata.

To represent a wider range of data, a recent TEI recommendation[26] has been adopted by SAWS, using the `<relation>` element to represent links from one object [27] (@active) to another (@passive), using link types (@ref) which can incorporate a domain ontology.[28] [29] This increases the expressiveness of the markup without requiring changes within TEI. `<relation>` is an established element; the more recent addition of @ref has enabled `<relation>` to be used for RDF triples, along with the assertion of responsibility using @resp.

An XSLT stylesheet for extracting information from TEI to RDF

Semantic information can be accessed in limited ways via a TEI document, but when extracted, it can be placed in a triple store for access, querying and reasoning.[30] New knowledge can be derived by traversing internal links, and following links to related external Linked Data sources.[31]

We offer an XSLT stylesheet that transforms TEI, rerepresenting the structural, semantic and metadata information as RDF/XML triples.[32] Acknowledging practical difficulties concerning the size of the TEI tagset, we take the minimal required version of TEI, TEIBare. This forms a base for future extension, e.g. to TEILite. [33] [34] Using Dublin Core terms[35] such as dct:creator and dct:title, statements in the TEI header are transformed into corresponding RDF triples, and structural ordering of blocks within the TEI document are encoded using dct:isPartOf and dct:hasPart triples. We have extended the XSLT to include transformation of triples encoded through the `<relation>` element into RDF syntax, and further extensions can be added.

Transformations from TEI to RDF for the SAWS use case

The SAWS TEI version of the Kitāb alḤaraka (“Book of Happiness”), held at Ankara Üniversitesi, contains various metadata in its header. Applying the XSLT generates the following triples:

```
<rdf:Description
rdf:resource="http://www.ancientwisdoms.ac.uk/mss/urn:cite:saws:HacıMahmu
d5683">[36]
     <dct:title>Hacı Mahmud Efendi 5683</dct:title>
     <dct:creator>(Pseudo-)Aristotle</dct:creator>
     <dct:type>TEI/XML</dct:type>
    <dct:conformsTo>http://www.tei-c.org/ns/1.0</dct:conformsTo>[37]
</rdf:Description>
```

Triples of the Data Model

Number	Subject	Predicate	Object
1	http://www.ancientwisdoms.ac.uk /mss/urn:cite:saws:HacıMahmud5683	http://purl.org/dc/elements/1.1/title	"Hacı Mahmud Efendi 5683"
2	http://www.ancientwisdoms.ac.uk /mss/urn:cite:saws:HacıMahmud5683	http://purl.org/dc/elements /1.1/creator	"(Pseudo-)Aristotle"
3	http://www.ancientwisdoms.ac.uk /mss/urn:cite:saws:HacıMahmud5683	http://purl.org/dc/elements/1.1/type	"TEI/XML"
4	http://www.ancientwisdoms.ac.uk /mss/urn:cite:saws:HacıMahmud5683	http://purl.org/dc/elements /1.1/conformsTo	"http://www.tei-c.org/ns/1.0"

Graph of the data model

The SAWS TEI version of the Corpus Parisinum manuscript, held in the Digby collection in Oxford's Bodleian library, contains a section `<div xml:id="Aristippus01">` which is contained by its parent, `<div xml:id="Part01">`. From this we can derive the following structural triples:

```
<rdf:Description
rdf:about="http://www.ancientwisdoms.ac.uk/mss/urn:cite:saws:CPIV.D:Aris
tippus01">
<dct:isPartOf
rdf:resource="http://www.ancientwisdoms.ac.uk/mss/urn:cite:saws:CPIV.D:P
art01"/>
</rdf:Description>

<rdf:Description
rdf:about="http://www.ancientwisdoms.ac.uk/mss/urn:cite:saws:CPIV.D:Part
01">
<dct:hasPart
rdf:resource="http://www.ancientwisdoms.ac.uk/mss/urn:cite:saws:CPIV.D:A
ristippus01"/>
</rdf:Description>
```

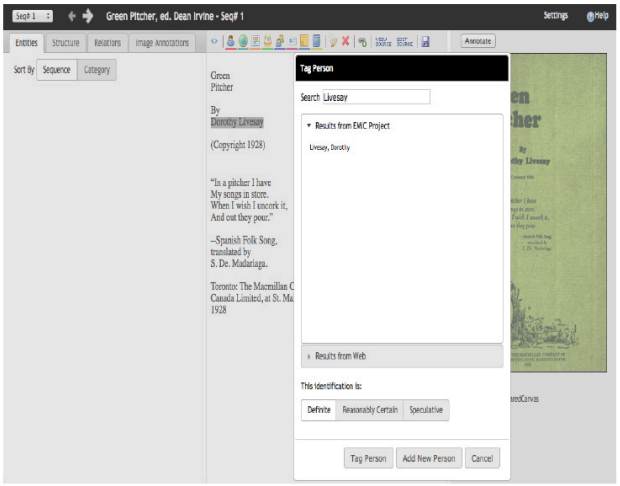
Feedback on the editing and linking process

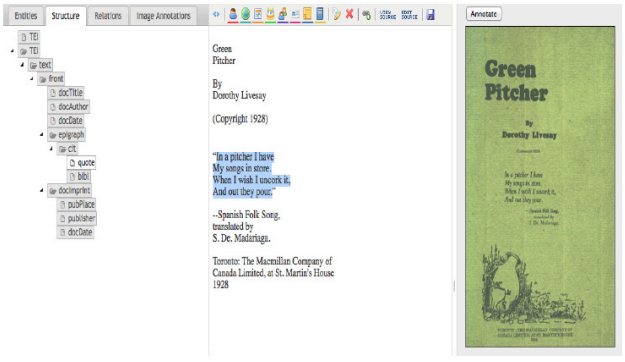
SAWS scholars studying documents in 'right-to-left' (RTL) languages noted the difficulties in working with standard XML editing software, and also requested more intuitive interfaces for editing documents and adding `<relation>` links.

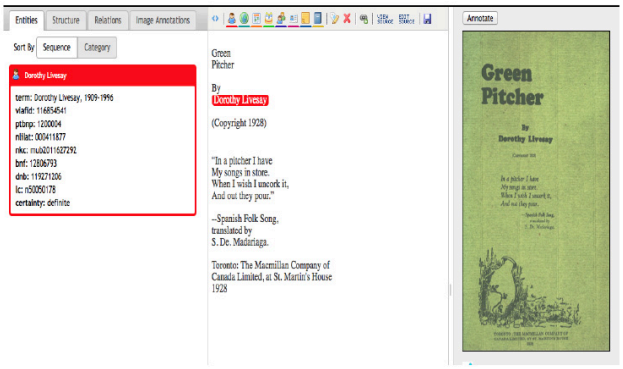
Islandora is an open source project allowing users to manage a Fedora repository through PHP using a Drupal front end. Fedora repositories are adept at maintaining and versioning metadata accompanying scholarly objects. Islandora provides an intuitive way to use Fedora to create, access and manage document collections, and is currently being used across a varied number of use cases.[38]

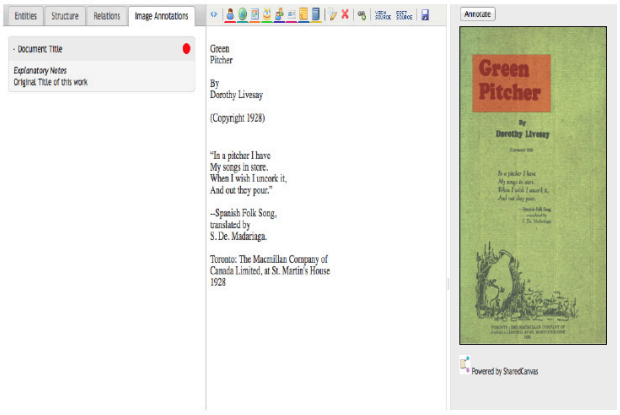
Various "solution packs" within Islandora are available for different types of projects. The Digital Humanities Solution Pack is specifically designed for text editing and annotation, based on Shared Canvas and CWRC (for editing TEI and adding links). These tools are used to access, edit, and retrieve information held in repositories, including TEI transcriptions of texts, OCR tools related images, annotations and metadata. This Digital Humanities project within Islandora is sponsored by EMiC to develop a suite of applications for managing and critically analysing Canadian modernism. As one of the authors of this paper is the lead programmer of both these projects, he can incorporate these transformations into the workflow to expose the data publicly. Of particular interest is the ability to extract data from TEI to build and maintain authority lists.

The Islandora Critical Editions module exposes a GUI allowing the addition and viewing of RDF entities and TEI tags. No knowledge of XML is required. Entities tie textual offsets to objects from authority lists, userentered notes, external links, or date ranges through RDF. Image annotations are OAC RDF annotations.









Concluding remarks

We offer a functional XSLT for converting TEI to RDF, incorporating the recent application of `<relation>` for encoding RDF within TEI and extracting TEI `<relation>` elements and selected structural markup as RDF files.

Future SAWS/Islandora collaboration will investigate the enhancement of TEIencoded documents and a more user-friendly environment for editing, managing and linking texts. The DH Solution Pack by Islandora is available by request but has not yet been released in beta version. It is intended that SAWS will have implemented and tested a working version of the DH Solution Pack by June 2013. Any DH project that wants to link TEI files with other sources of information will, we argue, benefit from investigating the DH Solution Pack. It has wide implementation possibilities and will be particularly useful for projects using right-to-left languages.

The outcomes of this SAWS/Islandora collaboration should apply across a wide variety of texts. It is hoped that this paper will stimulate further interest in RDF and Linked Data within TEI, particularly amongst Digital Humanists wishing to work with a broader range of Humanities scholars.

Notes

1. http://www.ancientwisdoms.ac.uk . Last accessed October 2012.
2. **John Unsworth.** (2003). Tool-Time, or 'Haven't We Been Here Already?' Ten Years in Humanities Computing. Delivered as part of "*Transforming Disciplines: The Humanities and Computer Science*," Washington, DC. Available at: http://people.lis.illinois.edu/~unsworth/carnegieninch.03.html (last accessed 20th July 2012).
3. **Anna Jordanous, K. Faith Lawrence, Mark Hedges, and Charlotte Tupman.** (2012). Exploring manuscripts: sharing ancient wisdoms across the semantic web. *In Proceedings of the 2nd International Conference on Web Intelligence, Mining and Semantics* (WIMS '12), Craiova, Romania.
4. **Tupman, Charlotte; Hedges, Mark; Jordanous, Anna; Lawrence, Faith; Roueche, Charlotte; Wakelnig, Elvira; Dunn, Stuart.** (2012). Sharing Ancient Wisdoms: developing structures for tracking cultural dynamics by linking moral and philosophical anthologies with their source and recipient texts. In *Proceedings of Digital Humanities* (DH2012), Hamburg, Germany.
5. **Hedges, Mark; Jordanous, Anna; Dunn, Stuart; Roueche, Charlotte; Kuster, Marc W.; Selig, Thomas; Bittorf, Michael; Artes, Waldemar;**(2012). "New models for collaborative textual scholarship,", *Proceedings of the 6th IEEE International Conference on Digital Ecosystems Technologies (DEST)*, Campione d'Italia, Italy.
6. **F. Rodríguez Adrados,** (1981). *Greek wisdom literature and the Middle Ages: the lost Greek models and their Arabic and Castilian Translations* (2001), English translation by Joyce Greer (2009), pp. 9197 on Greek models; D. Gutas, "Classical Arabic Wisdom Literature: Nature and Scope", *Journal of the American Oriental Society*, Vol. 101, No. 1, Oriental Wisdom (Jan. Mar., 1981), pp. 4986
7. **M. Richard,** (1962). "Florilèges grecs", *Dictionnaire de Spiritualité* V, cols. 475512
8. A full discussion of our TEI markup and use of the `<relation>` element can be found here: Tupman, Charlotte; Hedges, Mark; Jordanous, Anna; Lawrence, Faith; Roueche, Charlotte; Wakelnig, Elvira; Dunn, Stuart. Sharing Ancient Wisdoms: developing structures for tracking cultural dynamics by linking moral and philosophical anthologies with their source and recipient texts. In *Proceedings of Digital Humanities* (DH2012), Hamburg, Germany. 2012.
9. (http://www.homermultitext.org/hmtdoc/cite/)
10. **M. O. Jewell.** (2010). Semantic Screenplays: Preparing TEI for Linked Data. In Proceedings of Digital Humanities, London, UK.

11. 11 **K. F. Lawrence.** (2011). Wherefore Art Thou? Crowdsourcing Linked Data from Shakespeare to Dr Who. In Proceedings of Web Science, Koblenz, Germany.
12. **Blanke, Tobias; Bodard, Gabriel; Bryant, Michael; Dunn, Stuart; Hedges, Mark; Jackson, Michael; Scott, David;** (2012). "Linked data for humanities research — The SPQR experiment," *6th IEEE International Conference on Digital Ecosystems Technologies (DEST)*, Campione d'Italia, Italy
13. **S. Peroni and F. Vitali.** (2009). Annotations with EARMARK for arbitrary, overlapping and outof order markup. In Proceedings of the 9th ACM symposium on Document engineering, pages 171180, Munich, Germany.
14. **G. Tummarello, C. Morbidoni, and E. Pierazzo.** (2005). Toward textual encoding based on RDF. In Proceeding of the 9th International Conference on Electronic Publishing (ELPUB 2005), Kath. Univ. Leuven, June, pages 5763.
15. RDFTEF sourcecode: http://rdftef.sourceforge.net/ Last maintained 2007.
16. **P. Portier, N. Chatti, S. Calabretto, E. Egyed-Zsigmond, and J. Pinon.** Modeling, encoding and querying multistructured documents. Information Processing & Management. Forthcoming.
17. e.g. @rel, @rev, @href, @resource, @property, @vocab
18. A more detailed discussion of existing methods for encoding RDF within TEI markup can be found in: A. Jordanous, A. Stanley and C. Tupman. Contemporary transformation of ancient documents for recording and retrieving maximum information: when one form of markup is not enough. In *Proceedings of Balisage: The Markup Conference 2012*. Balisage Series on Markup Technologies, vol. 8 (2012), Montréal, Canada, August 2012.
19. TEIOntologies Special Interest Group http://www.teic.org/SIG/Ontologies/
20. **ChristianEmil Ore and Øyvind Eide.** (2009). TEI and cultural heritage ontologies: Exchange of information? *Literary and Linguistic Computing* 24(2): 161172
21. http://www.edd.uio.no/artiklar/tekstkoding/tei_crm_mapping.html , http://www.edd.uio.no/tei/teiontsig/test_crm_model.graphml
22. http://www.teic.org/release/xml/tei/stylesheet/rdf/
23. http://www.teic.org/SIG/Ontologies/guidelines/guidelinesTeiMappableCrm.xml
24. CIDOCCRM only direct represents textual material through one class (E33 Linguistic Object) and its two subclasses (E34 Inscription, E35 Title), but its selection as a base model is partially influenced by research aims of the SIG members in enhancing cultural heritage and museum documentation (http://www.teic.org/SIG/Ontologies/guidelines/guidelinesTeiMappableCrm.xml). Dicussions (see http://wiki.teic.org/index.php/SIG:Ontologies) about the use of FRBRoo, a bibliographical records model harmonised with CIDOCCRM (http://www.cidoccrm.org/frbr_inro.html) have not been acted upon, to date.
Some mappings from TEI to Dublin Core (a model of metadata information: http://www.dublincore.org) are occasionally present in stylesheets created by the SIG (http://www.teic.org/release/xml/tei/stylesheet/rdf/dc.xsl) but this output has not been highlighted, despite Dublin Core also being a realistic option for more detailed mappings of document metadata, especially from the TEI header.
25. http://wiki.teic.org/index.php/SIG:Ontologies
26. Sourceforge.net discussion: Encoding RDF relationships in TEI ID: 3309894, at http://tinyurl.com/lrbz53b
27. The application of `<relation>` to express RDF triples has been documented by TEI at http://www.teic.org/release/doc/teip5doc/en/html/refrelation.html with supporting examples.
28. The SAWS ontology (an extension of FRBRoo for representing relations of interest for study of wisdom manuscripts) is available at http://purl.org/saws/ontology .
29. S. Dunn, M. Hedges, A. Jordanous, K. F. Lawrence, C. Roueché, C. Tupman, and E. Wakelnig, Sharing Ancient Wisdoms: developing structures for tracking cultural dynamics by linking moral and philosophical anthologies with their source and recipient texts, *Digital Humanities 2012*, Hamburg, Germany.
30. For the SAWS use case, a SPARQL endpoint to access the RDF data is available (http://www.ancientwisdoms.ac.uk/sparql)
31. To date, SAWS links to various collections of ancient data interlinked through Pelagios (http://pelagios.blogspot.com) references to the Pleiades historical gazetteer (http://pleiades.stoa.org/). We are also in the process of linking to existing relevant documents such as in the Perseus Digital Library (http://www.perseus.tufts.edu /) and would like to link to information on people mentioned in the texts, such as through the Prosopography of the Byzantine World resource (http://www.perseus.tufts.edu/). The facility to traverse links between sets of data and discover related information serendipitously is a major benefit of Linked Data for the SAWS project.
32. XSLT available at http://www.ancientwisdoms.ac.uk/media/ontology/tei_to_rdf.xsl , with working versions available through https://github.com/ajstanley/TEI_to_RDF .
33. The Dublin Core Metadata Initiative is the main source model for structural and metadata mappings from TEIBare to RDF: http://dublincore.org/documents/dcmiterms/
34. http://www.teic.org/Guidelines/Customization/
35. The namespace 'dct' represents http://dublincore.org/documents/dcmiterms/
36. The manuscript ID is in CITE/CTS format for document citation (see http://www.homermultitext.org/hmtdoc/cite/)

37. The dct:conformsTo relationship requires the object of the triple to be a string, rather than a resource
38. List of current Islandora installations: http://islandora.ca/current_installations

Authorship problem of Japanese early modern literatures in Seventeenth Century

Uesaka, Ayaka
dil0015@mail4.doshisha.ac.jp
Doshisha University Graduate School,Japan

Murakami, Masakatsu
mamuraka@mail.doshisha.ac.jp
Doshisha University, Japan

I. Introduction

This study aims to focus on *Yorozu no humihougu* 万の文反古 ("An Old letter scrapbook"; 1696), a collection of posthumous works in the early modern Japanese genre of *Ukiyozoushi* 浮世草子, written by Saikaku Ihara 井原西鶴 (1642?–1693) as a classical, foundational document of Japanese culture; then it will examine the "authorship problem" in Saikaku's works using the tools of quantitative analysis.

In contrast to so-called scholarly books with named authors, graphic novels or storybooks called *soushi* 草子(realistic literature), were generally anonymous in this era. The earliest work generally acknowledged as *Ukiyozoushi* 浮世草子, and thus the first Japanese early modern novel, emerges with the publication of Saikaku's *Kousyoku ichidai otoko* 好色一代男 ("The life of an amorous man"; 1682) (Munemasa 1969), but there are almost no attributed (signed) works among the 23 *Ukiyozoushi* 浮世草子 considered to be Saikaku's. While work on Saikaku has proceeded, these fundamental doubts about his authorship remain.

Meanwhile, the potential of quantitative analysis of textual data and the related field of the digital humanities have also dramatically advanced. However, quantitative analysis of Japanese classical works has lagged behind. Delayed digitalization of classical works has been a problem due to complications regarding development of morphological analysis software for classical works. At this moment, adequate morphological analysis software for early modern Japanese literature does not yet exist.

Five Saikaku's collections of posthumous works were edited and published from 1693 to 1699 by followers of the author including Dansui Houjou 北条団水, on the basis of unpublished drafts considered by them to have been Saikaku's. As a result, some doubt arose subsequently about the authorship and publication history of these works.

Yorozu no humihougu 万の文反古 was published as the fourth such collection of posthumous works, in the third year after the death of Saikaku (1696). This work is a collection of epistolary novels, consisting of 17 chapters each telling a different short story. The next section considers the doubts about Saikaku's authorship.

II. Previous Studies

A. Doubts raised by Yamaguchi

Yamaguchi (1929) mentions that *Yorozu no humihougu* 万の文反古 may be an apocryphal work actually written by Dansui, for the following reasons.

1. While the handwriting is similar to Saikaku's, the lines are bolder and there is a slight lack of roundness due to the powerful strokes.
2. Unlike in the other posthumous publications, Dansui did not provide a preface to *Yorozu no humihougu* 万の文反古.
3. The publishers of *Yorozu no humihougu* 万の文反古 are the same as those for *Saikaku oridome* 西鶴織留, the unfinished second collection. There is doubt as to why they chose to publish *Yorozu no humihougu* 万の文反古, later than *Saikaku oridome* 西鶴織留 although it was in more complete form.
4. While it seems clear that some of the work was written by Saikaku, some chapters seem to have been rewritten by Dansui to impose an epistolary form and remove some descriptive passages. It is considered that these intrusions are less elegantly composed.

B. Arguments for Saikaku's scholarship

Teruoka (1953) thinks that the handwriting is similar to Saikaku's, but notes that even if this is not the case, it does not prove that Saikaku did not write the actual text, which he feels exhibits clear ideological commonalities with Saikaku's other work. In addition, Taniwaki (1981) raises various doubts but ultimately assumes that all 17 chapters were written by Saikaku since the inventive ideas would have been beyond the range of his imitators.

C. New doubts raised by Nakamura

Nakamura (1982) mentions Yamaguchi's concerns regarding the handwriting, concluding that the handwriting likely belongs to the author of *Tanba Taro monogatari* 丹波太郎物語 ("The story of Tanba Taro"), not Saikaku.

On the basis of similar arguments, Nakamura actually makes an argument for the identities of the authors of each chapter.

The existence/non-existence of the later insertions to the book has not yet been settled. For that reason, this study re-examines the text of *Yorozu no humihougu* 万の文反古 using a quantitative approach.

III. Database of Saikaku's Works

Table I shows part of the 578,617-word database used for this analysis (beginning with *Yorozu no humihougu* 万の文反古). We morphologically analyzed all of Saikaku's 23 works. This database is the only one on Saikaku's works at present and has a high degree of reliability.

TABLE I. Database of Saikaku's works

Work	Volume	Words	Part of speech	Other information		
万古	巻一	世帯	名詞	セタイ	せたい	せたい
万古	巻一	の	助詞	Δ	の	の
万古	巻一	大事	名詞	ダイジ	だいじ	だいじ
万古	巻一	は	助詞	Δ	は	は
万古	巻一	正月仕舞	名詞	Δ	しゅうがつじまい	しゅうがつじまい
万古	巻一	十二月九日	名詞	Δ	じゅうにがつここのか	じゅうにがつここのか
万古	巻一	の	助詞	Δ	の	の
万古	巻一	書中	名詞	ショチウ	しゅちゅう	しゅちゅう
万古	巻一	、	句読点	Δ	、	、
万古	巻一	伊勢屋十左衛門	名詞	イセー	いせやじゅうざえもん	いせやじゅうざえもん
万古	巻一	船	名詞	フネ	ふね	ふね
万古	巻一	、	句読点	Δ	、	、
万古	巻一	十二日	名詞	Δ	じゅうににち	じゅうににち
万古	巻一	に	助詞	Δ	に	に
万古	巻一	くだりにつき	動詞	連用 Δ	くだりにつく	くだりにつき

Table 1.
Database of Saikaku's works

IV. Analysis

In general, Saikaku's works are made up of many short stories(chapters). We used information of each chapter in our analysis. Then, we compared *Yorozu no humihougu* 万の文 反古 to *Kousyoku ichidai otoko* 好色一代男, as an authenticated work of Saikaku.

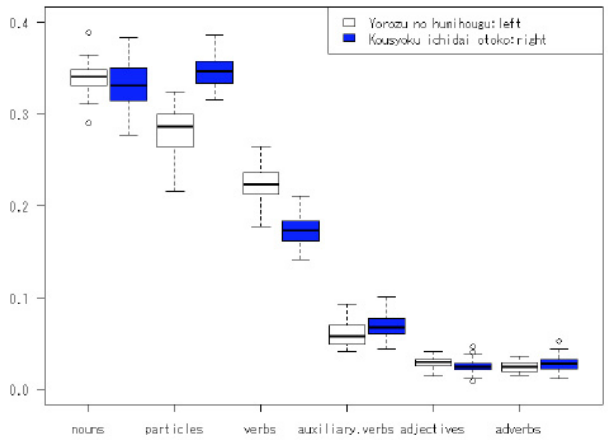


Figure I.
Boxplot (the top six most common classes)

The basic analysis considers appearance ratio of words by grammatical class, using the top six most common classes: nouns, particles, (main) verbs, auxiliary verbs, adjectives, and adverbs for appearance ratio were used in the analysis.

Figure 1 is a boxplot depicting the appearance ratio of these items in both works. We found differences among verbs and particles only.

Furthermore, we examined by welch's t-test at the 0.05 significance level. It concluded that in *Yorozu no humihougu* 万の文反古 and *Kousyoku ichidai otoko* 好色一代男 of verbs and particles using way is different.

Figure II represents the results of the analysis on appearance rate, using principal component analysis (PCA) with a correlation matrix. The horizontal axis shows the importance of first principal component and the vertical axis, the second. Proportion of variance the first principal component is 0.34, while the second is 0.29; the cumulative prpportion up to the second principal component is 0.64.

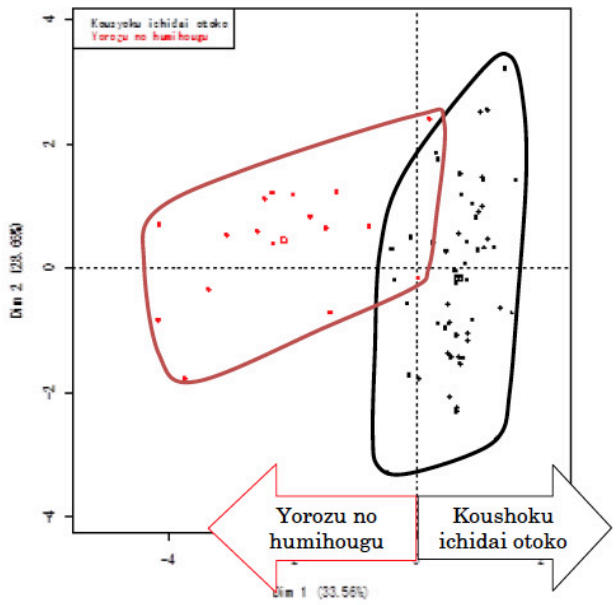


Figure II.
PCA results for Yorozu no humihougu 万の文反古 *and Kousyoku ichidai otoko* 好色一代男 *(the top six most common classes)*

In Figure II, indicating differences revealed by PCA, *Yorozu no humihougu* 万の文反古 is on the left and *Kousyoku ichidai otoko* 好色一代男 on the right.

Upon examining the principal component vector, we find that verbs, adverbs, auxiliary verbs, attributes, and adjectives affect in a positive direction, while particles and nouns affect in a negative direction. (Table II).

Yorozu no humihougu 万の文反古 showed a high appearance ratio for verbs and nouns, while *Kousyoku ichidai otoko* 好色一代男showed a high appearance ratio for auxiliary verbs, particles, and adjectives (Table II). Therefore, it can be said that the appearance ratio of attached words is high in *Kousyoku ichidai otoko* 好色一代男 compared to that in *Yorozu no humihougu* 万の文反古.

TABLE II. The result of PCA

	PC1	PC2	PC3	PC4	PC5	PC6
nouns	−0.23	0.63	−0.13	−0.25	0.54	0.43
particles	0.63	0.14	0.01	0.48	−0.14	0.58
verbs	−0.62	−0.28	0.04	0.02	−0.41	0.61
auxiliary.verbs	0.27	−0.3	0.7	−0.5	0.2	0.24
adverbs	−0.06	−0.61	−0.25	0.32	0.68	0.12
adjectives	0.32	−0.23	−0.66	−0.6	−0.17	0.17
Proportion of Variance	0.34	0.29	0.18	0.11	0.09	0.01
Cumulative Proportion	0.34	0.62	0.80	0.91	0.99	1.00

Table II.
The result of PCA

V. Conclusion

In this study, two works attributed to Saikaku, *Yorozu no humihougu* 万の文反古 and *Kousyoku ichidai otoko* 好色一代男, were compared and analyzed for word class appearance ratios. Significant differences were found.

However, it can be argued that the content and date of each work written will influence word class appearance ratios. Thus, we need to consider this issue from other perspectives and using other data.

References

Munemasa, I. (1969). “Kanazoushi kara Ukiyozoushi he (仮名 草子から浮世へ)” Shibundo.

Yamaguchi, T. (1929). "Saikaku meisakushu ge (西鶴名作集下)" Nihonmeityo zenshu kankoukai.

Teruoka, Y. (1953). "Saikaku kenkyu note (西鶴研究ノート)" Tyuuoukouronsha.

Taniwaki, M. (1981). "Saikaku kenkyu ronkou (西鶴研究 西鶴研究 論攷)" Sintensha.

Nakamura, Y. (1982). “Nakamura yukihiko cyogitsushu (中村幸彦著述集)", Tyuuoukouronsha.

Epistolary voices. The case of Elisabeth Wolff and Agatha Deken

van Dalen-Oskam, Karina
karina.van.dalen@huygens.knaw.nl
Huygens Institute for the History of the Netherlands

The Question

In his book *Computation into criticism* John Burrows analyzed the speech of 44 different characters from novels by Jane Austen using the 30 most frequent function words (Burrows 1987). He showed how even this small amount of high frequency words yielded clearly distinctive results for Austen’s different characters, more so than the characters in novels written by other authors from Austen’s time or from a later time period. Recently, John Burrows and Hugh Craig applied multivariate analysis to the speech of characters in a corpus of seventeenth-century plays (Burrows & Craig 2012). They showed that characters can be distinguished in this way, but that the characters of one playwright usually cluster together compared to the characters of another

playwright. In their research, stylometric analysis and authorship distinction are nicely intertwined.

The assumption behind their research seems to be that it is a natural wish of a literary author to make his or her characters speak thus on paper or on stage as one would expect of a character of that gender, age, social background, etc. But is this true? Could it be based on anachronistic expectations, based on our own horizon of experience? Or is it indeed to be seen as a universal characteristic of all kinds of fiction from all time periods and all cultures?

This assumption should be tested on other (historical) genres. In this contribution I want to find out whether the fictional letter writers in the epistolary novels of two famous Dutch women writers show a significant stylistic differentiation. The case also involves an authorship problem.

The Case

Elisabeth Wolff-Bekker (1738-1804) and Agatha Deken (1741-1804) met in 1776 and immediately became great friends. When Wolff's husband died in 1777, Agatha moved in with Elisabeth and from then on, they closely collaborated on many publications, most important of which are their epistolary novels (Buijnsters 1984). The first one, *The history of Sara Burgerhart*, was published in 1782 and was an immediate bestseller. It was followed by *The history of Willem Leevend*, a much larger and more complex epistolary novel, published in 1784-1785. The one was *The history of Cornelia Wildschut*, almost as long and certainly as complex as their *Willem Leevend* and published in 1793-1796. Much has been written about the two women and their work. Wolff is known to have been a highly educated and very smart and lively woman, whereas Deken, raised in an orphanage, is described as timid and dull. Based on these impressions, many readers and scholars assume that Wolff was responsible for the lively and funny letters (and/or letter writers), and Deken for the dull and simple letters (and/or letter writers). Even during their lifetime, this seems to have been the general idea. In their forewords and in some personal letters they explicitly stated that this was ridiculous: they did everything in close collaboration. Near the end of her life, Deken states she would like to draw up a list of the fictional letters she wrote, to prove these naïve assumptions wrong. She never got around to doing that. Can we, by using stylometric methods, establish how they distributed the work load between them?

The three epistolary novels mentioned above are digitally available at www.dbnl.org . They will be compared to three other epistolary novels that are available in digital form in the same digital library: *Het land, in brieven* written by Elisabeth Maria Post and published in 1792; *Charakters en lotgevallen van Adelson, Héloïse en Elius* by Anna Catherina van Streek-Brinkman (1804); and *De kleine pligten* by Margaretha Jacoba de Neufville (1824-1827). Finally, the fictional letters will be compared with a digitally available corpus of personal letters written by Deken and Wolff, based on the editions of Dyserinck (1904) and Buijnsters (1987).

The Results

We start with an overview, comparing all six epistolary novels with each other. I made use of the stylometric R-script developed by Eder and Rybicki, which performs Principal Components Analysis, Cluster Analysis, Multidimensional Scaling, and Bootstrap Consensus Trees (Eder & Rybicki 2011), choosing the last of these for my analysis since the bootstrap consensus tree is a harmonisation of as many different cluster analysis based on word frequencies as the scholar indicates. The six novels are diverse in length, with a maximum of 585,664 tokens and a minumum of 59,752. In Fig. 1 they have been analyzed in samples of 25,000 tokens.

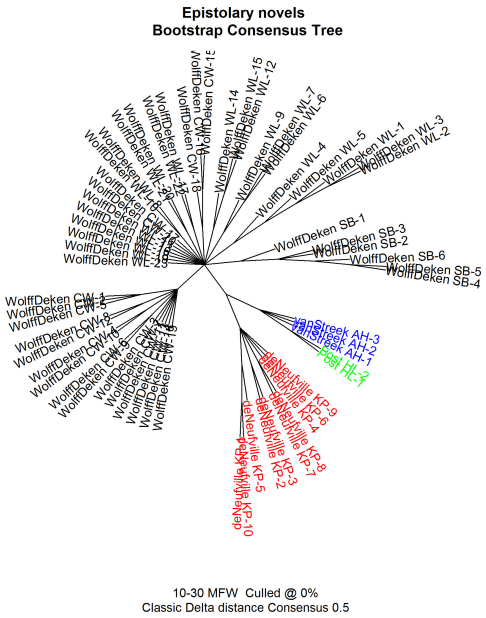


Fig. 1: Epistolary novels

The three Wolff & Deken novels are clearly separated from the novels by the three other (single) authors. This shows that co-authors Wolff and Deken have a distinctive style compared to the other three. A comparable picture arises when we analyse the text of the 30 main characters from all six novels, all having more than 17,000 tokens in their individual corpus.

In the next step we zoom in on one of the Wolff & Deken epistolary novels to find out if the letter writers can be distinguished, and additionally, if the letter writers clearly fall into two different clusters which could be linked to the two different authors. This will be done on their first joint publication, *Sara Burgerhart*, published six years after they met. This epistolary novel has letters and some other texts presumably written by 26 different characters, including one Anonymous and also text written by the authors/narrators. Fifteen of these have a corpus of tokens higher than 2,000 (including the authors/narrators). Fig. 2 shows that the different characters are indeed distinguishable when their complete corpora are analyzed.

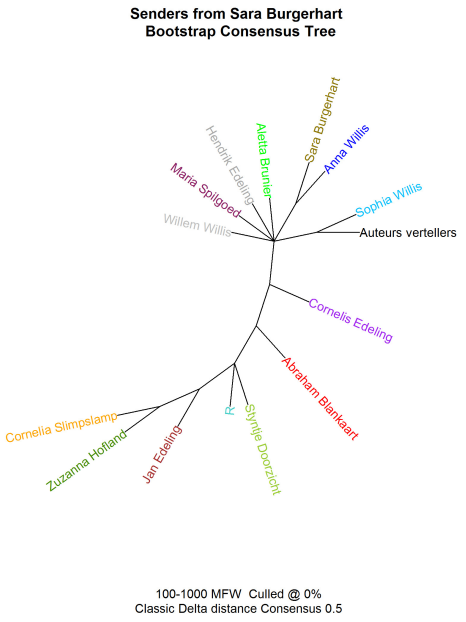


Fig. 2: Senders from Sara Burgerhart (no sampling)

Since the authors/narrators' text (Forewords and Afterwords) in Wolff & Deken's novels is usually undersigned by Wolff only, this consensus tree may show a certain work distribution between the two women. Wolff's style may be prevalent in the letters attributed to the main characters Sara Burgerhart, her women friends Aletta and Anna, and her husband-to-be Hendrik. Through extrapolation, Deken then could be responsible for some of the bad characters in the novel, such as Cornelia Slimpslamp, and Zuzanna Hofland. She could also be the ghostwriter of pious Styntje Doorzicht. And she would indeed be very lively and funny in her letters by Abraham Blankaart. But it is too soon to conclude this; when we use sampling (2,000 tokens), the picture is rudely disturbed and no clear distinction can be found (Fig. 3). Many of the samples are directly connected to the root, which means the software could not convincingly cluster them to any other sample. The only significant branch occurs for the samples of the character Abraham Blankaart, which suggests his letters have a clearly individual style.

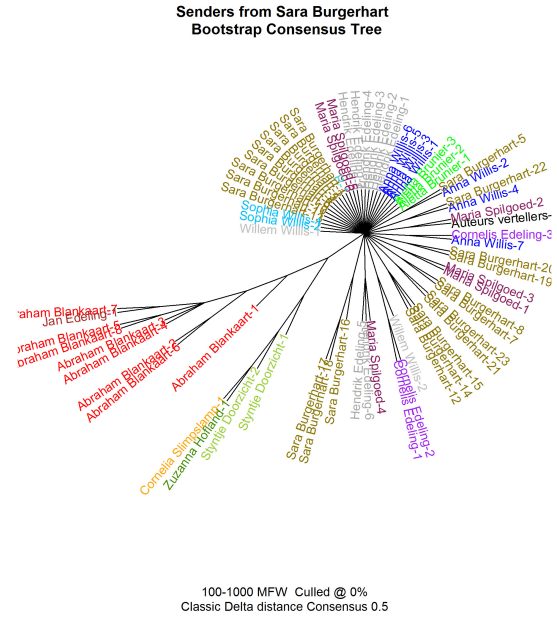


Fig. 3: Senders from Sara Burgerhart (sampling)

The implication is that Wolff and Deken either both worked on the same letters, revising each others' work all along, or that their style of writing is so much alike that they cannot be distinguished. The first option is difficult to prove; historical evidence is not available to confirm this. The second option can be explored by a comparison of the letters in the epistolary novels to the personal letters of Deken and Wolff from the time period in which they wrote and published *Sara Burgerhart* (Fig. 4).

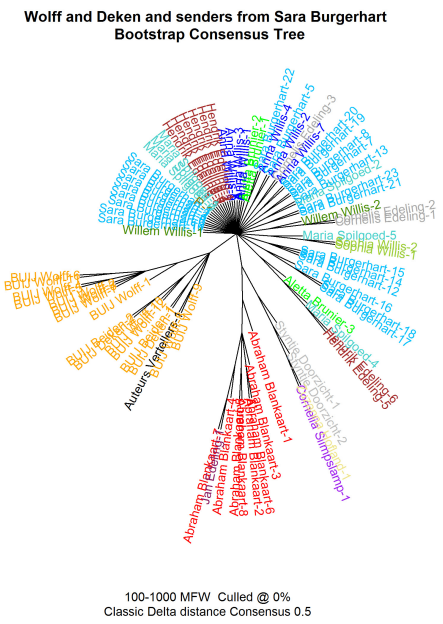


Fig. 4: Wolff and Deken and senders from Sara Burgerhart

Fig. 4 has three additional corpora in the comparison, from the years 1776-1782. Eleven letters written by Deken, 4,594 tokens in total; 35 letters written by Wolff, 25,790 tokens in total; and twelve letters that were written jointly, 5,878 tokens in total. The consensus tree again shows no clear distinction pointing to two clearly different authors. And, again, the letters of Abraham Blankaart show up as a rather distinctive set.

Conclusion

The measurements seem to imply that the writing styles of Deken and Wolff in their epistolary novels were very much alike. This confirms their own statements about their collaboration, and can be explained by their close working relation: they not only wrote the works together, but they were very close friends, even living together.

As to the other question I started out with, the style of different characters, something extraordinary showed up in the graphs. Only one of the main characters of *Sara Burgerhart* clearly had a more individual style than the other main characters, namely Abraham Blankaart. So for now, it seems that a distinctive style for all characters such as John Burrows has shown for the characters of Jane Austen, is not a prerequisite for these epistolary novels or these authors. Still, when reading these novels we do recognize the letter writers. Further research, e.g. with Burrows's Zeta, is needed to find out how the authors exactly did this. This is not the end of what we have to do, however: we also should continue research into the genre of epistolary novels in other languages and time periods, to find out whether stylistic characterization occurs elsewhere in the genre, or whether the aimiable character of Abraham Blankaart is indeed a clear exception.

References

Buijnsters, P. J. (1984). *Wolff en Deken. Een biografie*. Leiden: Martinus Nijhoff

Buijnsters, P. J. (1987). *Briefwisseling van Betje Wolff en Aagje Deken*. Uitgegeven met inleiding en aantekeningen. 2 Dln., Utrecht: HES Uitgevers

Burrows, J. F. (1987). *Computation into criticism. A study of Jane Austen's novels and an experiment in method*. Oxford: Clarendon Press

Burrows, J., and H. Craig (2012). Authors and characters. In *English Studies* 93 (3). 292-309.

Dyserinck, J. (ed). (1904). *Brieven van Betje Wolff en Aagtje Deken*. Den Haag: De Gebroeders van Cleef.

Eder, M., and J. Rybicki (2011). Stylometry with R. In "Digital Humanities 2011: Conference Abstracts." Stanford: Stanford University. 308-11.

Victorian Paratextual Poetics and Citation Analysis

Walsh, John
jawalsh@indiana.edu
Indiana University, United States of America

Sugimoto, Cassidy
sugimoto@indiana.edu
Indiana University, United States of America

This study combines elements of literary theory, citation theory, and techniques of citation analysis to study paratextual elements in the work of Victorian poets Algernon Charles Swinburne (1837–1909), Dante Gabriel Rossetti (1828–1882), and Alfred, Lord Tennyson (1809–1892). The study will focus on Swinburne, who provides the largest and richest collection of paratextual elements, but will also look at Rossetti and Tennyson for the sake of comparison. In his Foreword to the English translation of Gérard Gennette's *Seuils* (translated as *Paratexts: Thresholds of Interpretation*), Richard Macksey defines paratexts as

> those liminal devices and conventions, both within the book (peritext) and outside it (epitext), that mediate the book to the reader: titles and subtitles, pseudonyms, forewords, dedications, epigraphs, prefaces, intertitles, notes, epilogues, and afterwords…but also the elements in the public and private history of the book, its "epitext," that are analyzed in the latter part of this volume: "public epitexts" (from the author or publisher) as well as "private epitexts" (authorial correspondence, oral confidences, diaries, and pre-texts). (Macksey xviii)

In humanities documents, particularly primary source materials, analysis of paratextual elements may provide information about influence, impact, and networks shared by authors, artists, and their works. These networks of intertextual and interpersonal relationships contribute to the production of meaning in the text. Our study will focus on peritexts, that is, the paratexts *within* the document. We are further limiting our focus to paratexts that in some way or another resemble a citation, i.e., those paratexts that explicitly reference one or more bibliographic elements, such as author, title, date, or a quotation that identifies itself as such (e.g, by the use of quotation marks). We look at a number of specific textual elements, including titles, subtitles, epigraphs, dedications, and notes. Titles may reference specific authors (e.g., Swinburne's "Song for the Centenary of Walter Savage Landor" and "From Victor Hugo") and/or works (e.g., Swinburne's "Grand Chorus of Birds from Aristophanes"). Epigraphs include quoted text, often accompanied by more formal citations indicating author, title, and even line numbers. See the example below, one of the epigraphs to Swinburne's lyrical drama *Erechtheus* (1876):

> ΑΤ. τίς δὲ ποιμάνωρ ἔπεστι κ πιδεσπόζει στρατοῦχ;
> ΧΟ. οὔτινος δοῦλοι κέκληνται φωτὸς οὐδ' ὑπηκόοι.
> Æsch. *Pers.* 241 - 2.

By focusing our attention on these "citation-like" paratexts, we are able to apply the techniques of citation analysis to probe these poets' paratextual gestures. Our analysis is complicated and enriched by the presence of "invented" paratextual citations, such as the fictitious French epigraph to Swinburne's "Laus Veneris" (1866).

Dedications provide an interesting case in that they typically refer to a person rather than a specific work. But the overwhelming number of Swinburne's dedication were to men and women of letters and visual artists: Walter Savage Landor, Victor Hugo, Richard Francis Burton, Edward John Trelawny, William Bell Scott, Charles Lamb, Dante Gabriel Rossetti, William Michael Rossetti, Christina Rossetti, Edward Burne-Jones, and William Morris. The target of the dedication, posited as a bibliographic entity in a bibliographic and literary context, becomes a sort of document, like Suzanne Briet's antelope, and represents not just the individual but that individual's body of work.

Swinburne's notes, like the epigraphs, often include more formal citations. In fact, a number of long odes —to favorite authors Walter Savage Landor and Victor Hugo — may be characterized as bibliographies in verse, accompanied by notes containing formal citations to the works described in the verse. The "Birthday Ode" (1880) to Victor Hugo contains thirty-eight notes, each with a citation to one or more works by Hugo. Harold Nicolson called this poem "a complete rhymed bibliography of the works of Victor Hugo."

Swinburne is certainly not unique among nineteenth-century poets in buttressing his poems with a rich paratextual framework. Other noteworthy examples include the Preface to *Lyrical Ballads*, Coleridge's introductory note to *Kubla Kahn* and his glosses to *The Rime of the Ancient Mariner*, Byron's notes to *Childe Harold's Pilgrimage*, and Shelley's notes to *Queen Mab*. But Swinburne was incredibly widely read, a scholarly poet, and an enthusiastic bibliophile. The paratextual network that surrounds Swinburne's work is particularly large, diverse, and complex, and in the context of information studies his work provides fertile ground for this sort of bibliometric analysis. Swinburne's poems and paratexts reference a large number of explicitly identified works by other writers and artists. The six volumes of Swinburne's collected *Poems* contain over four hundred poems, large and small. Within these roughly four hundred poems, we have identified over four hundred and fifty paratextual citations, references embedded within Swinburne's titles, epigraphs, notes, and so on. The works cited by Swinburne cover the range of Western culture from the Old and New Testaments of the Bible to classical antiquity through the middle ages, Renaissance, and Enlightenment to contemporary 19th-century works by figures such as Landor, Hugo, Tennyson, Gautier, Robert Browning, Whitman, and Dante Gabriel Rossetti. Besides English, languages represented by Swinburne's paratexts include French, Greek, Italian, and Latin. His paratexts also reference visual art (Rossetti, Whister) and music (Wagner).

The Swinburne documents analyzed in this study have been digitized and are available in TEI/XMLencoded formats through the Algernon Charles Swinburne Project http://www.swinburneproject.org/ . The TEI provides markup elements that map fairly neatly to the types of paratexts and bibliographic entities under consideration here:

- `<epigraph>`
- `<title>`
- `<author>`
- `<date>`
- `<quote>`

- `<div type='dedication'>`
- `<ab type='dedication'>`

Since these elements were already encoded in the TEI/XML, it was relatively trivial to extract them from the enclosing texts and identify those that contain paratextual references. Swinburne's references, or "citations," were entered into a database that was augmented with more detailed bibliographic information about Swinburne's sources. This bibliographic information was then analyzed to provide different "views" of Swinburne's paratexts. The analysis provides us with feedback about:

- frequency of paratextual types (i.e., titles, epigraphs, notes, etc.)
- frequency of cited authors/artists
- genres of cited works
- literary/historical periods of cited works
- languages of cited works
- trends in paratextual practice across Swinburne's career

The text/document sits at the intersection of the humanities and information studies, and a consideration of the paratext brings together theoretical concerns about the material document and intertextuality from literary studies with theories about citation and methodologies of citation analysis from information studies.

In an effort to broaden the study and provide comparison data, we have also gathered information about paratexts in the poetry of Swinburne's contemporaries, Rossetti and Tennyson. Our paper will provide a comprehensive examination of a large, defined subset of the paratexts found in the work of a three important Victorian poets. Our analysis will provide different views of these poets as seen through their paratexts and the diverse collection of authors, artists, and works referenced by those paratexts. We will present the findings from a thorough analysis of paratextual reference in Swinburne, Rossetti, and Tennyson, and based on those findings, we will make some preliminary observations about Swinburnian Victorian paratextual poetics.

References

Briet, S. (2006). *What is Documentation?*. Trans. Ronald E. Day and Laurent Martinet. Lanham, MD: Scarecrow Press.

Genette, G. (1997). *Paratexts: Thresholds of Interpretation*. Trans. Jane E. Lewin. Cambridge: Cambridge University Press.

Macksey, R. (1997). Foreword. *Paratexts: Thresholds of Interpretation*. By Gérard Genette. Trans. Jane E. Lewin. Cambridge: Cambridge University Press. xi-xxii.

Nicolson, H. (1926). *Swinburne*. New York: Macmillan.

Swinburne, A. C. (1904). *The Poems of Algernon Charles Swinburne*. 6 vols. London: Chatto & Windus.

Walsh, J. A., (ed.) (2011). *The Algernon Charles Swinburne Project*. Indiana University. 1 Nov. 2011. http://www.swinburneproject.org/ .

User ethnographies: informing requirements specifications for Ireland's, national, trusted digital repository.

Webb, Sharon
sharon.webb@nuim.ie
An Foras Feasa, Ireland

Keating, John
John.Keating@nuim.ie
An Foras Feasa, Ireland

1 Introduction

The Digital Repository of Ireland (DRI) is an interactive national trusted digital repository for contemporary and historical, social and cultural data held by Irish institutions; providing a central internet access point and interactive multimedia tools, for use by the public, students and scholars. DRI is a four-year exchequer funded project, comprising six Irish academic partners, and is supported by the National Library of Ireland, the National Archives of Ireland (NAI) and the Irish national broadcaster, RTÉ.[1]

DRI is in its second year and has completed a number of project deliverables, including initial requirements statements, a lean prototype with core functionality and a national practice report, *Digital archiving in Ireland: National survey of the humanities and social sciences* (O'Carroll et al. 2012). This report is based on the findings from our requirements interviews which represent the first phase of engagement between DRI and its stakeholders and can be described as a 'descriptive ethnographic

contribution' (Wynne et al. 2012). The principle aim of the interviews was to inform the system's business, functional and nonfunctional requirements as well as to drive policy decisions and inform guidelines on issues such as metadata, file formats and access rights. Since DRI's remit extends to the humanities as well as the social sciences the interviews revealed a diverse problem domain (Hull et al. 2011). However, even within this diverse space the interviews uncovered many similarities, shared problems and challenges among the community of users.

The diversity of DRI's community, that is national cultural institutions, university libraries and other higher education institutions, as well as their respective research institutes, national broadcasters and various independent and state bodies, requires reconciliation between their various perspectives. This paper will discuss requirements specifications in light of these stakeholder interviews and the national report, which form a crucial part of the information gathering phase of requirements engineering, and consider how user ethnographies can enhance our understanding of the user and their software needs.

2 Methodology - requirements gathering and qualitative interviews.

The generation of use cases and use case scenarios are paramount to the successful development, and completion, of digital humanities projects, resources and artifacts and is crucial in the first phase of requirements engineering, that is requirements elicitation and information gathering. Use cases inform the system that will be and advance the development teams' knowledge of their end user, focusing attention on authentic users and their needs. These use cases, and the actors involved, must be linked to some, but perhaps augmented, reality. Analysis of the various actors is required to develop use case scenarios that consider the context in which the future system, that is the Digital Repository of Ireland (DRI), will be used.

To achieve this we used a qualitative approach to requirements elicitation and carried out extensive stakeholder interviews (we completed 40 separate interviews between Nov. 2011 - Aug. 2012). While alternative approaches to requirements elicitation exist, and include quantitative methods such as online voting and user questionnaires, as well as traditional JAD (Joint Application Development) methods, (Maciaszek et. al. 2001) we elected the use of semi-formal and topic driven requirements interviews because they enhanced our user engagement and helped to develop important ties and relationships with our community (O'Carroll et al. 2012). This approach allowed us to incorporate our requirements interviews with policy management and ensured that we addressed the key concerns of the community and developed strategies for digital rights management, digital preservation, access control, digital standards, among others. The interviews were 'an inquiry' rather than 'an inquisition', (Weigers et al. 2006) and as such required close analysis to extract requirements, as they were not explicitly stated or expressed. The result of this analysis is requirements statement as well as descriptive user ethnographies linked to specific activities, namely preservation, interaction and access.

Ethnographic research, as a core feature of anthropology and sociology, encourages us to observe social and cultural norms and can help us interpret individual, as well as group, behaviour, activities and practices. As a research methodology it is concerned with 'why' we perform and engage in certain cultural and social norms and as such can help us understand technological practices and workflows associated with particular software systems. Rönkkö et al. (2002), describe ethnographic research as a means to emphasise 'the members' point of view' and as a method which can help us 'understand the organisation of social, cultural and technical setting[s]'. From a software engineering view point user ethnographies can provide an 'inside perspective' (Rönkkö et al. 2002) and help reveal the different methods, practices, and indeed agendas, of the community. Some of these 'inside perspectives' are captured in our national report which considers key topics such as digital preservation, user tools, file formats and metadata standards (O'Carroll et al. 2012).

A user ethnography, in this context, is essentially a composite analysis of the methods and practices identified as the salient features of the system to be, which we considered as part of the social, cultural and institutional norms of particular methods, behaviors and user activity linked to digital archiving. It is a term borrowed from the anthropological practice which may involve participant observation or immersion in a particular group or society. However, this was not possible given resource limitations and the scope of this project. Instead, we chose requirements interviews, focus groups with a subset of users, feature development through executable requirements specification and online surveys to develop the user ethnography, the extended use case scenario, to provide a holistic view of the problem domain in order to inform the solution.

3 Reconciling 'the members' point of view' - requirements engineering and user ethnographies

Requirements engineering is a key aspect of the software development cycle and is a necessary activity for any small

or large scale project. The methods developed within DRI to capture the requirements were informed by the need to capture a 'panoramic view' of DRI's community (Passos et al. 2012). A high level consideration of requirements was necessitated given the broad range of users and their needs. The ethnographic approach, that is considering the setting, the field or indeed the culture of our designated community, allowed us to 'focus on the participants and their interactions in [the] system rather than the data, its structure and its processing' (Sommerville et al. 1993).

From this analysis we developed our requirements statements which were expressed as structured, natural language statements as well as executable specifications, developed as Cucumber tests or features which support user testing and the development of live documentation (Wynne et al. 2012). Cucumber features are also key to the ethnographic approach as they are specified collaboratively with the software development team and key stakeholders in DRI. This allows us to observe and analyse particular activities related to core functionality and provides essential feedback, validation and verification of the requirements specification.

Feature development is key to the reconciliation process as it allows us to test particular high level requirements in terms of specific user goals. For example, feature development with an audio archive highlighted a missing requirement in terms of time coding audio content in comparison to "tagging" associated with images based content. Through this process we try to reiterate the importance of the informants point of view (the problem rather than just the solution) and in this sense we view feature development as part of "participant observation" techniques.

4 Conclusion

To develop our requirements specifications, we used a qualitative methodology to drive requirements gathering and have similarly applied a qualitative method, that is user ethnographies, to reconcile the various perspectives of the community. Our national practice report represents the first analysis of DRI's requirements but reconciling the various perspectives requires further, more thorough analysis of each domain, that is humanities and social sciences, within the context of three separate activities, preservation, access and interaction. This paper will consider the development of these user ethnographies, a method which is an important aspect of software engineering and requirements engineering and how they inform the final stages of requirements specification.

References

O'Carroll, A., and S. Webb (2012). Digital archiving in Ireland. National survey of the humanities and social sciences. National University of Ireland Maynooth available at www.dri.ie/publications

Wynne, M., and A. Hellesøy (2012). The Cucumber book, behaviour-driven development for testers and developers. *Pragmatic Programmers*.

Hull, E., K. Jackson, and D. Jeremy. (2011). Requirements engineering. *Springer*.

Maciaszek, L. A. (2001). Requirements analysis and system design, developing information systems with UML. Addison-Wesley.

Weigers, K. E. (2006). More about software requirements, thorny issues and practical advice. Microsoft Press.

Rönkkö, K., O. Lindeberg, and Y. Dittrich. (2002). 'Bad practice' or 'bad methods'. Are software engineering and ethnographic discourses incompatible? *Proceedings of the International Symposium on Empirical Software Engineering*, Nara, Japan 204-10.

Passos, C., D. S. Cruzes, and T. Dybå (2012). Challenges of applying ethnography to study software practices. *Proceedings of the ACM-IEEE international symposium on Empirical software engineering and measurement* 9-18.

Sommerville, I., T. Rodden, P. Sawyer, R. Bentley, and M. Twidale (1993). Integrating ethnography into the requirements engineering process. *Proceedings of the IEEE International Symposium onRequirements Engineering* 165-73.

Sharp, H. (2012). Keynote: Using ethnography in empirical software engineering. 16th International Conference on Evaluation & Assessment in Software Engineering.

Notes

1. See www.dri.ie

Mapping Text: Automated Geoparsing and Map Browser for Electronic Theses and Dissertations

Weimer, Katherine H.
k-weimer@library.tamu.edu

Texas A&M University Libraries, United States of America

Creel, James
jcreel@library.tamu.edu
Texas A&M University Libraries, United States of America

Modala, Naga Raghuveer
raghuravi23@tamu.edu
Texas A&M University, Dept. of Biological and Agricultural Engineering

Gargate, Rohit
rohitvg@gmail.com
Texas A&M University, Dept. of Computer Science and Engineering

While texts contain extensive mentions of locations, traditional library catalogs are lacking when searching for geographic locations. Ahlers and Boll (2007) state that approximately twenty percent of web queries have a geographic relation. Buckland (2007), Hill (2006) and others have emphasized the need for collections to be searchable using geographic means. With the advent of visualization tools and web based mapping, there are numerous possibilities to gain insight into the geographic content of texts. Gregory, and Hardie (2011), Bodenhamer, Corrigan and Harris (2010), Dear, Ketchum, Luria and Richardson (2011) and others discuss the role of geographic information in the humanities. Texts, maps and photographs have been described using map interfaces; however, the grey literature of graduate scholarship output has not thus far been presented graphically. Researchers at Texas A&M University Libraries are taking theses and dissertations, geoparsing those texts, and creating a visual, map-based search interface in order to glean better understanding of the locations and topics presented in these scholarly works.

The ETDMap is a prototype which automatically discerns the places mentioned in digital documents(i.e. geoparsing)and through a series of automated steps creates a map to browse the collection. Researchers gained conceptually from work outlined by Grover, et al (2010) and Leidner (2007). This geoparser operates in the context of a DSpace institutional repository. Development has focused on the electronic theses and dissertations collection, although the software is applicable to any textual content in the repository. This abstract will present the current status and overview of the geoparser and map search interface tool.

The geoparser is implemented as a curation task in the DSpace repository, using the Java programming language. The geoparser automatically parses the text document, dividing it into sections and identifying prospective toponyms. The geoparser excludes certain portions of the document, such as bibliographies and appendices, since the toponyms found in those sections are typically not directly related to the subject matter of the text. The toponyms are filtered according to several heuristic criteria, and then are used to construct queries through the GeoNames Web-based Java API. The GeoNames server returns a set of locations that match the queries. The geoparser then applies disambiguation heuristics to score the locations and determine the most likely referent of the toponym. Finally, the top-scored locations, along with geospatial metadata (including coordinates) are written to metadata fields on the item under consideration. These metadata are then output to a KML file for viewing in a variety of interfaces. The term 'map' used in Fig. 1 refers to a data structure map, not a geospatial visualization.

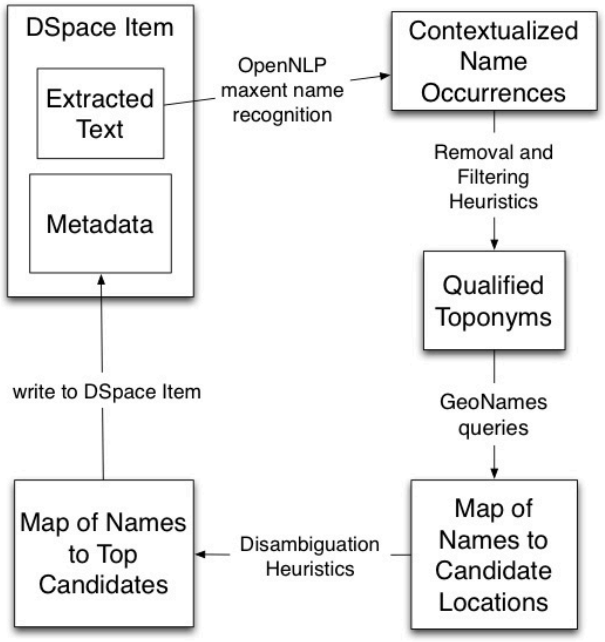


Figure 1.
Geoparsing Workflow [image credit: James Creel]

The geoparser uses regular expressions to partition the document text into sections. Theses and dissertations follow a predictable and regulated document format, which allows for clean results. Currently, the sections of interest are the abstract and main document body. These are processed by the geoparser, while the references, vita and appendices are ignored. The geoparser identifies toponyms in several stages using the OpenNLP maximum entropy software library[i] The ETDmap utilizes training data available on the OpenNLP Models page.[ii] The detected potential toponyms are stored along with contextual information,

such as the number of occurrences and the locations of those occurrences, all of which is used in subsequent heuristic processing.

The locations referred to are discerned by a variety of heuristics. The primary Java entities used in the process are the CandidateMapper, the RefinementImpl(Implementation) and DisambiguationHeuristic. Pruning heuristics, which eliminate spurious prospective toponyms, are being implemented, and are under review and refinement. Those include heuristics that ignore short or common words, ignore singe occurrences, and require exact matches to records found in GeoNames. Additionally, scoring heuristics add points to scores associated with the possible toponyms. Populated places receive higher scores, as do those closer to or contained within other candidate locations. Once the stock of heuristics has been exhausted, the candidates with the top scores are selected as referents of the toponyms.

The Generate KML curation task reads the geospatial metadata thus generated (including geospatial coordinates provided by the GeoNames server) and encodes it in a KML file attached to the item in DSpace. This KML file includes placemarks for each of the mentioned places and includes description on each placemark with the title, author, advisor, url, date and department. At the collection level, the repository supplies a link to a KML file generated on request that consists of the aggregation of all the KML files generated for items in the collection.

GeoNames.org was selected as the gazetteer for the project due to its inclusion of numerous official gazetteers from countries around the world, and because it is easily and freely downloadable, so therefore practical for use in this case. In terms of visualization, the initial map background was OpenLayers WMS. It provides a simple and easy to use interface and is open source, but did not provide great detail when zoomed in.(See Fig. 2.) Other map backgrounds included are GoogleMaps and Open Street Maps. Open Street maps is open source, but, includes names in the native language of the country, so is not universally user friendly. Google is not open source, but has a nicely displayed product. In our current version, the user has the choice to select which map background they would like displayed by clicking on a button at the upper right hand portion of the map.

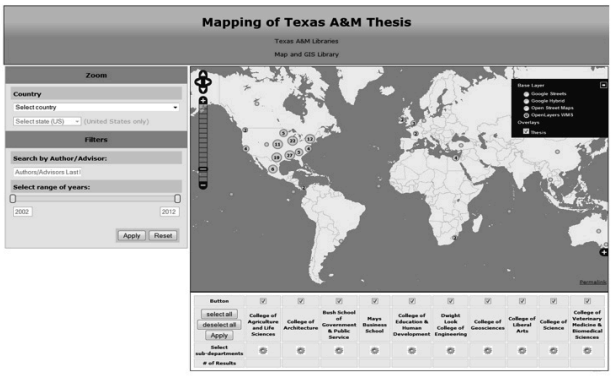


Figure 2.
Map Using OpenLayers WMS Interface

The metadata created through the KML curation task are used not only to create the points of interest for the map viewer, but to also serve as the database for the search filters. During the early development stage, metadata fields were expanded in the KML to include official place names and geographic coordinates for each location mentioned in the text as well as the work's author, title, advisor, url, date of publication, and academic department. The latest version of the map includes a time slider search for date of publication, a keyword search for author or academic advisor name, and check boxes for academic college and/or department. Clustering of results enables the user, once zoomed into their areas of interest, to further refine the browse as the results are then broken into smaller groupings(Fig. 3).

Figure 3.
Zoomed in View, Showing Expansion of Clusters

Once the user has pinpointed a location of interest, they may click on the pointer and be forwarded to the full text document located in the university's institutional repository, DSpace. (Fig. 4)

Figure 4.
Selected Item Showing Title and Metadata Linked to the Full Text

Future Research

Research continues on refinement and development of the geoparser. Two short-term goals figure prominently in current efforts: an evaluation of the tool, and implementation of a statistical classifier as an augmentation to heuristic-based geoparsing.

Evaluation of the tool presents complications for the traditional precision/recall metrics of information retrieval. While these metrics are easily applicable to the disambiguation task, their application to the name extraction task is less straightforward. The mere occurrence of a toponym in a text does not indicate its relevance to the subject matter. We recognize and deal with certain negative cases by ignoring particular document sections like vitae, references, and appendices, but passing mentions of places occur in body text as well. We plan to implement a statistical comparator for target documents and a set of pre-selected documents known to refer meaningfully to particular places(encyclopedia articles, for instance). The statistics used for the comparator will include term-vectors or other textual derivatives. Techniques gleaned from this development will likely find application in the disambiguation task as well.

We have prepared a set of manually identified and disambiguated toponyms for approximately 100 theses as a basis for our pending evaluation of the toponm disambiguation task. Evaluation of the toponym extraction task will require more subtlety, perhaps including a user study to assess human understanding of the relevance of particular place mentions to document subject matter. Additional user studies will be conducted to enhance usability of the map interface. Finally, we plan to apply the mature application to collections beyond electronic theses and dissertations.

Funding

This research was supported in part by AMIGOS Library Services [Fellowship 2010-2012 to K. Weimer].

References

Ahlers, D., and S. Boll. (2007). Location Based Web Search. In Sharl, A. and Tochtermann, K. (eds), *The Geospatial Web: How Geobrowsers, Social Software and the Web 2.0 are Shaping the Network Society.* London: Springer. 55-66.

Bodenhamer, D., J. Corrigan, and T. Harris (eds). (2010). *The Spatial Humanities: GIS and the Future of Humanities Scholarship.* Bloomington: Indiana University Press.

Buckland, M., A. Chen, F. Gey, R. Larson, R. Mostern, and V. Petras. (2007). Geographic Search: Catalogs, Gazetteers, and Maps. *College & Research Libraries* 68(5): 376-387.

Dear, M., J. Ketchum, S. Luria, and D. Richardson (eds). (2011). *Geohumanities: Art, History, Text at the Edge of Place.* New York: Routledge.

Gregory, I., and A. Hardie (2011). Visual GISting: Bringing Together Corpus Liguistics and Geographical Information Systems. *Literary & Linguistic Computing* 26(3): 297-314.

Grover, C., et al. (2010). Use of the Edinburgh Geoparser for Georeferencing Digitized Historical Collections. *Philosophical Transactions of the Royal Society A* 368. 3875-3889.

Hill, L. L. (2006). *Georeferencing: The Geographic Associations of Information.* Cambridge, MA: MIT Press.

Leidner, J. (2007). *Toponym Resolution in Text.* Doctoral Dissertation. Edinburgh: University of Edinburgh. http://hdl.handle.net/1842/1849 (accessed 3 November 2012).

Notes

i. http://incubator.apache.org/opennlp/
ii. http://opennlp.sourceforge.net/models-1.5

Computer Identification of Movement in 2D and 3D Data

Wiesner, Susan L.
slywiesner@virginia.edu
University of Virginia; Society of Dance History Scholars

Bennett, Bradford C.
bcb3a@virginia.edu
University of Virginia

Stalnaker, Rommie L.
rstalnaker81@gmail.com
Independent Scholar/Artist; San Diego United

Simpson, Travis
tts2q@virginia.edu
University of Virginia

Wang (2003) noted that the majority of research conducted into movement patterns has been for surveillance, while 'the breadth of uses for motion analysis beyond surveillance includes sports and ballet'. Although scholars have long recognized the need to employ digital analysis of human movement, especially in dance, the technology for such analyses has not been available. The ARTeFACT project, funded by an NEH Digital Start-up Grant and an ACLS Digital Innovation Fellowship, addresses this lack by developing methodologies for computer identification of complex dance movements in 3-D, through initial work towards the ability to have computer analysis and tagging of 2-D dance film.

Although other projects have had goals specific to dance research and the use of motion capture, we have discovered only one other project that considers a grammar of movement to enable the segmentation, recognition, retrieval, and qualitative analysis of movement from 3D data as we envision (Choensawat et al., 2009). This paper discusses the ARTeFACT approach: the methods used to describe movement, the relationship of NLP to the non-verbal language of dance, and the technology used to automatically identify dance movement in 3-D data.

Ontology, Mo-Cap and Movement Identification

One of the challenges in developing a grammar based on movement is the lack of any clear distinctions between movements, relative to the spaces and punctuation delimiting words in NLP. Thus we must find other means of identifying the patterns and features of movements within a corpus of movement texts. As a first step we developed an ontology, stored in XML files which contain fields and variables pertinent to an individual STEP. STEPS are defined as specific codified movements performed in isolation or in combination with other STEPS; the codified movements are typically performed as part of technique class, and are subsequently used in choreographic works as sections of movement phrases. Each STEP entry includes spatial level, body part, effort, genre, style, relationship to other STEPs (IS A, IS A PART OF, CONTAINS), terminology (folksonomic, codified), and any movement synonyms (movements in one genre synonymous with a movement in another genre, e.g. 'ecarte derrier' in ballet and 'tilt' in modern dance).

Using a motion capture system we captured over 100 codified steps and phrases from Ballet, Modern, Jazz and Tai Chi Chuan genres. At least three good trials of each move were captured from which we developed a movement database or library of codified moves, with *model moves* determined by selecting a representative trial for each move. Data was collected at 120 Hz with a VICON 8 camera motion capture system on two highly qualified performers — a professional ballet dancer of 15 years and a practicing Tai Chi Chuan expert of 25 years. Per the modified Plug-In Gait full body marker set, 38 infrared reflecting markers were placed on the performers. The library contains joint position data, classification information (e.g. relationship of feet to ground, number of occurrences, traveling movements, and rotation around the pelvic center). Movement identification was provided by custom MATLAB code, idMove, which reads the joint positions and uses a combination of classification and pattern recognition to identify a dance move.

Data analysis was performed with validated software written in VICON's Bodybuilder language, which generated the 3-D marker kinematics used in the analysis (Bennett et al. 2005). The output of this software contained 3-D positions of the joints over time, which was the basis for kinematic analysis. Although other parameters such as joint angles and movements were available, only marker position data was used as it was deemed the most compatible with the eventual goal of identifying moves from 2-D film. Classification used empirically pre-set threshold values to determine whether certain conditions were met, e.g. whether a foot was off the ground. idMove computed ten independent time series for the model and the test move — transverse proximity of knees and ankles (2) and both orientations of bilateral hip-knee and knee-heel elevation (8) and the appropriate time series of model and test move were cross-correlated.

A total of 181 trials of 93 unique codified movements were tested with and without classification filtering. Both methods correctly identified the majority of the trials; however, using correlation alone was more accurate. The algorithm using classification as a way to reduce potential

model moves produced a sensitivity of 84.5% over the 181 codified trials with an average combined correlation coefficient of 0.856. Additionally, there were 7 false positives and 21 trials not matched with any move. Overall, 88 of the total 93 codified moves were identifiable by at least one trial. When using only correlation in the process of matching moves, the results were greatly improved, demonstrated by a sensitivity of 97.3% and false positives were reduced to a total of five with all unique moves identifiable in at least one trial.

Abstract Movement and Conceptual Metaphor

As helpful as it is to identify codified movements, in reality many movements in dance and other movement-based activities are not codified. Further, earlier research has shown that dance viewers rarely describe movements; instead they note an interpretive, subjective response to movement. Therefore, while codified movements were a logical first step for our research, we chose to use Lakoff and Johnson's research on conceptual metaphor and embodied knowledge to develop another means of classifying and identifying movement patterns (Lakoff and Johnson 1980; Johnson 2007). By incorporating abstract movement into our library we are able to conduct research into semantic descriptions of human motion in complex unconstrained activities. This is the first time that conceptual metaphor has been used to study the meaning of movement for the purposes of automatic recognition and segmentation of 2D and 3D data strings.

With movements from seven dance works that represent conflict (war) and contain both codified and abstract movement, we categorised and captured 396 different sections of movement, performed by two dancers. We found that movements are shared across the seven dance works studied, specific to 19 different CONFLICT terms listed in the *Collins Cobuild Dictionary of Metaphor* (Diegnan 1995). Therefore, in order to overcome the challenge of segmentation inherent in abstract movement we have developed rules for identifying similar patterns within a specific context (in this case, metaphoric terms). For example, joint angles of specific body parts can be identified per conflict term (e.g. acute knee and hip angles in 'struggle'), as can the relation of foot to floor (jumps appear often with 'hero' and 'victory'). Also, as choreographers may include codified movements within an abstract string, we are able to identify those through the application of the model moves in the codified library. By creating the rules and clearly defining the movement specifics relative to the metaphor terms, we can generate 'start and stop' points within the string of data, thus replicating the use of spaces to delineate words in NLP methodologies.

In addition to developing rules, we have conducted statistical analysis on the frequency of movements and movement durations per term providing information regarding movement quality, as well as quantity. The most frequently occurring terms are Victim, Struggle and Attack in descending order, although Struggle consumes more time than Victim. Analysis of the empirical data and testing of idMove with the abstract movements associated with Conflict terms is ongoing, as is the development of a grammar to which NLP methods can be applied (i.e. the derivation of tokens from 2D data that are the building blocks for expressions of dance movement).

Conclusion

The ARTeFACT project — which brings together scholars and artists across the sciences, humanities, and arts — is making several scientific and technical advancements in the fields of motion analysis and image pose recognition. This initial work demonstrates our ability to identify codified dance moves using 3-D data from film of a single performer. Consistent with our goal of using only 2-D data, the analysis relied on elevation changes as the primary criteria for recognition. With the assistance of pose recognition software for film, elevation changes in joint centers could be directly applicable to a similar analysis on single camera video. In addition to adding to the ability to identify the semantics of movement in 2D and 3D data and to provide data mining of abstract movement based on concepts as well as codified movements, this is a unique use of conceptual metaphor against a corpus of movements. Further, the knowledge gleaned from this project extends to research into movement-based disciplines (dance, kinesiology, sports medicine, anthropology, etc.)

References

Ahmad, K., A. Salway, J. Lansdale, H. Selvaraj, and B. Verma (1998). (An)Notating Dance: Multimedia Storage and Retrieval. in *Conference Proceedings, International Conference on Computational Intelligence and Multimedia Applications*. held in Singapore. *World Scientific*. 788-793.

Allen, F. R., E. Ambikairajah, N. H. Lovell, and B. G. Celler (2006). Classification of a known sequence of motions and postures from accelerometry data using adapted Gaussian mixture models. *Physiological Measurement*. 1-17.

Bao, L., and S. S. Intille (2004). Activity recognition from user-annotated acceleration data. In *Proceedings of*

the 2nd International Conference on Pervasive Computing. 1-17.

Bennett, B. C., M. F. Abel, A. Wolovick, T. Franklin, P. E. Allaire, and D. C. Kerrigan (2005). Center of Mass Movement and Energy Transfer During Walking in Children With Cerebral Palsy. *Archives of Physical Medicine and Rehabilitation* 86. 2189-2194.

Bobick, A. F. and J. W. Davis (2001). The recognition of human movement using temporal templates. *IEEE Transactions on Pattern Analysis and Machine Intellligence* 23 (3).

Campbell, L. W. and A. F. Bobick (1995). Recognition of human body motion using phase space constraints. *ICCV, 5th Int'l. Conf on Computer Vision. (ICCV '95)*. 624.

Choensawat, W., W. Choi, and K. Hachimura (1995). A quick filtering for similarity queries in motion capture databases. *PCM 2009, LNCS 5879*. 404-415. In Diegnan, A. (ed). *Collins Cobuild English Guides 7: Metaphor*, New York: HarperCollins.

Ermes M., J. Parkka, J. Mantyjarvi, I. Korhonen (2008). Detection of daily activities and sports with wearable sensors in controlled and uncontrolled conditions. *IEEE Transactions on Information Technology in Biomedicine* 12. 20-26.

Golshani, F., P. Vissicaro, and Y. Park (2004). A Multimedia Information Repository for Cross Cultural Dance Studies in *Multimedia Tools and Applications.* 89-104.

Goulermas, J. Y., A. H. Findlow, C. J. Nester, P. Liatsis, X. J. Zeng, L. Kenney, P. Tresadern, S. B. Thies, and D. Howard (2008). An instance-based algorithm with auxiliary similarity information for the estimation of gait kinematics from wearable sensors. *IEEE Transactions on Neural Networks* 19. 1574-1582.

Johnson, M. (2007). *The Meaning of the Body: Aesthetics of Human Understanding.* Chicago: The University of Chicago Press.

Kanan, R., F. Andres, and C. Guetl (2009). DanVideo: an MPEG-7 authoring and retrieval system for dance videos. *Multimed Tools. Appl.* 46. 545-572.

Lakoff, G., and M. Johnson (1980). *Metaphors We Live By*. Chicago: The University of Chicago Press.

Lau, H. Y., K. Y. Tong, and H. Zhu (2008). Support vector machine for classification of walking conditions using miniature kinematic sensors. *Medical and Biological Engineering and Computing*. 46. 563-73.

Lausberg, H., and H. Sloetjes. (2008). Gesture Coding with the NGCS-ELAN System. *Proceedings of Measuring Behavior 2008.* Maastricht, The Netherlands.

Lester, J., T. Choudhury, N. Kern, G. Borriello, B. Hannaford. (2005). A hybrid discriminative/generative approach for modeling human activities. *Proceedings of the International Joint Conference on Artificial Intelligence*.

Lester, J., T. Choudhury, and G. Borriello. (2006). A practical approach to recognizing physical activities. *Pervasive Computing LNCS*. 3968. 1-16.

Miles-Board, T., Deveril, W. Hall, and J. Lansdale. (2012). Decentering the Dancing Text: From dance intertext to hypertext. http://eprints.soton.ac.uk/id/eprint/257304 Accessed on 10-08-2012.

Moeslund, T. B., A. Hilton, and V. Krüger. (2006). A survey of advances in vision-based human motion capture and analysis. *Computer Vision and Image Understanding*. 104. 90-126.

Qian, G., F. Guo, T. Ingalls, L. Olson, J. James, and T. Rikakis. (2004). "A gesture-driven multimodal interactive dance system." *Multimedia and Expo, 2004. ICME '04 IEEE Int'l. Conference*. 3. 1579-1582.

Ramadoss, B., and K. Rajkumar. (2007). Semi-automated Retrieval of Dance Media Objects in *Cybernetics and Systems: An Int'l. Journal*. 38. 349-379.

Reng, L., T. Moeslund, and E. Granum. (2006). Finding motion primitives in human bodygestures. In S. Gibet, N. Courty, and J.-F. Kamp. (eds). *Gesture in Human-Computer Interaction and Simulation, Lecture Notes in Computer Science* 1. 133-144. Berlin: Springer.

Reside, D. (2007). The AXE Tool Suite: Tagging across time and space. Proceedings from *Digital Humanities*. Urbana-Champaign: Illinois. 178-179.

Salway, A. (1999). *Video Annotation: the Role of Specialist Text*. PhD Thesis. University of Surrey.

Wang, L., W. Hu, and T. Tan. (2003). "Recent developments in human motion analysis." *Pattern Recognition.* 36. 585-601.

Wiesner, S., B. Bennett, and R. Stalnaker. (2011). *ARTeFACT Movement Thesaurus.* [white paper] NEH Office of Digital Humanities.

Literary Geography at Corpus Scale

Wilkens, Matthew

mwilkens@nd.edu
University of Notre Dame, United States of America

Space is important in literary studies. This was true even before postmodernism's spatial turn a generation ago, and our collective interest in spatial issues has only grown in recent years. Of course, what we mean by space varies widely across the discipline. We have studies — some historically oriented, some not — of the relationship between literature and geography at scales ranging from the local to the global. We're also interested in the somewhat

smaller scales of built space and the lived environment. And then there's the long-standing problem of mapping between space and time as organizing principles of narrative and other forms of cultural production (Giles; Hsu; Heise; Orvell and Meikle; Jameson).

We now have methods by which to work with large bodies of text and to extract at least some types of spatial information from them. These methods, which involve computational data mining of hundreds or thousands of books, make it possible for us to address large-scale spatial questions, questions of the type that once seemed unthinkable, in new and robust ways. This is especially true because in many cases we can then combine the evidence produced through these new approaches with our well-established critical judgments.

What follows is an example of such hybrid scholarship. It begins with a question: How can we define and assess the "geographic imagination" of American fiction around the Civil War, and how did the geographic investments of American literature change across that sociopolitical event? To preview quickly the most important results, we find that there is significant national and international dispersion of geographic reference in American novels written between 1851 and 1875; that the distribution of place references tracks closely but not perfectly with population; that changes in literary investment in specific places and regions tends to lag changes in population; and that although there are important shifts in the geographic distribution of literary interest occasioned by the Civil War, such shifts are smaller than established theories would lead us to expect, emphasizing the need to rethink the contours of large-scale cultural change in light of more inclusive textual analysis.

Technical Details

The literary corpus is based on the volumes catalogued by Lyle Wright in his *American Fiction*, 1851-1875. Of the 2,925 titles listed by Wright, 1,050 have been digitized, thoroughly hand-corrected, and contain firmly established dates of publication between 1851 and 1875. The present work is based on these 1,050 volumes, which together contain over 80 million words. The research corpus thus comprises 36% of all known American long-form fiction produced during the generation spanning the Civil War. Of these, 489 volumes (36 million words) were published before 1861; 561 volumes (44 million words) were published in 1861 or later.

Text strings representing named locations in the corpus were identified using the named entity recognizer of the Stanford Core NLP package (Finkel et al.) with supplied training data. To reduce errors and to narrow the results for human review, only those named-location strings that occurred at least five times in the corpus and were used by at least two different authors were accepted. The remaining unique strings were reviewed by hand against their context in each source volume. After corrections were applied, there remained 143,499 occurrences of 1,577 unique location strings in the corpus. The location strings extracted from the corpus texts were then associated with geospatial information via Google's geocoding API. Geocoding results were further reviewed and a small number of errors corrected.

Precision and recall measures were calculated on a small sample of the data. With the above corrections, precision was 0.60, while recall was 0.85; F_1 was 0.70. This is a good result, given the complexity of the problem and the state of the art (Leidner).

Results

What did the literary-geographic imagination of mid-nineteenth-century American fiction look like? It was global, certainly, making use of international locations nearly as often as domestic ones. It was also surprisingly and disproportionately urban; although the twenty largest American cities by population made up only about 10% of the national headcount at the time, the twenty most frequently occurring US cities in the corpus accounted for well over a third of all US place-name occurrences. Literary attention was most heavily concentrated along the eastern seaboard, but not especially so in New England and not to the exclusion of the rest of the nation. While literary attention generally lagged the large and growing populations of the Midwest, it did not by any means ignore that region, nor did it overlook the South (especially — but not only — after 1861) nor the West. On the whole, the use of US place names in the fiction of the period correlated reasonably well with population; large places occurred more frequently than small ones to roughly the same degree as the population of the larger location exceeded that of the smaller. But this relationship, while strong, was interestingly imperfect, yielding numerous cases of under- and overrepresentation relative to population.

There exist mixed signals concerning the emergence of American literary regionalism in the years before 1875. The period's strong investment in urban locations suggests that there was at no point a marked preference for the types of rural locales generally associated with the regionalist impulse, nor was there a large-scale shift away from heavily populated regions in the years immediately following the Civil War, when one might have expected to find the early signs of emerging regionalism. Modest changes toward wider distribution of literary attention, especially at the

city level, did occur following the war, however, and it remains the case that both before and after 1861 there existed widespread literary use of locations outside the northeast corridor. Whether or not these facts point toward an earlier or later emergence of regional writing — or indeed toward any regionalist flowering at all — remains an open question in the absence of a broader historical extension of the current research, but they provide important contextual information concerning the distribution of literary-geographic attention in the generation leading up to what we have long considered the regionalist era.

The American Renaissance as a phenomenon rooted primarily in New England is also only partially supported by the data. While New England locations were overrepresented relative to the population of the region both before and after the Civil War, the extent of their overrepresentation actually increased after 1861, a trend that's difficult to reconcile with standard periodizations derived from Matthiessen, which associate the phenomenon with the first half of the 1850s. At the same time, the fraction of all US location uses that fell within New England was hardly overwhelming at around 15%, a figure that indicates the breadth and depth of literary investment elsewhere in the nation and world at the time.

Finally, the literary-geographic imagination of the period was largely — perhaps surprisingly — stable over time. True, there were small overall shifts toward greater diversity of locations used after the Civil War and away, on a percentage basis, from some of the largest cities, but these and other changes were on the order of single percentage points in most cases. They were potentially important, but they were not overwhelmingly large. This fact doesn't necessarily suggest that significant shifts weren't taking place over the 25 years in question; indeed it's hard to imagine that the Civil War didn't result in meaningful cultural reconfigurations that are traceable through the period's literary output. But it does suggest that at least in the literary-geographic cases studied here, intellectual significance and the absolute magnitude of the observed effect may be best measured on separate scales.

The work presented here represents the first broadly inclusive survey of American literary-geographic usage in the mid-nineteenth-century, one that casts light on — and complicates — our long-standing narratives concerning two of the most important periods of American literary history. Increased focus on the international, urban, and slowly evolving nature of the literary-geographic imagination in the United States around the Civil War is warranted by the current results, which plot a significant path for future work in both conventional and computationally assisted American literary studies.

References

Finkel, J. R., T. Grenager, and C. Manning (2005). Incorporating Non-Local Information into Information Extraction Systems by Gibbs Sampling. *Proceedings of the 43nd Annual Meeting of the Association for Computational Linguistics*. held 2005. 363-370.

Giles, P. (2011). *The Global Remapping of American Literature*. Princeton: Princeton UP.

Heise, T. (2010). *Urban Underworlds: A Geography of Twentieth-Century American Literature and Culture*. New Brunswick, NJ: Rutgers University Press.

Hsu, H. (2010). *Geography and the Production of Space in Nineteenth-Century American Literature*. Cambridge: Cambridge University Press.

Jameson, F. (2003). The End of Temporality. *Critical Inquiry* 29.4. 695-718.

Leidner, J. L. (2006). Toponym Resolution: A First Large-Scale Comparative Evaluation. *Institute for Communicating and Collaborative Systems*: n. pag.

Orvell, M., and J. L. Meikle. (eds), (2009). *Public Space and the Ideology of Place in American Culture*. New York: Rodopi.

Wright, L. H. (1965). *American Fiction, 1851-1875: A Contribution Toward a Bibliography*. 2nd edn. San Marino, CA: Huntington Library.

Scientific Visualization for the Digital Humanities as CLARIN-D Web Applications

Zastrow, Thomas
thomas.zastrow@uni-tuebingen.de
Universität Tübingen, Germany

Hinrichs, Erhard
erhard.hinrichs@uni-tuebingen.de
Universität Tübingen, Germany

Hinrichs, Marie
marie.hinrichs@uni-tuebingen.de
Universität Tübingen, Germany

Beck, Kathrin
kathrin.beck@uni-tuebingen.de
Universität Tübingen, Germany

The importance of scientific visualization for basic and applied research has been recognized as an importance aspect of scientific practise in many disciplines. Recent research trends in the Humanities in general and in Digital Humanities in particular are no exception in this respect (Culy and Lyding, 2010). The goal of the present paper is threefold: (i) to survey different types of scientific visualizations needed for language data, (ii) to describe a set of web applications that have been implemented in the context of the CLARIN–D project[1] , and (iii) to demonstrate the added value of visualization.

CLARIN offers language resources on a large scale, with text corpora often exceeding 100 million words, with spoken and multi–modal data recorded and annotated at different tiers, and with structured language resources of high complexity. In all instances, the querying of such resources will result in new data sets of considerable quantity and complexity. These results are typically rendered in raw data formats that are not conducive to direct inspection by the user. This lack of readability provides a major obstacle for Humanities scholars who are not accustomed to perusing large amounts of data in such a raw, digital form. To overcome this impasse, it is crucial to render data sets in a form that is cognitively more accessible and that highlights the central characteristics of the data in an intuitive fashion.

One area of language–related research where visualization is particularly useful concerns the domain of language variation and language diachrony. The web application CiNaViz (short for *City Name Visualization*) has access to names and geographical coordinates of 1.162.040 geographical locations all over Europe. With its query interface, researchers can search for specific distributions of city names and visualize them on a map. As an example of language variation, Figure 1 shows on the left the distribution of city names ending with *bach* (red), *beck* (blue) and *bek* (green) in Central Europe. One can see that there are clear separators between the three variations, where the separator between bach and beck follows the so called *Benrath Line* which divides the nothern from the southern dialects in Germany. The map on the right hand side of Figure 1 shows locations containing the substring *schwab*. It is evident that these locations are not located in the region called *Schwaben* (Swabia) today (the region around Stuttgart/Tübingen/Ulm, marked on the map by a blue ellipse). This is because the Swabian people were relocated during the Middle Ages from their original places of residence.

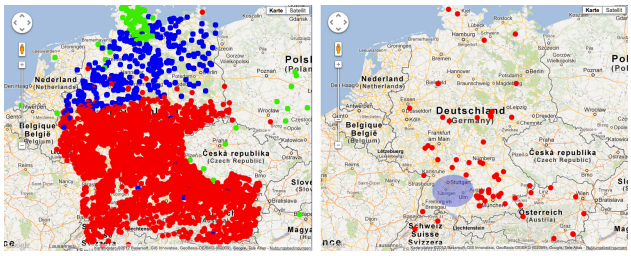


Figure 1:
Visualization of language variation and diachrony

The use of language data is, of course, not only relevant for linguistic and philological research. The social sciences also draw on language data for empirical investigations of various kinds. One area where visualization can help is in tracking the dynamics of culture as reflected in language use. Michel et al. 2010 have coined the term *culturomics* for this new terrain of digital humanities at the interface of humanities and social sciences. One area discussed by Michel et al. concerns the tracking of celebrity names over time by frequency of mention in the Google Books corpus. Such data are, of course, of immediate relevance for historians, sociologists, as well as researchers in media and cultural studies. While Michel et al. based their visualizations on a very large, closed data set, we have applied the same type of techniques to a much smaller and dynamically updated corpus of news articles harvested from the online news feeds of major German newspapers and magazines.

Our *WhoIsInTheNews* [3] web application consists of two parts: (i) a web crawler, which downloads German news feeds everyday and extracts the contained named entities with the help of a chain of WebLicht web services[4] . (ii) A graphical user interface to the stored named entities which allows the user to analyze and visualize the appearances of named entities over time and geographical diffusion. For a morphologically rich language like German, linguistic–preprocessing of the raw data is necessary and is performed by a WebLicht workflow which consists of the following automatic annotation steps: tokenization, part of speech tagging and named entity recognition.

Figure 2 shows the occurrences of the names *Romney* (red) and *Santorum* (blue) in German newsfeeds over a 12 month timeframe, from November 2011 until October 2012. The visualization captures in a concise way the dynamics of the German news coverage of the two leading candidates in the Republican primaries for the 2012 U.S. presidential election. Despite the sometimes unexpected victories by Santorum in several primaries such as Minnesota, Missouri and Colorado, Romney had a consistently higher coverage in the German media, with Santorum dropping out of the German news alltogether shortly after abandoning his campaign on April 10. The WhoIsInTheNews application

shows that visualization techniques can be used not just to plot unrelated career paths of celebrieties as in the case of Michel et al., but also to track the interdependence of such paths in textual materials harvested from online sources in a continuous and incremental fashion.

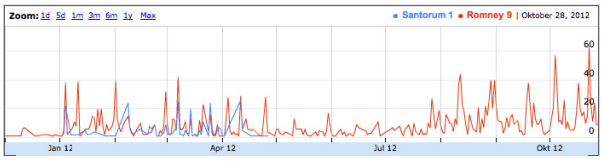


Figure 2:
Tracking of celebrity names over time

The WhoIsInTheNews web application also supports the visualization of named entities that refer to geographical locations. Figure 3 shows the distribution of the 1000 most frequent city names, referenced in the German news feeds harvested over the November 2011 to October 2012 timeframe. Not surprisingly, the density of locations is highest among European cities, followed by the Middle East, the east coast of the United States and the Pacific Rim. Equally noteworthy are the omissions: much of the Midwestern states of the United States, the vast Russian territory outside St. Petersburg and Moscow, as well as much of Africa and South America.

Figure 3:
Tracking the 1000 most frequent city names

Technical realization in CLARIN–D

All visualizations described in this abstract are embedded in the CLARIN–D infrastructure. More specifically, they are implemented as Web 2.0 Ajax driven web applications which make use of annotation web services included in WebLicht[5] , a Servic Oriented Architecture for the orchestration of RESTstyle web services (Hinrichs et al. 2010 and Dima et al. 2012).

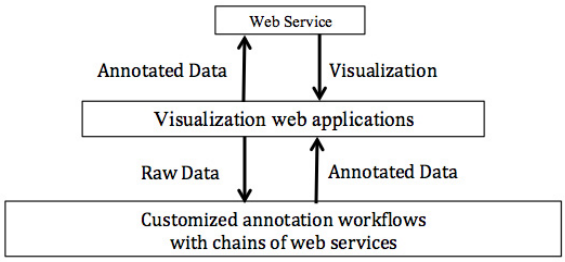


Figure 4:
Embedding web services into web applications

Conclusion and Outlook

Space limitations do not allow us to present the full range of visualization web applications and their significance for Digital Humanities research that are offered in the CLARIN–D infrastructure. For an overview of the visualization tools offered in the CLARIN–D project, we refer interested readers to the WebLicht tools suite (http://weblicht.sfs.uni-tuebingen.de). Therefore, we concentrated on those visualizations that are of relevance to disciplines beyond Linguistics.

The easy web availability of the WebLicht tools is a crucial advantage over existing visualization tools, which typically require expertise in software installation and customization beyond the competence of ordinary digital humanities users. We therefore view web availability as a crucial advantage over existing solutions based on geographical information systems.

References

Culy, C., and V. Lyding (2010). Visualizations for exploratory corpus and text analysis, in *Proceedings of the 2nd International Conference on Corpus Linguistics CILC–10*, May 13–15, 2010, A Coruña, Spain, pp. 257–268.

Dima, E., E. Hinrichs, M. Hinrichs, A. Kislev, T. Trippel, and T. Zastrow (2012). Integration of WebLicht into the CLARIN Infrastructure. *Proceedings of the joint CLARIN–D/DARIAH Workshop* "Service–oriented Architectures (SOAs) for the Humanities: Solutions and Impacts" at Digital Humanities Conference 2012. Hamburg. 17–23.

Hinrichs, E., M. Hinrichs, and T. Zastrow (2010). WebLicht: Web–Based LRT Services for German. In: *Proceedings of the Systems Demonstrations at the 48th Annual Meeting of the Association for Computational Linguistics* (ACL–2010). Uppsala, Schweden. 25–29.

Michel, J. B. et al. and The Google Books Team (2010). Quantitative analysis of culture using millions

of digitized books. *Science* 331, 176 – 82. (doi:10.1126/science. 1199644)

Notes

1. http://www.clarin-d.de
2. CiNaViz is freely available as a web application (see http://weblicht.sfs.uni-tuebingen.de/CityViz/). The Java code of CiNaViz is available under the GPL v. 3 generally used within CLARIN–D
3. WhoIsInTheNews can be accessed at http://weblicht.sfs.uni-tuebingen.de/ne/ . The source code of WhoIsInTheNews is free available under GPL 3. The only IPR restriction concerns the particular data set of news articles, harvested from the news feeds of German newspapers and further processed by the WhoIsInTheNews application. Scholars who want to utilize WhoIsInTheNews for their own data sets and offer this web application on their own server, are free to do so under GPL.
4. https://weblicht.sfs.uni-tuebingen.de
5. The WebLicht acronym stands for *Web based Linguistic Chaining Tool* (https://weblicht.sfs.uni-tuebingen.de)

Combining tailor made research solutions with big infrastructures: The speaking map of the Netherlands

Zeldenrust, Douwe
douwe.zeldenrust@meertens.knaw.nl
Meertens Instituut (Royal Netherlands Academy of Arts and Sciences), Netherlands, The

Van Oostendorp, Marc
m.van.oostendorp@hum.leidenuniv.nl
Leiden University, Netherlands, The

Introduction

Since the middle of last the decade investments in large-scale e-infrastructures for the humanities have risen enormously. Projects such as CLARIN, DARIAH and more recently CLARIAH, received funding. But there is growing scepticism concerning the value of these big infrastructures. In 2012, at the Digital Humanities conference, it was questioned if it is possible and desirable to have an infrastructure for the humanities (Bellamy, 2012). At the Cologne Dialogue on Digital Humanities 2012, this perception was even taken a step further. It was argued that digital infrastructures could be regarded as a dead end for digital humanities. According to this view methodological innovation and advancing the modelling of humanities data and heuristics is better served by flexible small-scale research focused development practices (Zundert, 2012).

While this discussion focuses on the question whether or not e-infrastructures are theoretically readily usable for specific research questions, the current challenge is much more concrete. It lies in catering for specific research needs and making the resources available for future and potential interdisciplinary research. This paper will focus on the possibility of creating tailor made solutions for researchers or Virtual Research Environments (VREs), while at the same time connecting, using and contributing to the big infrastructures. To make this tangible a use case will be presented: ‘The speaking map of the Netherlands’. First of all this paper will give an overview of this project. Next it will go into detail concerning the connections with regard to the access, sharing and storage of the data. Finally, the paper will conclude with a reflection on future steps in connecting research data and tools to infrastructures.

Text for the speaking map of the Netherlands

The Meertens Institute, an institute of the Royal Netherlands Academy of Arts and Sciences, studies the diversity in language and culture in the Netherlands (Meertens). It possesses a large library and numerous (audio) collections. The institute has more than a thousand hours, or six weeks non-stop listening, of audio recordings of dialects from all parts of the Netherlands. The recordings are of conversations between two or more people, without interference from the researcher. The institute started in 1950 to collect the data. In the eighties collecting stopped when the recordings were sufficiently spread out over the Netherlands. Since 2009 the dialects (in total 2216 recordings are available) can be found on the website of the Meertens Institute as the 'speaking map of the Netherlands’ (Soundbites).

The Meertens Institute also has typescripts of 660 of the available recordings. These typescripts have been digitized and the collection contains in total more than 11,000 scans (Archives). Optical character recognition (OCR) has been performed on these typescripts and samples of the produced texts have been corrected. While this collection is digitized, it is not yet readily available for researchers.

In July 2012 the Royal Netherlands Academy of Arts and Sciences funded the project 'text for the speaking map of the Netherlands' to provide open access to the typescripts and to facilitate research with the entire resource (including the audio files). The project started in September 2012 and it will run until May 2013.[1]

A tailor made solution

The project 'Text for the speaking map of the Netherlands' incorporates the lessons learnt from the construction of previous VREs (Berry et al., 2012). One of the key issues of constructing VREs is the implementation of tailor made interfaces for interaction between research questions, data, tools and infrastructure (Zeldenrust, 2011). To establish direct communication and to bridge the gap between research question and technical possibilities a small team has been formed. In conjunction with a phonologist, a programmer and the audio curator of the Meertens Institute, a dedicated interface for phonologic research has been designed.

The phonologic interface will be a web-based system. Its core is a MySQL database containing the metadata and the web locations of the audio files and the scans of the typescripts. The web application will present various ways of exploring the data. First of all, in the cartographic tradition of the Meertens Institute, the site offers the visual interface of the speaking map. It is a representation of the data using the geographic locations. Next, the web application provides access to the datasets using the metadata fields. This will allow scrolling through the data. And finally, using previous added keywords and the ORCs of the typescripts, a text search will also be available. These dedicated interfaces need a relative small budget, are quick to set up and are able to serve as a stepping-stone for innovative research.

Connecting to a digital humanities infrastructure

The interface is specially designed for phonological research of the Dutch dialects. This field of study of research is currently flourishing. One could for instance perform a large-scale phonetic study into vowel quality and vowel length using the combination of sound and typescript, or research into word frequency in Dutch dialects. In addition, the collection offers a unique representation of the Netherlands that no longer exists. The conversations are about poor living conditions in rural areas, welfare during the Depression, the Second World War, local customs et cetera. This collection presents a rich resource for interested parties other than the traditional dialect researchers and is yet to be discovered by for example historians and ethnologist. While the phonological interface makes the resource available for a specific field of research, the project intends to explore the full potential of the resource and to open it up for a wide variety of research possibilities. To reach this goal the resources will be made available through the Common Language and Resources Infrastructure (CLARIN). Each resource will be described using CMDI (Component Metadata Infrastructure) and will be assigned PIDs (Persistent Identifiers). To allow others to harvest our metadata records an OAI-PMH (Open Archives Initiative Protocol for Metadata Harvesting) provider will serve the DCMI (Dublin Core Metadata Initiative) and the CMDI documents. The metadata documents are furthermore indexed and made searchable using an open source search platform.[2] It features full-text search, faceted search and geospatial search. CLARIN will handle the long term archiving of the data using the facilities of The Language Archive (TLA).[3] These big infrastructures are expansive and it takes time and mass to reach a critical usable level. In connecting the dataset to the CLARIN infrastructure it will be disseminated not only for phonology but also for other types of research.

Conclusion

In the introduction it was questioned if big infrastructures are potential platforms for methodological innovation. Some even take it a step further and state that big infrastructures could be regarded as a dead end and that flexible small-scale solutions serve humanities research better. In the case of project 'Text for the speaking map of the Netherlands' audio files, typescripts and metadata will be made available via a tailor made web interface. Big infrastructure CLARIN will provide dissemination, storage and the possibility of combining the resources. The latter functionality is methodological not innovative, however, it may lead to new insights and knowledge. Using the standards of CLARIN also provides easy to use building blocks for future VREs. The conclusion is that at the moment both tailor made research solutions and big infrastructures are of value for research. This current opportunity is restricted to the use of resources; how we deal with tools in this respect is still a matter to be resolved.

References

Bellamy, C. (2012). Opportunity and accountability in the 'eResearch push'. In: *Digital Humanities 2012, conference abstracts* 111–112.

Berry, D. M. (ed). (2012). *Understanding Digital Humanities* London: Palgrave Macmillan.

Zeldenrust, D., and M. Kemps-Snijders (2011). Establishing connections: Making resources available through the CLARIN infrastructure. In: *Supporting Digital Humanities 2011, Copenhagen, proceedings*. http://cst.ku.dk/sdh2011/ papers (accessed 5 October 2012).

Zundert, J. (2012). If you build it, will we come? Large scale digital infrastructures as a dead end for digital humanities. In: *The Cologne Dialogue on Digital Humanities 2012, Controversies around the Digital Humanities, proceedings*. www.cceh.uni-koeln.de/events/CologneDialogue (accessed October 05, 2012).

Archives

The Archives of the Meertens Institute, collection no. 62.

Websites

www.clarin.eu (Accessed October 05, 2012).

www.meertens.knaw.nl (Accessed October 05, 2012).

www.meertens.knaw.nl/soundbites (Accessed October 05, 2012).

www.mpi.nl/research/research-projects/the-language-archive (Accessed October 05, 2012).

Notes

1. In December 2012 the Speaking Map received additional external funding. More than 240 hours of Dutch spoken in France and the USA will be added in 2013. Plans to add other collections are in the making.
2. In this case SOLR is used.
3. The data of the Speaking Map is used by the TLA in a pilot project concerning the archiving of data. This pilot started in November 2012.

Posters

Great Parchment Book project

Avery, Nicola
Nicola.Avery@cityoflondon.gov.uk
London Metropolitan Archives, City of London Corporation, UK

Campagnolo, Alberto
a.campagnolo1@arts.ac.uk
Ligatus Research Centre, University of the Arts, London, UK

De Stefani, Caroline
Caroline.DeStefani@cityoflondon.gov.uk
London Metropolitan Archives, City of London Corporation, UK

Pal, Kazim
K.Pal@cs.ucl.ac.uk
Department of Computer Science, University College London, UK

Payne, Matthew
Matthew.Payne@westminster-abbey.org
London Metropolitan Archives, City of London Corporation, UK

Smith, Philippa
Philippa.Smith@cityoflondon.gov.uk
London Metropolitan Archives, City of London Corporation, UK

Smither, Rachael
rachaelsmither@gmail.com
London Metropolitan Archives, City of London Corporation, UK

Stewart, Ann Marie
ann.stewart@liverpoolmuseums.org.uk
London Metropolitan Archives, City of London Corporation, UK

Stewart, Emma
Emma.Stewart@cityoflondon.gov.uk
London Metropolitan Archives, City of London Corporation, UK

Stewart, Patricia
Patricia.Stewart@cityoflondon.gov.uk
London Metropolitan Archives, City of London Corporation, UK

Terras, Melissa
m.terras@ucl.ac.uk
Centre for Digital Humanities and Department of Information Studies, University College London, UK

Ward, Laurence
Laurence.Ward@cityoflondon.gov.uk
London Metropolitan Archives, City of London Corporation, UK

Weyrich, Tim
t.weyrich@ucl.ac.uk
Department of Computer Science and Centre for Digital Humanities, University College London, UK

Yamada, Elizabeth
lizandchihiro@hotmail.com
London Metropolitan Archives, City of London Corporation, UK

Project outline

The Great Parchment Book of the Honourable The Irish Society is a major survey, compiled in 1639 by a Commission instituted by Charles I, of all the estates in Derry, Northern Ireland, managed by the City of London through the Irish Society and the London livery companies. Damaged in a fire at London's Guildhall in 1786, it has been unavailable to researchers for over 200 years (Moody 1939; Curl 2000). The damaged manuscript has however remained part of the City of London's collections held at London Metropolitan Archives (LMA). As part of the commemoration of the 400th anniversary of the building of Derry's city walls in 1613, it was decided to attempt to make the document available as a central point of the planned exhibition. The book represents an important source for the City's role in the colonisation and administration of Ulster and, given the relative paucity of archival records for early modern Ireland, the manuscript should also reveal key data about landholding and population in 17th-century Ulster.

This ambitious project has attracted support from several funders, including the UK's National Manuscripts Conservation Trust, the Marc Fitch Fund, the Engineering and Physical Sciences Research Council (EPSRC), a number of London livery companies and the Irish Society itself. University College London (UCL), Derry Heritage and Museums Service (DHMS), and LMA have also provided funds and staff time.

Physical description and conservation issues

The manuscript consists of 165 separate parchment membranes, all damaged in the fire. Uneven shrinkage and distortion has rendered much of the text illegible.

Traditional conservation alone would not produce sufficient results to make the manuscript accessible or suitable for exhibition, the parchment being too shrivelled to be returned to a readable state. Much of the text is visible but distorted; following discussions with conservation and imaging experts, it was decided to flatten the parchment sheets as far as possible, and to use multi-modal digital imaging to gain legibility and enable digital access.

The project

A partnership with the Department of Computer Science and the Centre for Digital Humanities at UCL established a four year EngD in the Virtual Environments, Imaging and Visualisation programme in September 2010 (jointly funded by the EPSRC and LMA) with the intention of developing software that will enable the manipulation (including virtual stretching and alignment) of digital images of the book rather than the object itself. The aim is to make the distorted text legible, and ideally to reconstitute the manuscript digitally.

Conservation work on the membranes encompassed cleaning, humidification, and tension drying, using magnets placed on top of the parchment above a metal sheet to hold creases open during the drying process. This opened out areas of parchment where the camera could not reach the text (De Stefani 2012).

The practical conservation of the membranes was the essential first step, followed by the imaging work being carried out by UCL, where a set of typically 50-60 22MP images is captured for each page and used to generate a 3D model containing 100-170MP, which allows viewing at archival resolution. These models can be flattened and browsed virtually, allowing the contents of the book to be accessed more easily and without further handling the document. A readable and exploitable version of the text is also being prepared, comprising a searchable transcription and glossary of the manuscript. This element of the project has received a grant from the Marc Fitch Fund towards the employment of a palaeographer who is also encoding appropriate terms using TEI to capture structural and semantic information about the texts enabling comprehensive searching of the document.

The transcript and images of the document will be published online. We are currently working with web-designers Headscape to develop a website to enable sophisticated online presentation and searching of the document contents.

From 2013, both DHMS and LMA plan to use the document in their interpretation and outreach programmes, developing resources for schools and colleges based on the information it contains. There is also considerable interest from academics, including the University of Ulster. Our work on the computational approach to model, stretch, and read the damaged parchment will be applicable to similarly damaged material as we believe we are developing best practice computational approaches to digitising highly distorted, fire-damaged, historical documents.

Summary

The digital imaging and transcription will provide a lasting resource for historians researching the Plantation of Ulster in local, national and international contexts. The progress of the project is being recorded on a blog (LMA 2012).

References

Curl, J. S. (2000). The Honourable the Irish Society and the Plantation of Ulster, 1608-2000: the City of London and the colonisation of County Londonderry in the Province of Ulster in Ireland: a history and critique. Chichester, West Sussex: Phillimore.

De Stefani, C. (2012). *Conservation of the Great Parchment Book. Presentation held at ARA Annual Conference 2012,* Brighton August 29-31.

London Metropolitan Archives (2012). The Great Parchment Book: Conserving, Digitally Reconstructing, Transcribing, and Publishing the Manuscript Known as the Great Parchment Book. http://greatparchmentbook.wordpress.com (accessed 1 November 2012).

Moody, T. W. (1939). *The Londonderry Plantation, 1609-41: the City of London and the Plantation in Ulster.* Belfast: W. Mullan and son.

Data Driven Documentation of Digital Humanities Discourse

Burton, Matt
mcburton@umich.edu
University of Michigan, United States of America

This poster presents a work-in-progress investigating the use of social media in scholarly communication and the role such technologies play in the formation of scholarly communities. The digital humanities have emerged as a focal point for debates about the impact of information technology in the humanities.[1] While the digital humanities has its roots in the computational processing of text,[2] the landscape today is far richer and more complicated than early practitioners of humanities computing could have ever imagined (except perhaps Father Busa whose grand visions have yet to be realized).[3] Today, the digital humanities encompasses transformative methods of inquiry, radically new kinds of research objects, and potentially destabilizing shifts in scholarly publishing. However, beyond a metamorphosis of method, object, and account, digital humanities leverage information communications technology in unique ways to constitute themselves as a *community in-formation.*

Social media, especially blogs, have been eagerly adopted by the digital humanities community.[4] Blogs are pregnant with promise and peril as platforms for serious (and silly) scholarly communication. They are quick for publishing, support multimedia, and enable rapid interaction, yet, the low barrier of entry and lack of peer review puts blog's credibility and quality in doubt. Outside of the digital humanities, blogs are not necessarily seen as modes of serious scholarly communication, instead they are considered a place for gossip.[5] Such totalizing perspective ignores the diverse uses and meanings of blogs for scholars in a variety of disciplines.[6] The value of scholar's blogs and the vibrant communities of discourse around them should not be understated or ignored.

The seriousness of blogs as a mode of scholarly communication is evident in the creation of initiatives such as Digital Humanities Now[7] and the Journal of Digital Humanities.[8] These projects treat blogs as legitimate forms of proto-scholarship and provide a filter function to the community; finding high quality discourse within their curated selection of digital humanities blogs, the Compendium of digital humanities. As a model for scholarly publishing, the Journal of Digital Humanities presents a reversal of the traditional dynamics of scholarly discourse. Technology has flipped the flow of scholarly communication from one of scarcity to surfeit. It is impossible to keep up with the flood of blogs and Tweet, yet, scholars ignore this "cool kids table" at their peril.[9] In the face of such information overload, scholarly communities must change not only their means of knowledge production, but their information seeking behavior and the ways in which membership and identity are constituted as well.

This study presents a data driven analysis of scholarly discourse focusing on the sociotechnical dynamics of blogs and their role constituting the digital humanities as a community-in-formation. There have been a few data driven approaches to understanding the digital humanities, Melissa Terras' beautiful infographic, Quantifying the Digital Humanities, was an important first step towards surveying the community writ-large.[10] Matthew Jockers and Elijah Meeks have both done some initial work combining topic modeling and digital humanities blogs. Jockers analyzed one year's worth of Day of DH blog posts[11] and Meeks produced a model and visualizations of a variety of texts discussing the question "What are the digital humanities?"[12] This study continues these initial works with a broader breadth of data and deeper analysis of the results.

This poster presents initial findings and an innovative mixed methodological approach combining topic modeling,[13] a form of computational text mining, with grounded theory,[14] a method for developing analytical concepts from interpretivist social science. This mixture of methods enables both a "distant reading" of a vast textual corpus while also rigorously reading individual texts to better analyze and articulate the content of scholarly discourse. Leveraging the Compendium of Digital Humanities,[15] a curated list of blogs produced by Digital Humanities Now, I archive, mine, visualize, and interpret these communications paying special attention to the discursive work of community constitution.

The contribution of this poster is twofold. First, it presents a *data driven* landscape of scholarly communication. Using tools and techniques from information visualization, I represent a topic model of discourse on digital humanities blogs using a javascript visualization framework, Data Driven Documents.[16] Second, it presents a rigorous methodological procedure for the analyzing topic models rooted in an interpretivist qualitative analysis framework, grounded theory. Leveraging grounded theory this study informs our

understanding of scholarly communities in-formation with an interpretive, grounded, empirical analysis of a computational model and its concomitant texts.

Notes

1. Gold, Matthew K., ed. (2012). Debates in the Digital Humanities. Univ Of Minnesota Press.
2. Hockey, S. "The History of Humanities Computing." *A Companion to Digital Humanities*. 2004.
3. Busa, R. "Foreword: Perspectives on the Digital Humanities." In A Companion to Digital Humanities. 2004.
4. Cohen, Dan. "Professors, Start Your Blogs." 2006. http://www.dancohen.org/2006/08/21/professors-start-your-blogs/
5. Saper, C. "Blogademia." Reconstruction 6, no. 4 (2006). http://www.citeulike.org/group/1736/article/1108357 .
6. Hank, C. F. "Scholars and their Blogs: Characteristics, Preferences, and Perceptions Impacting Digital Preservation". University of North Carolina, 2011. http://ils.unc.edu/~wildem/ASIST2011/Hank-diss.pdf .
7. http://digitalhumanitiesnow.org
8. http://journalofdigitalhumanities.org/
9. Pannapacker, William. "Pannapacker at MLA: Digital Humanities Triumphant?" 2011. http://chronicle.com/blogs/brainstorm/pannapacker-at-mla-digital-humanities-triumphant/30915
10. Terras, Melissa. "Infographic: Quantifying Digital Humanities." 2012. http://melissaterras.blogspot.com/2012/01/infographic-quanitifying-digital.html
11. Jockers, Matthew. "Who's Your DH Blog Mate: Match-Making the Day of DH Bloggers with Topic Modeling" 2010. http://www.matthewjockers.net/2010/03/19/whos-your-dh-blog-mate-match-making-the-day-of-dh-bloggers-with-topic-modeling/
12. Meeks, Elijah. "Comprehending the Digital Humanities." 2011. https://dhs.stanford.edu/comprehending-the-digital-humanities/
13. Blei, D. M., A. Y. Ng, and M. I. Jordan. "Latent Dirichlet Allocation." The Journal of Machine Learning Research 3 (2003): 993–1022.
14. Glaser, Barney G., and Anselm Strauss. The Discovery of Grounded Theory: Strategies for Qualitative Research. Aldine Transaction, 1967. Charmaz, Kathy. Constructing Grounded Theory: A Practical Guide Through Qualitative Analysis. 1st ed. Sage Publications Ltd, 2006.
15. http://tinyurl.com/kaj2vpl
16. http://d3js.org/

Expanding the Interpretive and Analytical Possibilities for Understanding Slavery and Emancipation in Washington, DC

Cayer, Janel
jcayerdo@gmail.com
University of Nebraska-Lincoln, United States of America

McMullen, Kevin
mcmullen.kevinm@gmail.com
University of Nebraska-Lincoln, United States of America

In 1862 the United States' Congress passed two bills which legislated the liberation of slaves in the District of Columbia, making Washington, DC the first emancipated city in the nation. The inaugural act, passed on April 16, called for the compensated emancipation of all persons "held to service or labor" in the District. Provisions of the April act required slave owners to submit written petitions to receive compensation. The second, supplemental act of July 12 allowed slaves who were freed under the April act to file for certificates of freedom on their own behalf if their former masters had not petitioned for compensation. Offering valuable descriptions about the lives of former slaves, the documents filed pursuant to these acts have been the focus of our recent work to examine race, slavery, and emancipation for the digital thematic research collection, *Civil War Washington* (civilwardc.org ; *CWW*). This poster presentation will summarize the digitization process that brought the petitions to *CWW* (microfilm digitization, transcription and TEI markup, XSLT transformation, and SOLR search/indexing) and, more importantly, it will demonstrate the compelling work yet to be done with the petitions; specifically, the poster will present a case study of what enriched encoding of the petitions — including encoding of monetary values and person roles — contributes to historical scholarship and to the broader understanding of these documents.

In preparing the petitions for publication, the project team was guided by our scholarly understanding of these documents as well as our desire to make them broadly

available to students, scholars, and the general public as soon as possible. Our TEI markup distinguishes between handwritten and printed content, identifies personal and place names, records values for dates, notes instances where petition forms have been left blank, and acknowledges illegible words and characters in the text. In developing our encoding practices, we took a pragmatic approach, balancing the urge to encode all complexities that make the petitions such significant documents, with the utility of presenting complete transcriptions of heretofore largely inaccessible texts. However, a longer term project goal is the addition of more detailed encoding. Therefore, working with a subset of files, we have expanded the encoding as described below. Our poster explains and illustrates this expanded encoding and showcases what such additional encoding enables for both the visual presentation and the computer-assisted analysis of the documents.

The April act required petitioners to declare the monetary value of their claim and provide a justification for this valuation. A special board of commissioners was established to determine how much remuneration slaveholders would receive. The board's decisions are documented in their *Final Report on Compensated Emancipation in the District of Columbia* (1864). Encoding these monetary values allows us to compare these amounts and enables readers to see the results of the petition in this financial sense: what did a petitioner claim and what did he receive? Similarly, by performing calculations with the encoded figures, we can provide users of the site with the monetary equivalent of each award in today's dollars. These calculations are performed in the XSLT stylesheet used for transforming the TEI files into HTML. Additionally, this encoding, when taken alongside further encoding of details about each slave, enables drawing correlations between the value assigned to each slave and other features. To be sure, we recognize the complex set of problems raised by highlighting the monetary value of people claimed as property. Our intention is to bring this sensitive issue to the forefront in order to encourage discussion and analysis of the complicated calculus at work in the texts and in American history. We mean for our treatment of the financial aspects of these materials to be thought-provoking, not cold and quantitative.

We have also expanded the encoding of the participant description in the TEI header to include a regularized version of the name of every person who appears in a petition. Our long-term goal is to connect every instance of person names to records in the project database. The expansion of the participant description is an intermediate step toward standardizing person names across and within petitions and making people/social connections across documents. The enriched encoding we have implemented here enhances the interpretive and analytical possibilities for understanding slavery and emancipation in Washington. We encourage others to download our TEI files and add encoding that is relevant to their own research interests. Our poster presentation illustrates some of the possibilities that micro-level encoding offers for the study of complex historical documents.

Exploring social tags in a digitized humanities online collection

Choi, Youngok
choiy@cua.edu
Catholic University of America, United States of America

Syn, Sue Yeon
syn@cua.edu
Catholic University of America, United States of America

Proposal:

The recent development of Web 2.0 technologies has been implemented in many applications with an emphasis on user contribution, active user participation, and harnessing of collective intelligence. In particular, interest has grown in the use of tagging as tagging allows users to add their own keywords or tags to online documents and images so that they can organize resources for themselves, share them with others, and find resources that others have tagged.

With this trend, libraries, archives, and museums are providing digitized collections of primary resources to support learning, research, and scholarly activities along with social tagging software to collect user's terms. Digital archival collections and tagging are backbone for the infrastructure of digital humanities as digital archival collections have become vital for scholarly research and teaching in humanities for resources access (Sinn 2012) and as tagging provides users a tool to hold a personal interpretation of resources to make meanings in a personal activity (Golder and Huberman 2006; Trant and Wyman 2006).

While an implementation of tagging has been increased in many digital archival collections, most studies investigated tagging in social networking systems such as Flickr, where broad communities contribute to content for different motivation often based on personal intentions. Few studies have paid attention to the tagging on primary

resources from digitized collections for research and educational activities. Research thus is needed to develop a basic understanding of how users tag primary resources, and how tagging may function in digital humanities.

The purpose of this study is to find the value of tags in a digital collection for research and educational activities by investigating the way users describe digital resources in tags. Specific research questions are: What kinds of terms do users assign to primary historical and classical resources as tags? In what ways can tags be used to supplement for retrieval purpose and resource representation? Do tags in digital scholarly collections go beyond content description? How do social tags differ from textual annotation of the object?

The research is an empirical study collecting data to find descriptive evidence of tagging values and tagging behavior of scholars and users in conducting online research in the 19th-century British and American literature of the NINES (the Networked Infrastructure for Nineteenth-Century Electronic Scholarship). NINES is a federated online collection of peer-reviewed digital objects from more than 110 sites where humanities scholars can find and bring together primary sources, images, literary and cultural documents, and literary criticism in the field of 19th-century studies (Earhart 2010).

Tag analysis will be done based on several categories to identify functional and linguistic aspects of tag usage as well as to find the relationship between tags and annotation of the resources. Tags will be analyzed based on syntax and variations in spelling. Tags will be categorized based on subject-related tags describing the content of the resource and resource-related tags referring to the resource itself. The degree of overlap between tags and annotation will be examined. The findings will help understand users' tagging behavior and resource interpretation in primary and historical resources in humanities.

References

Earhart, A. (2010). Using NINES Collex in the classroom. *ProfHacker, the Chronicle of Higher Education Blog,* http://chronicle.com/blogs/profhacker/using-nines-collex-in-the-classroom/23829 10 May 2010.

Golder, S. A., and B. A. Huberman (2006). Usage Patterns of Collaborative Tagging Systems.*Journal of Information Science,* 32(2): 198-208.

Sinn, D. (2012). Impact of digital archival collections on historical research. *Journal of the American Society for Information Science and Technology,* 63(8): 1521-1527.

Trant, J., and B. Wyman (2006). *Investigating social tagging and folksonomy in art museums with steve.museum, World Wide Web 2006: Tagging Workshop*. Edinburgh, Scotland: ACM.

Modelling the Interpretation of Literary Allusion with Machine Learning Techniques

Coffee, Neil
ncoffee@buffalo.edu
Dept. of Classics, State Univ. of New York at Buffalo

Gawley, James
james.gawley@gmail.com
Dept. of Classics, State Univ. of New York at Buffalo

Forstall, Christopher
forstall@buffalo.edu
Dept. of Classics, State Univ. of New York at Buffalo

Scheirer, Walter
wscheirer@fas.harvard.edu
Harvard University, United States of America

Johnson, David
davidjoh@buffalo.edu
Dept. of Computer Science and Engineering, State Univ. of New York at Buffalo

Corso, Jason
jcorso@buffalo.edu
Dept. of Computer Science and Engineering, State Univ. of New York at Buffalo

Parks, Brian
bparks@synapsesoftware.com
Dept. of Computer Science, Univ. of Colorado, Colorado Springs

A Computational Perspective on Allusion

Most literary allusion, the deliberate evocation by one text of a passage in another, is based upon text reuse. Yet most instances of textual similarity are not meaningful literary allusions. The goal of the Tesserae project (http://tesserae.caset.buffalo.edu) is to automatically detect allusion in a corpus of literary texts, primarily Classical Latin poetry. We begin with a large set of textual parallels, and then attempt to model which of these instances of text reuse are meaningful literary allusions and which are not, according to a group of human readers. While initial attempts with a few basic textual features have proven surprisingly effective, here we employ a more complex feature set and machine learning techniques drawn from the field of computer vision in an attempt to improve the results. Novel applications of machine learning, beyond the well known but constrained textual classification tasks of attribution and categorization, have the potential to be transformative for complex analysis tasks in the Digital Humanities.

Benchmark Data

As an illustration, we consider textual parallels between Book 1 of Lucan's *Bellum Civile* and the entirety of Vergil's *Aeneid* (Coffee et al. 2012). Our benchmark dataset comprises a list of 3,400 pairs of sentences that share at least two different words. Each of these pairs has been read and graded for its literary significance by a group of students and faculty working in small teams. These annotator rankings range from 1 (no literary significance) to 5 (pointed literary allusion).

Learning Relevant Features

Earlier work showed that high-ranked parallels could be distinguished from the others with modest accuracy using only word frequency, distance between words, and the presence of exact form matching versus differently-inflected forms of the same word (Gawley et al. 2012). Nevertheless, others have recommended more sophisticated approaches to this problem (Bamman and Crane 2008). Here we consider an expanded feature set including bi-gram frequency, frequency of individual words, character-level n-grams and edit distances. Our goal is to learn relevant combinations of features in the presence of often incomplete data.

Recent work by members of our team has developed new methods for tuning machine-learning using support vector machines (Scheirer et al. 2012) and random forests (Xiong et al. 2012). Random forest is of particular interest, providing robust feature selection that shows promise for literary analysis (Tabata 2012). The problem of missing data is prevalent in all areas of literary study, but is not well addressed by existing algorithms in common use by digital humanists. This is especially true for ancient texts, where we often find a significant gap in the manuscript tradition. Using principled strategies for imputation and marginalization, we reduce the impact on the results.

Results and Implications

Our ability to learn the difference between high-ranked parallels (ranks 4 & 5) and low-ranked parallels (ranks 1 & 2) for *Bellum Civile* and the *Aeneid* is strong: random forest achieves an average AUC score between 82% and 83%, while linear SVMs yield an average score of 81.5%. This suggests that quantifiable patterns do exist across allusions, which can be captured algorithmically. In this ongoing research we seek a more successful model of literary significance that will allow our software to put interesting allusions at the top of the list; at the same time, we hope it will also cast new light on the underlying structures of our experience of literature.

An interactive demonstration of the Tesserae allusion detection tool accompanies this poster.

Notes

1. **Bamman, D., and G. Crane** (2008). The Logic and Discovery of Textual Allusion. LaTeCH
2. **Coffee, N., J.-P. Koenig, S. Poorima, C. W. Forstall, R. Ossewaarde, and S. Jacobson** (2012). Intertextuality in the digital age. *Transactions of the American Philological Association*, 142(2).
3. **Gawley, J., C. W. Forstall, and N. Coffee** (2012). Evaluating the literary significance of text re-use in Latin poetry. *DHCS*
4. **Scheirer, W. J., A. Rocha, J.Parris, and T. E. Boult** (2012). Learning for meta-recognition. IEEE T-IFS 7.
5. **Tabata, T.** (2012). Approaching Dickens' style through random forests. *DH*.
6. **Xiong, C., D. Johnson, R. Xu, and J. J. Corso** (2012). Random forests for metric learning with implicit pairwise position dependence. *ACM SIGKDD*.

*Work supported by NEH Start-Up Grant Award No. HD-51570-12

"Where do you need us?" — The National Library in the Digital Humanities

Conteh, Aly
aly.conteh@bl.uk
British Library, United Kingdom

Wilms, Lotte
lotte.wilms@kb.nl
Koninklijke Bibliotheek, Netherlands, The

In the past two decades or so, national libraries have been digitising millions of pages of books, newspapers, magazines and other text based collections. In this digital age, the research landscape is changing rapidly, with scholars able to ask new types of questions and answer them in novel ways by working with a wide variety of materials and in new collaborative modes.

As a national library, the British Library (BL) holds over 150 million items dating as far back as 2000 BC and is responding to this climate by realigning its services and structure, including the creation of a new digital scholarship department, but there are still fundamental changes the Library needs to make in order to allow researchers to fully exploit digital resources. We are passionate about working with researchers and scholars so that they can use our digital collection to create new knowledge.

The National Library of the Netherlands (KB) has planned to have digitised and OCRed its entire collection of books, periodicals and newspapers from 1470 onwards by the year 2030. But already in 2013, 10% of this enormous task will be completed, resulting in 73 million digitised pages, either from the KB itself or via public-private partnerships as Google Books and ProQuest. Many are already available via various websites (e.g. kranten.kb.nl , statengeneraaldigitaal.nl , anp.kb.nl , earlydutchbooksonline.nl) and we are working on a single entry point to (re)search all sets simultaneously.

Of course, as an institution that serves the community, these (vast) digitisation projects are not done for ourselves and we make everything we digitise publicly available. Unfortunately, as a library that is not connected to a university — with researchers of their own — a place for us in the Digital Humanities landscape is not as naturally formed as that of a university library. But we do have interesting material and want to get our data out there and have it used by researchers, the general public or anyone who is interested in large corpora of text. But how can we best achieve this?

What does the Digital Humanities community need from us? What do we have to do be as institutions to serve researchers who want to get their hands dirty with this data? How would you like to access this data? Would you prefer a lab with support from material experts and programmers? Would you rather have quantity over quality? Which formats should we offer?

This poster will present the KB as the National Library of the Netherlands, and the collections we(currently)have, but also our efforts to make our data available as complete sets by setting up a Data Services team that focuses on the questions raised by these new activities. This poster will also present the British Library, the data and services it provides and the BL Labs project which is designed to achieve the transformational steps that will change the way the Library provides access to its digital collections and enable scholars to research entire collections rather than just individual items.

We hope to stir up a discussion with the digital humanists at DH2013 to ascertain whether our work is going into the right direction, and what researchers need and expect from us as national libraries. How can *we* help you?

Exploring Digital Humanities Collaborations in the CIC

Courtney, Angela
ancourtn@indiana.edu
Indiana University, United States of America;

Long, Christopher
cpl2@psu.edu
Pennsylvania State University, United States of America

Mueller, Martin
martin.mueller@mac.com
Northwestern University, United States of America

Rehberger, Dean
rehberge@matrix.msu.edu
Michigan State University, United States of America

Walter, Katherine L.
kwalter1@unl.edu
University of Nebraska-Lincoln, United States of America

Winet, Jon
jon-winet@uiowa.edu
University of Iowa, United States of America

The CIC — the Big 10 plus the University of Chicago — is a consortium of some of the largest universities in the Midwest region of the United States. Its members include:

The University of Chicago	http://www.uchicago.edu
University of Illinois Urbana Champaign	http://illinois.edu
Indiana University	http://www.indiana.edu/
University of Iowa	http://www.uiowa.edu
University of Michigan	http://www.umich.edu
Michigan State University	http://www.msu.edu
University of Minnesota	http://www1.umn.edu
University of Nebraska–Lincoln	http://www.unl.edu
Northwestern University	http://www.northwestern.edu
Ohio State University	http://www.osu.edu
Pennsylvania State University	http://www.psu.edu
Purdue University	http://www.purdue.edu
University of Wisconsin–Madison	http://www.wisc.edu

The CIC schools, enroll approximately half a million students each year and have over $7 billion in funded research, over 79 million library volumes, and employ 46,000 faculty. The CIC is home to the HathiTrust Research Center, funded by Indiana University and the University of Illinois at Urbana-Champaign, and boasts three major digital humanities centers: the Illinois Center for Humanities, Arts and Social Sciences (I-CHASS) at the University of Illinois at Urbana-Champaign (http://ichass.illinois.edu/), Matrix: The Center for Humane Arts, Letters, and Social Sciences at Michigan State University (http://matrix.msu.edu), the Center for Digital Research in the Humanities (CDRH) at the University of Nebraska-Lincoln (http://cdrh.unl.edu), and, as of July 1, 2103, Maryland Institute for Technology in the Humanities (MITH) at the University of Maryland (http://mith.umd.edu). CIC institutes are also home to a lively and varied assortment of digital humanities labs, institutes, and initiatives.

Despite the range of digital humanities activities in the CIC, little discussion concerning broad collaboration at a consortial level had occurred among the schools until the 2012 CIC Digital Humanities Summit was held to foster exchange of knowledge and experience across campuses and to discover basic common needs and vibrant shared objectives that could benefit from cross institutional collaboration.

In our poster session members of the CIC Digital Humanities Committee will discuss the impetus for the CIC Digital Humanities Initiative; describe recommended actions from the Summit; report on some responses from other CIC committees (specifically those on data storage and geospatial research) as well as responses of the deans and provosts of the universities to the action items in the report. We will outline preliminary actions of the universities and faculty as a result of the scan and the recommendations. This poster session provides an opportunity to discuss efforts to build digital humanities infrastructure as well as research within the context of a large academic consortium. Other CIC schools are beginning to invest, with widening DH education and research opportunities that will extend far outside the CIC.

In Fall 2011, a new CIC Digital Humanities Initiative was announced with the intent to form a stronger CIC faculty community and networks in digital humanities. Concurrently, an interdisciplinary CIC Digital Humanities Committee was appointed. Initially, this group developed an environmental scan that identifys CIC digital humanities centers and institutes; initiatives and laboratories; conferences and workshops; degrees, specialization & certificate programs; recent grants; publishing initiatives; and digital humanities faculty members. Meant to be a dynamic report, the Scan has been shared widely within the CIC universities, especially among the leadership and the digital humanities faculty. Each school is being encouraged to add to the CIC Digital Humanities Environmental Scan, with the intent that this will be a living, changing document. One school is developing a database version for tracking their own activities and is sharing the code with others.

In April 2012, the CIC Digital Humanities Summit was co-hosted by the University of Nebraska-Lincoln and the CIC. Sixty representatives from twelve of the universities met, including deans, librarians, faculty conducting research in the humanities using technology, and directors of digital humanities centers, studios and labs. Among topics were the environmental scan and possible opportunities and challenges for digital humanities collaboration among the schools; and how to improve the environment and resources for the digital humanities in the CIC. Keynote addresses regarding trends in digital humanities and success and failure of digital humanities centers, and short talks on university publishing and on promotion and tenure characterized much of the Summit's group discussions. It is notable that there was a great deal of enthusiasm expressed for collaboration and for open source approaches, with very few challenges identified by the participants.

Following the summit, the CIC Digital Humanities Committee developed a report with action items based on Summit participants' recommendations. The report is leading to campus-specific efforts as well as consortial activity. The CIC's focus on digital humanities is significant for several reasons. The environmental scan demonstrates that traditionally strong humanities computing efforts in the CIC include an increasingly diverse range of digital humanities endeavors. Moreover, there appear to be a diversity of strengths across the CIC universities with excellent opportunities to provide mutual support and sharing of expertise, perhaps through a broadly conceived initiative such as a CIC Digital Humanities Commons. Faculty are spread across all the humanities disciplines and are in CIC libraries, archives, museums and iSchools, bringing substance to new coalitions and allowing us to reimagine relations with different audiences inside and outside the university. There was a broad acceptance of the importance of collaboration and transdisciplinarity, and a recognition that while it may be difficulty to create community around Digital Humanities at the local level, there is strength at a consortial level.

In the poster, we will address these points and also present brief but concrete examples of mature and robust research programs and nascent efforts among the CIC universities.

The Long Road Home: conversion and transformation of the Text Creation Partnership corpus

Cummings, James
James.Cummings@it.ox.ac.uk
University of Oxford

Rahtz, Sebastian
Sebastian.Rahtz@it.ox.ac.uk
University of Oxford

This poster addresses some of the practical problems of working with the underlying digital files prepared by the Early English Books Online-Text Creation Partnership (EEBO-TCP) using generalized tools developed for used with files following the recommendations of the Text Encoding Initiative (TEI).[1] This conversion work, at the University of Oxford, was initially driven by a desire to experiment with the creation of truly usable ePub editions of the Eighteenth Century Collections Online - Text Creation Partnership (ECCO-TCP) corpus released into the public domain in 2010. It was undertaken at the University of Oxford's IT Services independently of the EEBO-TCP team at the Bodleian Library or in Michigan.

The Text Creation Partnership (TCP) digitization programme is a large and complex operation, with very detailed guidance and standards (http://www.textcreationpartnership.org/docs/) worked out over the last decade. When the project started the decision was made to use SGML markup as the archival storage format, and a variation on the TEI Guidelines, version P3, for the encoding vocabulary. Unfortunately, SGML-aware software is increasingly hard to come by, and the advantages of doing this kind of work in XML these day are self-evident. The TEI has also developed significantly since TEI P3, with the current TEI P5 releases making many changes and improvements (some of them owing to proposals arising from work rationalizing the TCP markup). Interchange or comparison with other TEI texts, or use of TEI-aware software, suggests that we should have a way to transform the TCP texts into valid TEI P5 XML. This does not mean that once texts are in TEI that they are inherently interoperable, at least not without some effort, but this should be a vastly simpler task with a converted version of the EEBO-TCP corpus because they have all been created by a single project following a single set of encoding guidelines. These TCP guidelines have been developed over the course of the project and while not always perfect have gradually been increasing in standards of consistency.

The transformation of these EEBO-TCP texts to a basic and conventional web site alongside the facsimile page images is generally a straightforward task. However, if we want to take advantage of some of the tools now commonly used to process digital files, particularly those based on the current TEI P5 recommendations, this is much more problematic. At very least this involves transforming the SGML markup to XML, and then to the latest edition of the TEI (TEI P5). This poster will document these stages in conversion with examples of some of the problems encountered in this sort of conversion. This has sometimes necessitated changes to the TEI Guidelines themselves in order to be able to consistently encode textual phenomena that has been identified by the TCP project which cannot adequately be described using the current TEI recommendations. In other cases, decisions have needed to be made in the appropriate way to map some of the encoding variants adopted by EEBO-TCP back onto the existing and TEI P5 markup guidelines.

As well as the process, this poster presents some of the software that we have developed for converting ECCO-TCP and EEBO-TCP files. The exercise of transformation

gives an interesting opportunity to examine the nature of the encoding of TCP texts, analyze the range of textual phenomena which are recorded in the corpus, and predict which structures which will be amenable to discovery by future scholars. The approximately 40000-text corpus of TCP also provides a good testbed for the more generalized TEI tools that we have developed. For this poster we describe some of the tools that we've used for the TCP conversions and the results of analysis of the converted TCP texts. As a case study we examine and demonstrate the generation of ebook editions (ePub format) of the ECCO-TCP and EEBO-TCP texts from the converted TEI. The results of such conversions will be discussed with regard for their usefulness for contemporary readers and any failures in representing the intellectual content of the original text.

Notes

1. Significant thanks are owed to Paul Schaffner for his very patient and understanding help in explaining decisions made by the TCP project. We are also grateful to Martin Mueller, Stephen Ramsay and Brian Pytlik Zillig of the Monk and Abbott projects, who wrestled with some of the same dilemmas before and in parallel with us, for discussions of minutiae of the markup. See http://www.tei-c.org/ for more information about the TEI.

MapServer for Swedish Language Technology

Dannélls, Dana
dana.dannells@svenska.gu.se
University of Gothenburg, Sweden

Borin, Lars
lars.borin@svenska.gu.se
University of Gothenburg, Sweden

Olsson, Leif-Jöran
leif-joran.olsson@svenska.gu.se
University of Gothenburg, Sweden

Introduction

Digital maps can today ensure a convenient and efficient rendering of geographical information in real time. Perhaps the best-known source for providing a search interface to the contents of tens of millions of computers on the internet, together with free of charge digital maps for anyone to use, is the Google Map server. The development of digital maps supported by Google is to a large extent driven by the needs of the industry whose requirements range from weather maps to driving instructions obtained from GPS information. Because of this focus, geographical locations which are found in literary texts — e.g. no longer existing places or older name variants – are not guaranteed to be available in this pool of modern digital maps. Moreover, freely available digital maps are naturally not optimized for all kinds of applications. The maps available on the internet are often copyright-protected. This lack of flexibility and the need to point to geographical locations of places that are found in literary texts are two of the main reasons why our group at *Språkbanken* 'the Swedish Language Bank'[1] have decided to investigate an alternative open-source solution, a platform called MapServer (Kropla 2005).[2]

Språkbanken

The Swedish Language Bank is a research unit which focuses on developing linguistic resources and tools for use by researchers and online visitors from different research fields such as linguistics, language technology, and language learning (Borin et al., 2012a; 2012b). It offers access to a vast amount of written natural language text resources including historical and literary texts. Recently, we have recognized the need of combining place-name recognition with geographical information systems as an alternative source of valuable information about the texts and a way to increase text understanding. The role of geographic visualization in the language learning and usage has been explored in various projects (Lieberman, et al. 2010; Gregory and Hardie 2011; Bibiko 2012).

MapServer at Språkbanken

MapServer is an open source Geographic Information System (GIS) development environment for producing maps from geographic data on the Web.[3] Its overall architecture is depicted in figure 1. The user interface provides two different ways to render geographic information: (1) static maps to present a small number of places that appear within the same geographical location and (2) dynamic maps that allow the user to change the amount of data appearing on the map in real time. The geographical dataset consists of both spatial and attribute data and is acquired from Geofabrik.[4] It contains raw data (Open Street Map format) and shape files (Haklay and Weber, 2008).[5]

Figure 1.
The client-server architecture of MapServer.

MapServer accesses an MySQL database to perform coordinate place-name search. This database is automatically maintained by accessing and extracting place-names from the GeoNames database.[6] Because the GeoNames database contains redundant information about place locations, disambiguation is performed. The disambiguation algorithm calculates the Euclidian distance (using the Pythagorean theorem) between places with the same name. If two or more places with the same name occur within three kilometers of each other, only one of these occurrences is added to the database, the one with the most recent modification date. In the case of modification date ties, the place with the highest latitude value is kept.

Visualization of automatically recognized place-names

The named entity recognizer (NER) and annotator used for annotating place names from Swedish literature (Borin et al., 2007; Borin and Kokkinakis 2010) automatically extracts names across large collections of texts with very high precision. In the kind of texts we are dealing with, the distribution of place entities across the whole document is quite common. The reader may sometimes lose track of where the story takes place, in particular if the plot of the story ranges over a wide geographical area. For example, in classical Swedish author Selma Lagerlöf's famous two-part work *The wonderful adventures of Nils* more than 50 place names are mentioned. A reader interested in learning more about these places will benefit from a visualization of their locations.

With the NER tool it is possible extract the place names mentioned in the text and formulate a query with these names to the MapServer application which will render a dynamic map pointing to the geographical locations corresponding to the names. An example is provided in figure 2.

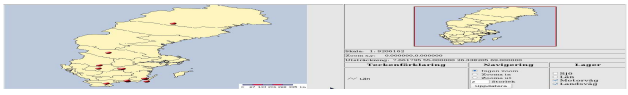


Figure 2.
A dynamic map of the place names mentioned in The wonderful adventures of Nils.

Summary

The MapServer application used by the Swedish Language Bank provides new opportunities for visualizing geographical information found in its large repository of written texts, in particular literary texts. The application is capable of performing coordinate search on the basis of recognized place names and rendering both static and dynamic maps that display their geographical locations.

References

Bibiko, H. J. (2012). Visualization and online presentation of linguistic data. In Frank Seifart, Geoffrey Haig, Nikolaus P. Himmelmann, Dagmar Jung, Anna Margetts, and Paul Trilsbeek (eds), *Potentials of Language Documentation: Methods, Analyses, and Utilization.* 105–110. Honolulu: University of Hawaii Press.

Borin, L., and D. Kokkinakis (2010). Literary onomastics and language technology. In Willie van Peer, Sonia Zyngier, and Vander Viana (eds), *Literary Education and Digital Learning: Methods and Technologies for Humanities Studies.* 53–78. Hershey/New York: Information Science Reference.

Borin, L., D. Kokkinakis, and L. J. Olsson (2007). *Naming the past: Named entity and animacy recognition in 19th century Swedish literature. ACL 2007 Workshop on Language Technology for Cultural Heritage Data (LaTeCH 2007).* 1–8

Borin, L., M. Forsberg, and J. Roxendal (2012a). *Korp — the corpus infrastructure of Språkbanken. Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12),* Istanbul, Turkey, May 2012.

Borin, L., M. Forsberg, L. J. Olsson, and J.Uppström (2012b). *The open lexical infrastructure of Språkbanken. Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12),* Istanbul, Turkey, May 2012.

Gregory, I. N. and A. Hardie (2011). Visual GISting: bringing together corpus linguistics and geographical information systems. *Literary and Linguistic Computing,* **3**:297–314.

Haklay, M. and P. Weber(2008). Open street map: User generated street maps. *IEEE Pervasive Computing,* 7: 12-18.

Kropla, B. (2005). *Beginning MapServer: Open source GIS development (expert's voice in open source).* Berkely, CA: Apress.

Lieberman, M. D., H. Samet, and J. Sankaranarayanan (2010). *Geotagging with local lexicons to build indexes for textually-specified spatial data.*

Proceedings of the 26th International Conference on Data Engineering, California, USA. 201-212.

Notes

1. http://spraakbanken.gu.se
2. There are other open-source GIS alternatives, such as GeoServer and PostGIS, that presumably would serve these needs equally well.
3. http://www.mapserver.org
4. http://www.geofabrik.de/data/download.html
5. http://wiki.openstreetmap.org/wiki/Sv:Map_Features
6. http://download.geonames.org/export/dump/

The Lethbridge Journal Incubator: Aligning digital open access scholarly publishing with the teaching and research missions of a public university.

Donnell, Daniel Paul
daniel.odonnell@uleth.ca
University of Lethbridge, Canada

Hobma, Heather
heather.hobma@uleth.ca
University of Lethbridge, Canada

Ayers, Gillian
gillian.ayers2@uleth.ca
University of Lethbridge, Canada

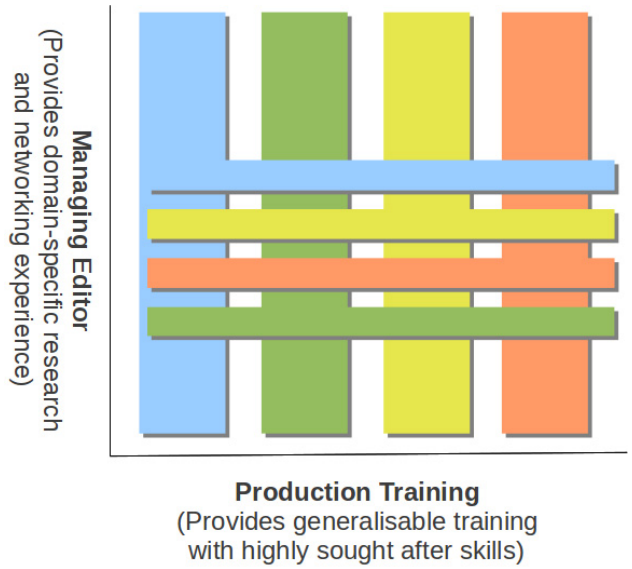
Devine, Kelaine
kelaine.devine@gmail.com
University of Lethbridge, Canada

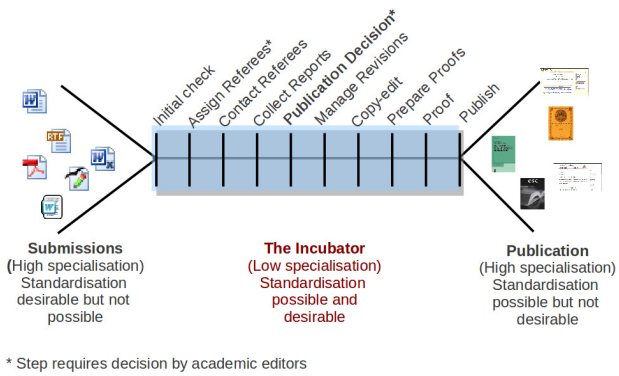
Ruzek, Jessica
jessica.ruzek3@uleth.ca
University of Lethbridge, Canada

The Lethbridge Journal incubator is an experiment in the sustainability of academic publishing. The incubator attempts to ensure this sustainability by aligning the publishing processes with the research, teaching, and service missions of the University. Instead of drawing resources away from these central missions, academic communication under this model become a resource that materially improves the University's ability to carry out these core functions.

The basic premise of the incubator is that the skills and experiences involved in contemporary scholarly journal production are both generalisable across disciplines and of significant value to graduate students whether they pursue post-graduate careers within or without the academy.

Through their work in the incubator, students will acquire training, managerial experience, and networking opportunities that are both of immediate use to them in their research domains and easily transferred to other aspects of their academic or professional careers. These skills are, moreover, highly sought-after by public and private sector employers, especially when combined with the higher-level analytic skills acquired in the course of their graduate studies.





* Step requires decision by academic editors

The incubator works by training graduate students in technical and managerial aspects of journal production. On the one hand, academic journals are highly specialised publications that require high-level, research-domain-specific skills and knowledge from their authors, editors, and readers. On the other hand, however, the actual process by which journals are produced is relatively standard and requires very little research-domain knowledge

Under the supervision of academics, professional librarians, and a professional office manager, students are introduced to the core elements of the workflow that underlies the production of all academic journals and trained in detail both in one or more technical aspects of journal production (copy-editing, preparation of proofs, document-encoding, the use of standard journal-production software), and, more broadly, in the duties of an academic journal managing editor (supervising the progress of articles through the workflow from receipt to publication, corresponding with authors and referees, keeping minutes of editorial meetings, and the like). Students then assume managerial responsibility for one or two titles from their broad area of domain expertise while also working as production assistants specialising in one or more technical aspects of journal production across all titles, regardless of discipline, in the incubator as a whole.

The incubator has been in prototype for just over a year. This poster describes the basic approach and reviews the lessons learned from the first year as well as plans for the coming year, in which we will be working on librarianship and business-model issues.

Live Coding Music: Self-Expression through Innovation

Dussault, Jessica V.
j.dussault.11@ucl.ac.uk
University College London

Gold, Nicolas E.
n.gold@ucl.ac.uk
University College London

Musical live coding is a relatively new discipline that explores the manipulation of audio and music through altering a program in real-time, frequently while projecting the music-generating code in front of an audience (TOPLAP, 2012). Live coding falls under both the realm of computer music and computer science in a blend of programming and artistic output. Pioneering live coders often write their own languages, adapt existing environments, and create unique interfaces and control mechanisms for their systems (McLean and Wiggins, 2010). The frequency of this personalization gives rise to two important questions: in spite of the presence of existing, similar environments, why are live coders creating their own systems and what types of changes are they making to the musical and programming functionality of their environments?

The questions are addressed by a comparative study of a wide range of live coding environments, the literature on live coding, and by interviews with active live coders. The types of implementations the environments use for musical expression and the rationale provided by their creators illustrate the many routes that live coders are taking to explore musical creativity and programming intricacies (for discussion of this creativity, see Collins, 2011; Magnusson, 2011; and McLean and Wiggins, 2010). From community supported languages such as SuperCollider (McCartney, 2002, 2012) and MAX/MSP (Puckette, 2002) to individually crafted, younger languages like Sorensen's Impromptu (2013), Fluxus (Griffiths and Papp, 2012) and Tidal (McLean, 2011), the study shows that all live coding environments must find solutions to the issues of musical representation, time passage, audio creation, and programming paradigm, and it is in these differences and similarities between environments that individual strategies of the live coders become apparent. Figure 1 shows a single note expressed in multiple environments, illustrating the differences between them.

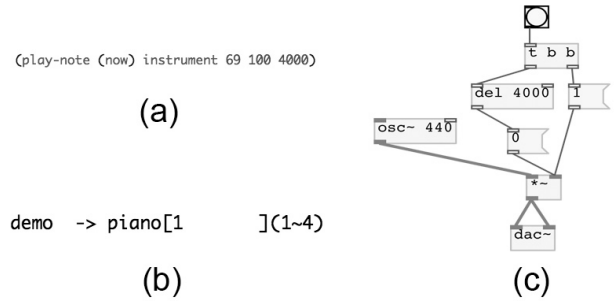


Figure 1:
A sustained note at 440 Hz in three environments used for live coding audio:(a) Impromptu, (b) ixi lang, and (c) Pure Data (see Sorensen, 2013; Bausola et al., 2013; and Puckette 2013, respectively).

The dedication to individuality in live coding has led to a wide range of methods, goals, and musicality that, as of yet, has no formal theory to describe or categorize the products and notations of live coding (McLean and Wiggins, 2010). Therefore, this study categorizes methods used by live coders in a prototypical music theory. The study concludes that live coders create and modify their own environments because musical improvisation and on-the-fly composition demand a certain familiarity with a programming environment that is not easily obtained by adopting an existing system (see also Brown, 2006). Additionally, the study proposes that live coders prefer their own environments to others' because live coders are as interested in the processes behind making the music as they are in the music itself(see also Brown, 2006; Brown and Sorensen, 2009; and Aaron et al., 2011). By adapting their programming environments, live coders explore the boundaries of their processes and the impact that constraints, interfaces, and techniques have upon the musical output.

The presentation includes descriptions and examples of code illustrating the different treatments that live coders use when approaching musical or programming requirements. These include the passage of time, ensemble coordination, and types of code interfaces. A laptop will be provided to offer a hands-on interactive comparison of several live coding languages, and an ongoing, paper-based live coding game will encourage participation throughout the course of the conference (Nilson, 1975).

References

Aaron, S., A. F. Blackwell, and R. Hoadley (2011). A Principled Approach to Developing New Languages for Live Coding. *Proceedings of the International Conference on New Interfaces for Musical Expression (NIME),* 381–386. Oslo, Norway.

Bausola, D., E. Hurtado, and T. Magnusson (2013). "Ixi Audio." http://www.ixi-software.net/ (accessed 14 March 2012).

Brown, A. R. (2006). Code Jamming, *M/C Journal,* 9(6). http://journal.media-culture.org.au/0612/03-brown.php.

Brown, A. R., and A. Sorensen (2009). Interacting with Generative Music Through Live Coding. *Contemporary Music Review.* 28(1): 17–29. doi:10.1080/07494460802663991.

Collins, N. (2011). Live Coding of Consequence. *Leonardo Music Journal.* 44(3): 207–211.

Griffiths, D. and G. Papp (2012). "Fluxus." http://www.pawfal.org/fluxus/ (accessed 30 October 2012).

Magnusson, T. (2011). Confessions of a Live Coder. *Proceedings of the International Computer Music Conference,* 609–616. University of Huddersfield, England. http://quod.lib.umich.edu/i/icmc/bbp2372.2011.122.

McCartney, J. (2002). Rethinking the Computer Music Language: SuperCollider. *Computer Music Journal* 26(4): 61–68.

McCartney, J. (2012). "SuperCollider." http://supercollider.sourceforge.net/ (accessed 30 October 2012).

McLean, C. A. (2011). *Artist-Programmers and Programming Languages for the Arts.* Ph.D. thesis, Goldsmiths, University of London. http://yaxu.org/writing/thesis.pdf.

McLean, A., and G. Wiggins (2010). Live Coding Towards Computational Creativity. *Proceedings of the First International Conference on Computational Creativity (ICCC-X).* Lisbon, Portugal.

Nilson, C. (1975). An Instructional Game for 1 to Many Musicians. Composition. http://toplap.org/index.php?title=Click_Nilson%27s_text_piece.

Puckette, M. (2002). Max at Seventeen. *Computer Music Journal,* 26(4): 31–43.

Puckette, M. (2013). "Pure Data." http://puredata.info/ (accessed 14 March 2013).

Sorensen, A. (2013). "Impromptu." http://impromptu.moso.com.au/ (accessed 14 March 2013).

TOPLAP (2012). "Toplap." http://toplap.org/about/ (accessed 30 October 2012).

Stylometry with R: a suite of tools

Eder, Maciej
maciejeder@gmail.com
Pedagogical University, Krakow, Poland

Kestemont, Mike
mike.kestemont@gmail.com
University of Antwerp, Belgium

Rybicki, Jan
jkrybicki@gmail.com
Jagiellonian University, Krakow, Poland

Stylometry today uses either stand-alone dedicated programs, custom-made by stylometrists, or applies existing software, often one for each stage of the analysis. Stylometry with R can be placed somewhere in-between, as the powerful open-source statistical programming environment provides, on the one hand, the opportunity of building statistical applications from scratch, and, on the other, allows less advanced researchers to use ready-made scripts and libraries. In our own stylometric adventure with R, one of the aims was to build a tool (or a set of tools) that would combine sophisticated state-of-the-art algorithms of classification and/or clustering with a user-friendly interface. In particular, we wanted to implement a number of multidimensional methods that could be used by scholars without programming skills. And more: it soon became evident that once our R scripts are made, provided with a graphic user interface and more or less documented, they are highly usable in class; experience shows that this is an excellent way to work around R's normally steep learning curve without losing anything of the environment's considerable computing power and speed.

The crucial point in building the interface was to keep all the stages of the entire analysis – from loading texts to final results in numeric and graphic form — in a single script. To exemplify, our Stylo script does all the work: it processes electronic texts to create a list of all the words used in all texts studied, with their frequencies in the individual texts; normalizes the frequencies with z-scores (if applicable); selects words from stated frequency ranges for analysis; performs additional procedures that (usually) improve attribution, such as Hoover's (2004a, 2004b) automatic deletion of personal pronouns and culling (automatic removal of words too characteristic for individual texts); compares the results for individual texts; performs a variety of multivariate analyses; presents the similarities/distances obtained in tree diagrams; finally, produces a bootstrap consensus tree — a new graph that combines many tree diagrams for a variety of parameter values. It was our aim to develop a general platform for multi-iteration attribution tests; for instance, an alternate script produced heatmaps to show the degree of Delta's success in attribution at various intervals of the word frequency ranking list (Rybicki and Eder 2011). The last stage of the interface design was to add a GUI, since some humanists might be allergic to the raw command-line mode provided by R — an observation shared by all three authors — and a host of various small improvements, like saving (and loading) the parameters for the most recent analysis, a wide choice of graphic output formats, etc.

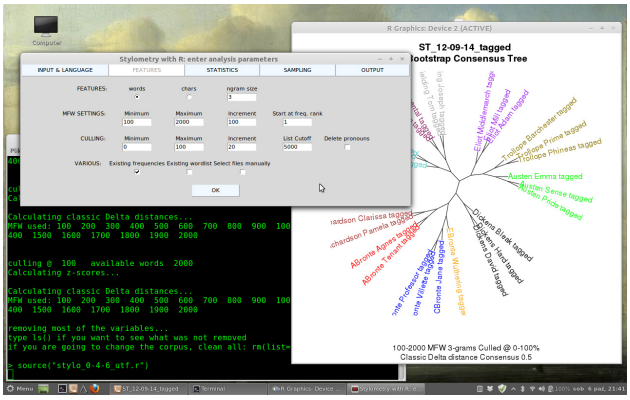


Fig. 1 The Stylo script with a bootstrap consensus plot.

The authors believe that at some point the suite of out-of-the-box scripts will cover a wide range of methods used in stylometry. So far, we offer the following tools:

(1) the **Stylo** script, now in version 0.4.8. This is the main tool, thoroughly tested and (partially) documented. It performs Principal Components Analysis, Cluster Analysis, Multidimensional Scaling, and Bootstrap Consensus Trees. The script reads plain text files, XML, or HTML; it supports explicitly nine languages, and implicitly many more (e.g. preliminary tests with a Chinese corpus were quite promising). Publication-quality plots can be exported in PDF, JPEG, PNG, or EMF formats. Additionally-generated files, such as a wordlist used and a table of word frequencies, can be re-used in other scripts or other statistical tools.
(2) the **Classify** script. It performs Delta (Burrows 2002), k-Nearest Neighbors classification, Support Vectors Machines, Naive Bayes, and Nearest Shrunken Centroids (Jockers, et al. 2008). Most of the options are derived from the above-mentioned Stylo script.
(3) the **Rolling Delta** script. It analyses collaborative works and tries to identify the authorship of their fragments. The first step involves a "windowing" procedure (Van Dalen-Oskam and Van Zundert 2007) in which each reference text is segmented into consecutive, equal-sized samples or windows. After "rolling" through the test text we can plot the resulting series of Deltas for each reference text in a graph.

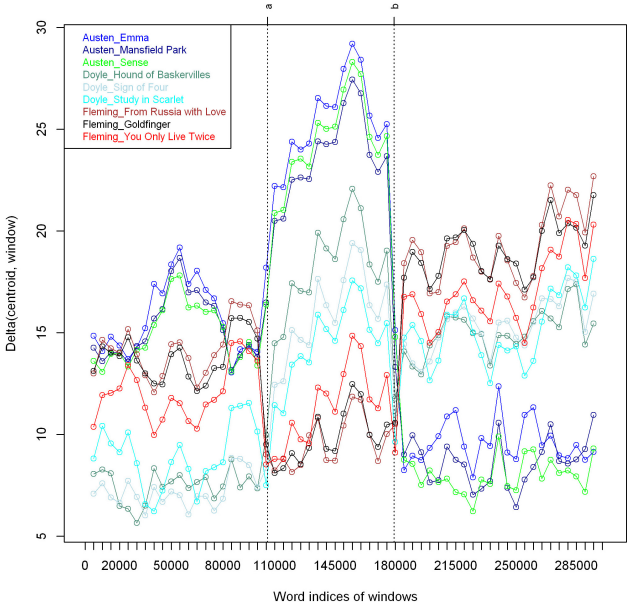


Fig. 2 Sample plot generated by the Rolling Delta script.

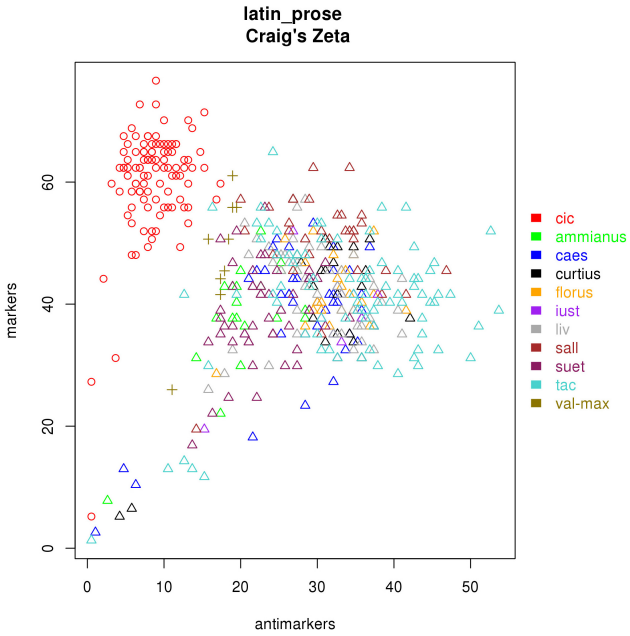


Fig. 3 Sample plot generated by the Oppose Test script.

(4) the **Oppose Test** script. It performs a contrastive analysis between two given sets of texts, using Burrows's Zeta (2006) in its different flavours, including Craig's extensions (Craig and Kinney, 2009). The script generates a list of words significantly preferred by a tested author, and another list containing the words significantly avoided.

(5) the **Keywords** script. This considerably simple tool is an implementation of the concept of "keywords", i.e. words appearing with a statistically significantly higher frequency in one text or collection of texts in comparison to another text or collection.

The scripts are available on https://sites.google.com/site/computationalstylistics/

References

Burrows, J. F. (2002). 'Delta': a measure of stylistic difference and a guide to likely authorship. *Literary and Linguistic Computing*, 17(3): 267–87.

Burrows, J. F. (2006). All the Way Through: Testing for Authorship in Different Frequency Strata. *Literary and Linguistic Computing*, 22(1): 27–48.

Craig, H., and A. F. Kinney (eds.) (2009). *Shakespeare, Computers, and the Mystery of Authorship*. Cambridge: Cambridge University Press.

Van Dalen-Oskam, K., and J. Van Zundert (2007). Delta for Middle Dutch – Author and Copyist Distinction in 'Walewein'. *Literary and Linguistic Computing*, 22(4): 345–62.

Hoover, D. L. (2004a). Testing Burrows's Delta. *Literary and Linguistic Computing*, 19(4): 453–75.

Hoover, D. L. (2004b) Delta Prime? *Literary and Linguistic Computing*, 19(4): 477–95.

Jockers, M. L., D. M. Witten, and C. S. Criddle (2008). Reassessing authorship of the 'Book of Mormon' using delta and nearest shrunken centroid classification. *Literary and Linguistic Computing*, 23(4): 465–91.

Rybicki, J., and M. Eder (2011). Deeper delta across genres and languages: do we really need the most frequent words? *Literary and Linguistic Computing*, 26(3): 315–21.

Introducing GeoBib: An Annotated and Geo-referenced Online Bibliography of Early German and Polish Holocaust and Camp Literature (1933–1949)

Entrup, Bastian
bastian.entrup@zmi.uni-giessen.de
Universität Gießen, Germany

Bärenfänger, Maja
Maja.Baerenfaenger@germanistik.uni-giessen.de
Universität Gießen, Germany

Binder, Frank
Frank.Binder@zmi.uni-giessen.de
Universität Gießen, Germany

Lobin, Henning
Henning.Lobin@germanistik.uni-giessen.de
Universität Gießen, Germany

1 Introduction and Goals

The Holocaust's aftermath on memory discourses still represents an important research field. Almost 70 years after the end of the Second World War, there will soon be no direct witnesses of the Nazi crimes anymore. What will remain are texts bearing witness to the crimes. These texts are to be collected, preserved, and made accessible in a systematic manner. Especially the early texts are of great interest: texts written or published between 1933 and 1949, particularly in the years 1944/1945. In this context Germany, bearing the guilt for the Holocaust, and Poland, as a Nazi-occupied country where large parts of the Holocaust took place, are of special relevance.

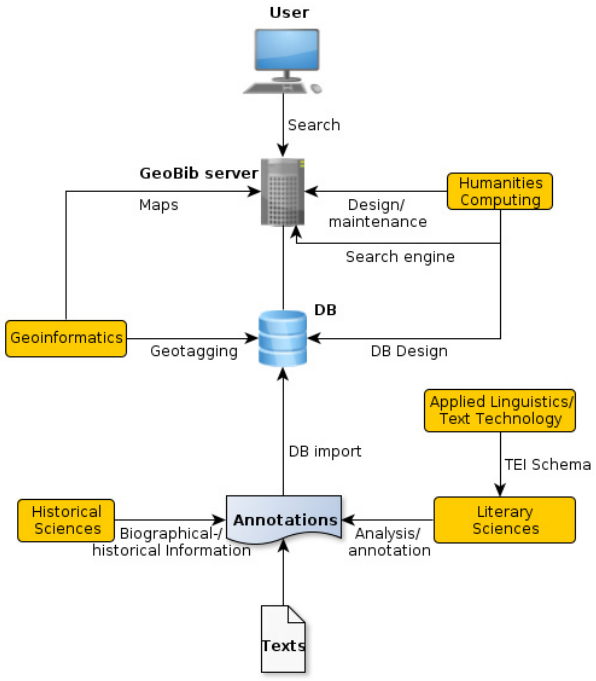


Figure 1:
Outline of the GeoBib project workflow

Unfortunately, the early texts on the Holocaust were soon forgotten or suppressed. With the formation of the two German states (1949), a process of suppression of the texts of the victims began (Cf. Hickethier 1986, p. 578). The official artistic doctrine in the GDR and Soviet Poland excluded certain authors, themes, and forms of presentation.

The goal of the interdisciplinary GeoBib project is to build a systematic and geo-referenced online bibliography of the early German and Polish Holocaust and camp literature: the first complete, bilingual, i.e. with respect to texts written both in German and Polish, research platform for information on the texts. On these, GeoBib seeks to collect annotations and metadata. This includes short summaries, keywords, biographical information on the authors, reviews, scientific literature, information on persons, geographical data, as well as time periods mentioned in the texts. The goal is to make information on these widely unknown texts accessible and searchable for a broad and interdisciplinary audience of researchers. Due to legal restrictions, the project does not, however, aim at annotating the whole texts. The focus lies on the collection and integration of these various kinds of information.

The combination of literary and geo-temporal annotations, as well as the implementation of different (geographical) search mechanisms, enables scholars to conduct innovative research in different scientific domains like literature or history, but is also intended to suit the needs of students, teachers, and the interested public. Maps and other visualizations of geo-temporal information are expected to reveal “historical relations that might otherwise go unasked” (White 2010, p.6).

2 Workflow and Technical Implementation

Since various research areas are involved, we need to aggregate a multitude of different information from the fields of literature, history, and geography — a challenge in data management and text technology. An adaptation of the TEI standard, using an ODD file, is being used to build a schema that serves to integrate the intended collection of the different kinds of metadata mentioned above.

Starting with the texts, annotation data is collected in TEI files that will be used as input for different processing steps (see Fig. 1): Bibliographical, biographical, and other metadata on the text need to be saved in a database and will be linked to the annotated geo-temporal data. Places need to be mapped to coordinates and annotated with the corresponding timestamp. The geographical data will be used in a *geographic information system* (GIS) and in an

online database to support search queries across all relevant annotation data (see Fig. 2). The combination of time and space, extracted from the texts, will be used to display the data in form of geographical maps.

Exemplary queries:

1. Find texts written from 1939 through 1945 in Polish language by or about children in the Warsaw Ghetto.
2. Find novels or dramas regarding the Kristallnacht.
3. Find texts that were published in the American Occupation Zone in 1948.
4. Find fictional texts by women authors written in Polish language.

Figure 2:
Exemplary queries that the GeoBib system is expected to support by returning a list of relevant bibliographic references.

3 Context and Outlook

The Holocaust was a traumatic event that has recently been examined with increasing focus on space and time. Or as *Beorn et. al.* (2009, p. 563) put it: the "Holocaust was a profoundly geographical event, rooted in specific physical spaces, times, and landscapes". Their project 'Geographies of the Holocaust' is actively working on analyzing the Holocaust from a geo-spatial point of view.

Others are looking at how geography is represented in texts (cf. Eide 2012, Appadurai 2010). Specialized software projects were developed to make the representation of, or the work with, geographical information in the humanities easier, e.g. Neatline (Nowviskie et. al. 2012).

Geo-spatial information extracted from literary texts, e.g. Google Ancient Places (GAP), can play a "vital role in improving efficiency for researchers" (Isaksen 2011, p. 82).

Hence, the GeoBib project is active in a growing field of interdisciplinary geo-spatial research as well as the European Holocaust research community (cf. Kahn 2011).

The GeoBib team is looking forward to collaborate in this context.

References

Appadurai, A. (2010). How Histories Make Geographies: Circulation and Context in a Global Perspective, Transcultural Studies. 1: 4–13.

Beorn, W., T. I. M. Cole, S. Gigliotti, et al. (2009). Geographies of the Holocaust, *Geographical Review.* 99(4): 563–574.

Eide, Ø. (2012). *Underspecified, Ambiguous or Formal. Problems in Creating Maps Based on Texts*. http://www.dh2012.uni-hamburg.de/conference/programme/abstracts/underspecified-ambiguous-or-formal-problems-in-creating-maps-based-on-texts/ (acessed 10 April 2012).

Hickethier, K. (1986). Biographie, autobiographie, Memoirenliteratur. In Fischer, L. (ed.), *Literatur in der Bundesrepublik bis 1967.* München 574–584.

Isaksen, L., E. Barker, E. Kansa, et. al. (2011). *GAP: a NeoGeo approach to classical resources, Leonardo*, 45(1): 82–83.

Kahn, R. (2011). The EHRI Project: building an online archive for European Holocaust research. SCONUL Focus. (52). 21–22.

Nowviskie, B., W. Graham, D. McClure, et al. (2012). Geo-Temporal Interpretation of Archival Collections Using Neatline. Digital Humanities 2012. *Conference Abstracts*. 299–302.

White, R. (2010). What is Spatial History? http://www.stanford.edu/group/spatialhistory/cgi-bin/site/pub.php?id=29 . (acessed 10 April 2012).

Notes

1. Funded by the German *Federal Ministry of Education and Research / Bundesministerium für Bildung und Forschung (BMBF)* for three years, starting July 2012 (FKZ: 01UG1238A).
2. Participating institutions: the Arbeitsstelle Holocaustliteratur, the Center for Media and Interactivity (ZMI), Applied and Computational Linguistics, Geoinformatics and Remote Sensing (all Justus-Liebig-Universität Gießen), and the Herder-Institut (Philipps-Universität Marburg).

Mapping Multispecies Temporalities: Experiments in Diagrammatic Representation

Gan, Elaine
eganuc@gmail.com
University of California, Santa Cruz, United States of America

Introduction

This paper presents an ongoing experiment in mapping temporalities of multispecies ecologies. Addressing the conference theme, "Freedom to Explore", the paper asks:

what happens to figurations of "freedom" and "exploration" when recontextualized as multispecies emergences (more-than-human freedoms), indeterminate patternings (temporal rather than spatial formations), and contingent, historically constituted mobilities (uneven trajectories and explorations constituted through difference and incommensurability)? Rather than proposing answers, my presentation offers two ways of thinking through such recontextualization. It unpacks a speculative methodology for digital and diagrammatic representations of time, and offers an early look at a working prototype for a fungal clock: a web-based visualization of polyrhythmic interactions between multiple organisms.

Diagrammatic Representation, Or Mapping Worlds Otherwise

Political ecologists describe maps as power tools: they both reveal *and* hide. As Arturo Escobar writes: "Who counts, draws, and narrates and how is of decisive importance." (Escobar 2008; Rocheleau 2005). Indeed, fast growing fields of data visualization, geographic information systems, simulation modeling, and scenario planning tap into silos of data to generate seemingly inexhaustible means to chart, count, tabulate, track, tag, and thus masterfully deploy a financialized Fix. Such representations are tethered to nineteenth century projects of calculation and seriality (Daston 1994; Hopwood, Schaffer, Secord 2010), epistemic tools constructed through historical materialities and discursive formations, but fraught nonetheless with violent legacies of occupation, extraction, and obsolescence. Given their rhetorical clout in the production and disruption of globalized hegemonies, how might visualizations — and specifically new diagrammatic forms and alternative cartographies — render worlds otherwise?

To describe freedom and exploration is to consider circulation. To study circulation is to unpack relationships that unfold not just in space, but through time. How might diagrammatic forms open up little-known, or perhaps long-forgotten, multiplicities of time or polyrhythms?

Unilinear orderings of time undergird scientific visualizations of growth, transformation, and exchange. Representations of species origins and variabilities — through historical timelines, energy cycles, network patterns, activity diagrams, genealogies, arborescent or rhizomatic animations — largely depend on a chronological sequence of standardized, homogeneous units of time. Entangled histories, indeterminate contingencies, and dynamic mobilities are distilled into teleological causes and effects. Qualitative differences and durational synchronies/sedimentations (Bergson 1922) that emerge from and constitute change, movement, and encounter are quantified into time series, or a measurable succession of seconds, minutes, hours, days, months, years, decades, centuries. Modernity configures and depends upon a temporality abstracted from lived experiences and meterized into a compressed, scalable, and human-centric coordinate of space (Harvey 1990). Against unilinear anthropogenic time, difference is obscured and becomes difficult to locate. As capitalist breakdowns come to a head, the need for critical attention to differential temporalities can no longer be ignored.

Multispecies Times: A Fungal Clock

The second thread of this paper offers a speculative methodology for diagrammatic representation. It presents an experimental web project that reframes unilinear time as entangled polytemporalities constituted through more-than-human assemblages. An interdisciplinary collaboration between anthropologist Anna Tsing and media artist Elaine Gan, it takes the form of a fungal clock. Written in html-5, and built through the lens of field research and multispecies ethnographies of matsutake (mycorrhizal fungi) worlds, the clock brings together a series of digital experiments in representing time. Instead of visualizing cycles and encounters between determinate species beings against a single (Western European) time scale, the clock diagrams interweaving temporal relations or topologies of differential rhythms. These relations are distinguished as three main folds: synchrony or coordination across recurring seasons; emergent becomings or interspecies webs that sediment into ecologies and worlds; uneven trajectories and aleatory encounters that crystallize into historical conjunctures.

Diversity does not unfold against a standard ruler of time. Nor does "it" branch through anthropogenic scales. Learning from Karen Barad, "temporality is produced through the iterative enfolding of phenomena marking the sedimenting historiality of differential patterns of mattering." (Barad 2007) What is at stake in visualizing temporality is not *how* formations change, but *which* properties, taxonomies, and relations become meaningful, intense, and valuable within and across particular regimes, niches or semiotic systems. What comes to matter is a matter of time.

Thus, this paper considers critical diagrams of temporalities as expanded ways of defining "freedom" and "exploration". It calls for ongoing development of conceptual-practical apparatuses that rethink human-centric movements and agencies as multispecies webs and temporal patternings. It highlights these questions as generative lines of inquiry for an exciting field of possibilities, increasingly known as "digital humanities": (1) What kinds of temporal relations materialize into and emerge through biocultural

assemblages? (2) How might interdisciplinary scholarship (theorizing, making, imagining) articulate these movements to work through incommensurable claims for engineering life or death? (3) What digital diagrammatic media might mobilize ethico-political practices beyond modernist master narratives and postmodernist relativisms — and towards worlds constituted through immanence and difference?

References

Barad, K. (2007). *Meeting the Universe Halfway*, Durham and London: Duke University Press. 180.

Bergson, H. (1922). *Duration and Simultaneity*, translation by Leon Jacobson in Durie, R. (ed.) *Duration and Simultaneity: Bergson and the Einsteinian Universe*, Manchester: Clinamen Press, Ltd., 1999.

Braidotti, R. (2002). *Metamorphoses: Towards a Materialist Theory of Becoming*, Cambridge: Polity Press.

Daston, L. (1994). Enlightenment Calculations, *Critical Inquiry* 21(1:) 182-202. Autumn 1994.

De Landa, M. (2000). Deleuze, Diagrams, and the Genesis of Form, *American Studies*, 45(1): 33-41.

Deleuze, G., and F. Guattari (1987). *A Thousand Plateaus*. translation by Brian Massumi, Minneapolis: University of Minnesota Press.

Escobar, A. (2008). *Territories of Difference*, Durham and London: Duke University Press. 56.

Galison, P. (2003). *Einstein's Clocks, Poincaré's Maps: Empires of Time*, New York and London: W.W. Norton & Company.

Harvey, D. (1990). *The Condition of Postmodernity: An Enquiry Into the Origins of Cultural Change*, Cambridge, MA and Oxford, UK: Blackwell Publishing.

Hopwood, N., S. Schaffer, and J. Secord (2010). "Seriality and Scientific Objects in the Nineteenth Century", *History of Science* 48, 161. 251-285.

Lima, M. (2011). *Visual Complexity: Mapping Patterns of Information*. New York: Princeton Architectural Press.

Rocheleau, D. (2005). Maps as Power Tools: Locating Communities in Space or Situating People and Ecologies in Place? in Brosius, P., A. Tsing, and C. Zerner (eds.), *Communities and Conservation*, Walnut Creek, CA: AltaMira Press. 326-362.

Rosenberg, D., and A. Grafton (2010). *Cartographies of Time*. New York: Princeton Architectural Press.

Digital Humanities Keywords: A Collaborative Community Web-based Project

Garfinkel, Susan
sgarfinkel@loc.gov
Library of Congress, United States of America

As the field of digital humanities has expanded dramatically in recent years, it has also struggled to define —or perhaps more correctly, to redefine—itself across a seemingly divergent set of its practitioners' backgrounds, interests, priorities, methodologies, and institutional settings that yet still have something fundamental in common. From critical code studies to corpus linguistics to tool creation to online pedagogy, digital humanities practitioners surely share the *digital* and the *humanities* yet may work within different fundamental paradigms, asking different questions and using different methodologies in the service of achieving very different final goals (Gold, Kirschenbaum). How, beneath the shared surface of this broadly cast "big tent" of digital humanities, can such a diverse community of practice come together in fruitful and mutually beneficial ways? What can we as scholars do to facilitate this convergence in a broad discursive space where the ethic of "more hack, less yack" can sometimes discourage the meaningful exchanges of ideas that still need to happen, the collective negotiations of purpose that so often serve as the source of new insights?

In his seminal book *Keywords: A Vocabulary of Culture and Society* (first published in 1975), Raymond Williams defines his project as "the record of an inquiry into a vocabulary" (Williams, 15) and presents that inquiry as a series of short interpretive essays on carefully selected but commonly used significant words. Between them, his chosen keywords comprise "a general vocabulary ranging from strong, difficult and persuasive words in everyday usage to words which, beginning in particular specialized contexts, have become quite common in descriptions of wider areas of thought and experience." (Williams, 14) They are “significant, binding words in certain activities and their interpretations,” and they are “significant, indicative words in certain forms of thought." (Williams, 15) Keywords, then, are those terms that contain and naturalize the categories through which we form our ideas; in our daily speech they

embody the assumptions through which our views of the world around us emerge.

In Williams's approach to the meaningful keywords of a community's shared discourse is an implicit challenge to any diverse yet conscientious community of practice: to deliberately and skillfully make ourselves aware of the multiple, hidden, sometimes divergent, and often hegemonic meanings of our basic shared terminology. Our task becomes, like Williams's, to inquire into our shared vocabulary rather than merely to use it. So, what are our "big tent" digital humanities community's strong, difficult, persuasive, yet overly-familiarized keywords? They are item-based terms like *code, data, object, document, archive, corpus,* and *collection.* They are activities such as *digitization, preservation, encoding, visualization, interpretation.* They are action words: *hack, make, curate, catalog, blog* and *tweet.* They are *metadata*; they are *keyword.* They may also reside as absences, a significant subset of keyword yet to be fully explored (Klein).

The Digital Humanities Keywords project, launching in June of 2013 at http://www.dhkeywords.org , seeks to take up that challenge by creating a shared space and collaborative conversation for working through and exposing the underlying assumptions—the points of agreement, the sites of tension, the unanswered questions —of the still-emergent and surprisingly complex digital humanities community, through focused attention to the building blocks of its shared discourse.

Like other keywords projects started in the wake of Williams's work (Bennett, Grossberg and Morris; Burgett and Hendler; García and Faherty), Digital Humanities Keywords will be a multi-authored work, but in addition will also engage the flexibility and collaborative features that many peer-based digital humanities projects now offer. As an open access, lightly edited, and openly peer-reviewed Web-site-turned-publication, Digital Humanities Keywords will function not only as an information resource, but as a site of active dialogue and a temporal record of shared communal content as it emerges over time. The project provides guidelines but no hard-and-fast-rules for contributions and it openly solicits beneficial interventions. Comments and dialogue and feedback will reside permanently side-by-side with the primary essays created for the site. Not so much a "how to" as a "how to think about," the Digital Humanities Keywords project explicitly seeks to bring some well-considered yack back to hack, in ways that take fullest advantage of the digital humanities community's existing strengths in open-source online collaboration.

The Digital Humanities Keywords project is sponsored by the Digital Humanities Caucus of the American Studies Association, but is offered to the entire Digital Humanities community with no expectation of an American Studies focus. The current poster presentation works to introduce the project to a broad cross section of the digital humanities community and to solicit participation—which is actively invited at all levels of involvement.

References

Burgett, B., and G. Hendler. (eds). (2007). *Keywords for American Cultural Studies.* New York: NYU Press.

Garcia, E., and D. Faherty (eds). (2011). Critical Keywords in Early American Studies. *Early American Literature* 46(3): 601-632.

Gold, M. K. (2012). The Digital Humanities Moment. *Debates in Digital Humanities.* University of Minnesota Press. ix-xvi.

Kirschenbaum, M. (2012). What is Digital Humanities and What is it Doing in English Departments? *Debates in Digital Humanities.* University of Minnesota Press. 3-11.

Williams, R. (1985). Keywords: A Vocabulary of Culture and Society. Oxford University Press.

Klein, L. F. (2012). American Studies after the Internet. *American Quarterly* 64(4). 861-872. http://muse.jhu.edu/ (accessed 15 March 2013).

Bennett, T., L. Grossberg, and M. Morris (eds). (2005). *New Keywords: A Revised Vocabulary of Culture and Society.* Blackwell.

DH@WIT: Digital Humanities for Undergraduate Design, Engineering, and Management Students

Gleason, Christopher Scott
gleasonc@wit.edu
Wentworth Institute of Technology, United States of America

Wentworth Institute of Technology (WIT), an independent, co-educational, technical design and engineering college located in Boston, Massachusetts, offers a comprehensive interdisciplinary, project-based education that integrates classroom, laboratory, studio, cooperative and experiential learning. Our department of Humanities and Social Sciences (HUSS) is currently working to develop and promote a digital humanities-

inflected undergraduate curriculum. Unlike more traditional humanities programs, we are using digital humanities to prepare students for careers *outside* of academia. The proposed poster presentation will highlight the aspects of DH@WIT that we consider to be unique.

We are already a technology and design based institution immersed in a studio/lab-based culture, so, whereas most institutions with an interest in DH tend to bring technology to humanities, we are *bringing humanities to technology.*

Our department is already highly interdisciplinary and collaborative (with faculty offering courses in Literature, Art History, Film, Music, Philosophy, History, Psychology, Sociology, Political Science, Economics, Cultural Studies, and Communications), and thus we are not hampered by disciplinary boundaries that exist at many other institutions.

In collaboration with our Office of Institutional Research, we have, to date, commissioned three feasibility studies from the Hanover Research Group: ***New Media Programs in the Liberal Arts*** (an initial industry survey of institutions that have led the way nationally in incorporating the study of new media into traditional higher education); ***Market Analysis for a BS in Digital Humanities*** (a report analyzing the market for a bachelor's degree program in digital humanities); and ***Job Opportunities for Digital Humanities Program Graduates*** (which analyzes occupational demand for DH graduates in related industries, builds profiles of relevant positions in related industries, lists relevant local employers, examines the educational and work experience of current professionals with a professed interest in DH, and surveys the sponsored funding environment for DH programs).

Our research suggests that DH graduates with a strong background in computer application design and programming are likely to have the best prospects for a job outside of academia. A more tech-training-oriented program would also align well with W.I.T.'s other institutional offerings. Including training in advanced technical skills such as graphic design or computer engineering, the DH@WIT curriculum offers significant advantages to students entering the workforce because their interests and skills are applicable to many opportunities in a wide range of industries. Overall, nationwide employment for occupations directly related to DH degree programs appears to be growing steadily. Thus, we are proposing a new degree program at Wentworth: a Bachelor of *Science* in Digital Humanities. This would be an interdisciplinary program of study that combines theoretical and practical courses, with the goal of educating new digital and media specialists for the growing knowledge and information economy. Such a program would provide students with a multi-disciplinary foundation in visual and digital literacy and competency.

Because the nature of DH work is applied and project-based, students in the BS program will have hands-on training in studio-based classes, in addition to the theoretical, critical, social, and ethical contexts for thinking about the making, and critique, of new knowledge through Digital Humanities/New Media research and applications. Learning outcomes will include digital literacy, visual literacy, rhetorical competence (visual and verbal), cultural awareness, creative self-direction, and intellectual curiosity.

A BS in Digital Humanities will include a cooperative education (co-op) requirement. Wentworth offers one of the most comprehensive co-op programs of its kind in the nation. The experience complements traditional classroom learning with a chance to build important skills and professional connections. Co-op experiences are directly related to our students' fields of study. Students earn income and do not pay tuition during co-op terms. This kind of co-op opportunity is rare for an undergraduate program in humanities.

DH@WIT is conceived as our own unique response to the "two cultures" problem noted by Thomas Bartscherer in his introduction to *Switching Codes: Thinking Through Digital Technology in the Humanities and the Arts:* "an attempt to bring scholars and artists into more robust dialogue with computer scientists and programmers"(2). We will prepare our students to serve as cultural-technological inter*media*ries.

References

Bartscherer, T., and R. Coover (2011). Switching Codes: Thinking Through Digital Technology in the Humanities and the Arts. Chicago: University of Chicago Press.

Job Opportunities for Digital Humanities Program Graduates. Hanover Research Group. Washington, DC. July 2012.

Market Analysis for a BS in Digital Humanities. (2011). Hanover Research Group. Washington, DC. September.

New Media Programs in the Liberal Arts. (2011). Hanover Research Group. Washington, DC. February.

Debates in the Digital Humanities: Scholarly Publishing Across Print/ Digital Streams

Gold, Matthew K.
mgold@gc.cuny.edu

CUNY Graduate Center, United States of America

Armato, Douglas
armat001@umn.edu
University of Minnesota Press

Davis, Zach
zach@castironcoding.com
Cast Iron Coding

Slaats, Matthew
mslaats@gc.cuny.edu
CUNY Graduate Center, United States of America

Abrams, Mark
lookmark@earthlink.net
CUNY Graduate Center, United States of America

Debates in the Digital Humanities, an edited collection featuring contributions from over forty DH scholars and practitioners, straddles the line between print and digital publication. The first edition of the printed text, which was published by the University of Minnesota Press in January 2012, was composed predominately of essays but also incorporated a variety of web-based materials such as blog posts, tweets, and wiki pages. The printed book was, from the earliest stages of the publication process, intertwined with digital platforms: following the model of peer-to-peer review described by Kathleen Fitzpatrick in *Planned Obsolescence*, all essays in the book were part of a semi-public, web-based review process that mixed new forms of peer-to-peer review with more traditional models of publisher-based blind peer review.

In an attempt to move the university-press based print publication process along at a rapid pace, the book went from initial conception and solicitation of essays to printed publication in the space of a single calendar year, a timeline that involved substantial efforts from contributors and from the Press. The book was conceived of less as an attempt to create a monumental, standard reference guide for the field than as a snapshot of current conversations within it a key moment of growth, with the primary purpose of introducing DH to scholars unfamiliar with its projects, practitioners, and debates.

Debates within the field have not stopped with the publication of the book, of course, so the challenge now facing the editorial team is to create an open access (OA) edition of the text that goes beyond the basic task of making the contents of the print edition available in digital form. In January 2013, an expanded, OA edition of the text will be published on a new platform created by a team of technologists associated with the GC Digital Scholarship Lab at the CUNY Graduate Center, in partnership with the University of Minnesota Press. This platform is being created from a fork of the Prism tool for collective interpretation that was released by the University of Virginia's Scholar's Lab in 2012 (Lestock) as part of its innovative Praxis Fellows Program (Nowviskie). The publication platform will feature enhanced social reading experiences that will include shared/social highlighting of the text, a dynamic index of the book built from reader interactions, and a fine-grained commenting system that will allow readers to associate comments with specific words and phrases within the text. Technical challenges for the platform include the creation of a system that will allow reader markup and commenting to be associated with a text even as the text itself is updated over time. Like Prism itself, all code associated with the new publication platform will be shared freely on Github, thus enabling others working on similar projects to use and reshape the codebase. In this respect, the online edition of *Debates in the Digital Humanities* is itself a DH project.

Beyond the technical infrastructure of the platform, this poster will examine the ways in which the *Debates in the Digital Humanities* project is moving from a printed book to an open-access digital stream that will be harvested at various points for new print and digital publications. These will include new printed editions of *Debates in the Digital Humanities*, targeted publications on specific DH topics and themes, and a series of DH annuals that attempt to capture important moments of debate on a yearly basis. And as the expanded edition of the text takes new forms, the publication process will be opened up and will become more inclusive, with open calls for submissions and fully public rounds of peer review. The project is thus attempting to stake out ground as a publication space that lies somewhere between a printed book, a book series, an academic journal, and a blog, one that is actively engaged in exploring both print and digital instantiations of a text as *events* that take place along a larger continuum of reading and writing around a set of shared issues. The project provides an important example of experimentation by a university press, in partnership with another academic institution, at a moment when the future of such presses is consistently being called into question.

This poster will present the project in its many forms and will attempt to delineate the ways in which the project itself instantiates the values of DH (defined by Lisa Spiro in *Debates in the Digital Humanities* as the values of openness, collaboration, and connectedness, diversity, and experimentation).

References

Fitzpatrick, K. (2011). Planned Obsolescence: Publishing, Technology, and the Future of the Academy. New York: New York University Press.

Lestock, B. (2012). Announcing Prism! Scholar's Lab, University of Virginia Library. 1 May 2012. Web. http://www.scholarslab.org/announcements/announcing-prism/

Nowviskie, B. (2011). Announcing the Praxis Program. Scholar's Lab, University of Virginia Library. 24 August 2011. Web. http://www.scholarslab.org/praxis-program/announcing-the-praxis-program/

Spiro, L. (2012). 'This Is Why We Fight': Defining the Values of the Digital Humanities. In Gold, M. K. (ed.) *Debates in the Digital Humanities*. Minneapolis: University of Minnesota Press.

Knitic — The Revolution of Soft Digital Fabrication

Guljajeva, Varvara
varvarag@gmail.com
Estonian Art Academy, Estonia

Canet Sola, Mar
mar.canet@gmail.com
Interface Cultures, University of Linz for Art and Design

The paper points out the rapid development of digital fabrication, the influences and importance of open source in this field, and overlooked manufacturing method that is textile fabrication. By introducing our practical work and research we demonstrate the potential of craft in the era of digital fabrication. Also the works by other artists and designers involved in improving and applying obsolete electronic knitting machines are covered.

We have started our research on knitting machines in the beginning of 2012 through our art project SPAMpoetry (Guljajeva 2012) (see Fig. 1.). We have purchased an old Brother knitting machine (Fig. 2.) in order to hack the uploading system and knit poems from SPAM. The research on reverse engineering of knitting machine made us realize that the electronic knitting machine was the first digital manufacturing tool at home that has been totally overlooked in the age of digital fabrication. Hence we got an idea and motivation for developing Knitic — an open hardware and integrate it to the field of digital fabrication (Canet and Guljajeva 2012).

Fig. 1
SPAMpoetry

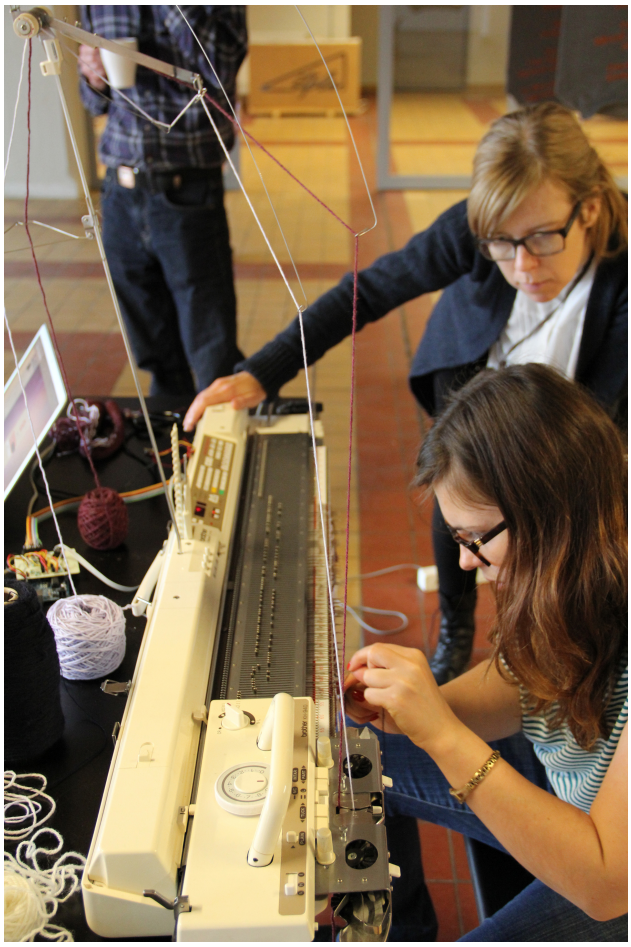


Fig. 2
making a workshop on application of a modified Brother KH930 knitting machine.

What is Knitic? It is an open hardware (see Fig. 3.), which controls an obsolete Brother knitting machine from

1980s via Arduino (open source micro controller). Knitic does not use a floppy emulation or knitting machine's keypad simulation, like previous hacks. Instead, the open hardware is the new 'brain' of a knitting machine that allows real-time control over the needles (see Fig. 4.). It means, one can knit as long patterns as desired and modify the pattern on the fly. Knitic has one more important advantage: it is compatible with all Brother electronic machines. Maybe also with punch-card ones, this needs to be tested though. How come? Because we do not use any Brother electronics but just sensors' output and solenoids' input of a knitting machine.

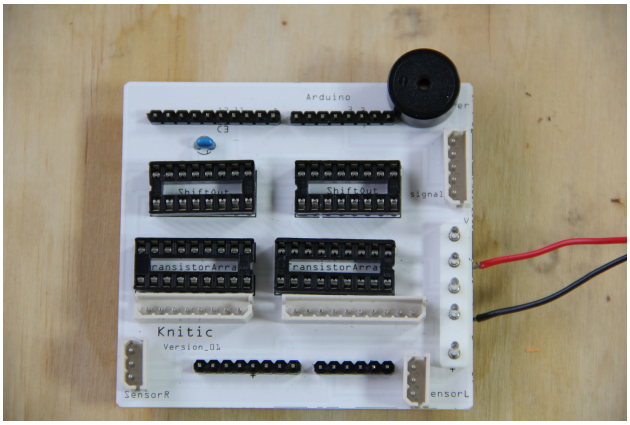


Fig. 3
PCB of Knitic

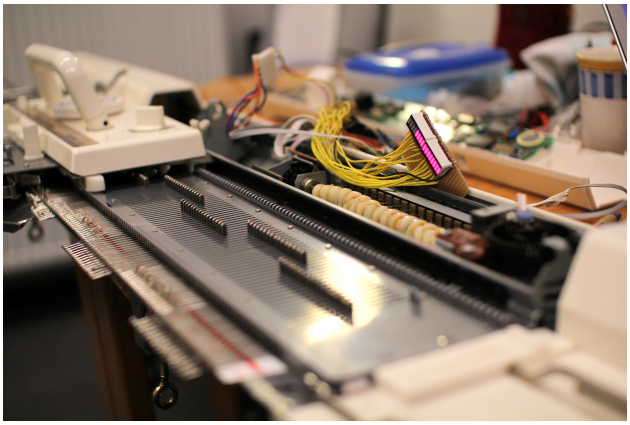


Fig. 4
Testing Knitic. Knitting machine's new brain is able to control needles according to the pattern.

Why are we developing open hardware for knitting? Digital fabrication is gaining importance. The numbers of Fab Labs, persons possessing digital fabrication tools, and open hardware are increasing. Furthermore, the number of start-ups and small-scale companies applying digital fabrication devices as their core business idea is increasing. Makers, designers and artists, who have invested in buying a 3D printer and/or laser cutter, in addition to their work are manufacturing for others, too. Now makers also replicate machines and sell, which all in all pays back the investments sooner.

Hence, society is shifting towards personal and custom manufacturing that is strongly supported by the information age. In the words of Neil Gershenfeld the founder of Fab Labs' model: 'the real impact of digital communications and computation came in giving ordinary people control over the information in their lives; digital fabrication will likewise give individuals control over their physical world by allowing them to personally program its construction.' (Gershenfeld 2007, 241).

However, all this innovation is around certain tools, mainly laser cutters, 3D printers, and CNC machines. At the same time textile fabrication has been overlooked. In the end, it is a shame to forget early fabrication methods, which can be adjusted for digital age needs. Also, re-application of obsolete media and integration of craft are interesting and novel approaches in the field of digital fabrication.

We believe that all these results could be augmented if textile fabrication is added as an option for open manufacturing. Individuals, who are experimenting with and making their living from digital fabrication practices could have more possibilities for creation as well as business. And what is more important, more people could be involved, especially the ones who are skilled in handcrafts like knitting and sewing. Hence, introducing this overlooked manufacturing field will certainly bring innovation, as well as novel business and collaboration models.

To tell more, there is a growing community of artists and designers working with and improving hacked obsolete electronic knitting machines. For example, Becky Stern from MAKE Magazine introduced the first tutorial how to modify a knitting machine (Stern, 2010). Andrew Salamone an artist from New Year has knit a body of work on those machines. The practice of Fabienne Serriere is a good example of one-person manufacturing. She has modified a knitting machine and now produces knitted items by applying parametric design approach while making her own patterns (Serriere 2011).

In addition to that, the machine that was produced for home-use in late 1980s, actually allows to knit big-scale and custom-made items. For instance, we have knit a car Kombi on Brother KH930 machine with Knitic (see Fig. 5,) and a number of SPAMpoetry pieces. Concerning innovation, knitting has a big potential in the field of smart textiles. At the moment the most of work in this field is done on fabrics. Hence, knitting is completely unexplored field from this point of view. For example, thermo chromic and UV pigments could be applied on yarn. Also conductive yarn is a thing to try out. Hence, we see lots of room for creativity and innovation in the field of knitting.

Fig. 5
Knitted Kombi on the streets of Belo Horizonte, Brazil.

Why does open source matters? In our point of view, open code, hardware, and design are the reasons for the success of digital fabrication field. For instance, Lipson and Kurman write about the phenomena of a factory at home and one-person industries, which is not a vision or future prediction but already a reality. There are a number of proofs for such a claim, but the most vital ones are open source hardware and software, and an active community around the rising paradigm. For example, 3D printers that were for industrial use and not affordable for individuals, can now, in 2012, be purchased for 1000 euros. Obviously, an industrial machine has better specifications from an open source one, but still a self-assembled RepRap can be applied for prototyping, a small-scale and customised production, and finally for self-replication. Moreover, the price of the machine is dropping and features improving because the machine is an open hardware! There are lots of 3D printers that are open source and through the innovation and contribution of the whole community the development curve is extremely rapid.

Concerning further reasons for the advent of digital fabrication, open design as well as software play an extremely important role. Thanks to the database of designs that are available online, like Thingiverse.com, one can find a huge number of 3D models as well as share their own designs freely. Hence, even non-experts are able to start experimenting and producing desired items. In addition to that, open code is also crucial for understanding and improving the performance of digital fabrication machines.

Coming back to Knitic, we have opened all our research in order to achieve similar effect in the community of makers as described above. Our ultimate goal is to contribute with completely open source knitting machine that can be produced by laser-cutting and 3D printing its parts. Hence, an open source knitting machine will not depend on availability of discontinued Brother electronic knitting machines.

We believe that textile fabrication has a huge potential in the age of digital fabrication and customisation. Moreover, knitting is a skill that humanity has been using for ages. Hence, there are lots of experts, knowledge, learning and production material, tools, etc. On the contrary, the ability to 3D print or laser-cut is the competence of very few people. It means, introducing craft in general to the desktop manufacturing communities and Fab Labs will bring more people and gender balance to these networks. Furthermore, the encounter of different skills and disciplines will most likely constitute innovation and creativity.

In the end, it is curious how an electronic knitting machine, the first digital manufacturing tool at home has been forgotten by digital fabrication labs and open hardware developers. Therefore, we are confident in the importance of our research project and contribution to the field of personal manufacturing. Moreover, our research and development of open source knitting machine is a perfect example of artists developing their own tools for their work. And that is what is happening in the world of open source hardware and software that affects greatly art, design, and manufacturing fields.

And finally, in our point of view it is impossible to talk about the shift of production paradigm by observing and describing the phenomenon of Fab Labs and novel open source machines that are able to produce hard-surface items, while excluding all other areas of manufacturing.

To sum up, since knitting is a well-known craft and there are lots of experts, it is a shame to run after new technology and forget good old skills. On the contrary, innovation should take advantage of existing knowledge.

References

Canet, M., and V. Guljajeva (2012). Knitic. http://www.knitic.com/ (accessed 14 March 2013).

The Economist. (2011). 3D printing. The printed world. http://www.economist.com/node/18114221 (accessed 1 September 2012).

Fab Lab International. http://fablabinternational.org/ (accessed 20 February 2013).

Gershenfeld, N. (2007). *Fab: The Coming Revolution on Your Desktop — from Personal Computers to Personal Fabrication*. Basic Books.

Goodspeed, T. (2010). Hacking a Knitting Machine's Keypad. http://travisgoodspeed.blogspot.com.br/2010/12/hacking-knitting-machines-keypad.html (accessed 14 January 2013).

Guljajeva, V. (2012). SPAMpoetry. http://www.varvarag.info/spampoetry/ (accessed 14 March 2013).

Igoe, T., and C. Mota (2011). Astrategist's Guide to Digital Fabrication. m.strategy-business.com/article/11307?gko=63624 (accessed 1 September 2012).

Kurman, M., and H. Lipson (2010). Factory @ Home: Emerging Economy of Personal Fabrication. http://web.mae.cornell.edu/lipson/FactoryAtHome.pdf (accessed 1 March 2013).

Sierrere, F. (2011). Mate cosies: warm hands, cold mate. http://fabienne.us/ (accessed 2 February 2013).

Solon, O. (2013). Digital Fabrication is so much more than 3D printing. In *Wired Magazine*, published 13 March 2013. http://www.wired.co.uk/news/archive/2013-03/13/digital-fabrication (accessed 14 March 2013).

Stern, B. (2010). How-To: Hack Your Knitting Machine. http://blog.makezine.com/craft/hack_your_knitting_machine/

TXM Portal: Providing Online Access to Textometric Corpus Analysis

Heiden, Serge
slh@ens-lyon.fr
ICAR Laboratory, ENS de Lyon - CNRS, France

Lavrentiev, Alexei
alexei.lavrentev@ens-lyon.fr
ICAR Laboratory, ENS de Lyon - CNRS, France

This poster presents the TXM portal, a software providing online access to textometric corpus analysis. Textometry is a computerized methodology of corpus analysis combining qualitative and quantitative tools applicable in various fields of the humanities (linguistics, literary studies, geography, philosophy, history, etc.). This methodology was initially developed in France in the 1980's under the name of lexicometry and a number of software products implementing various analytical tools were developed. TXM is a new generation of open-source software built on a modular basis bringing together previous textometric techniques and state-of-the-art text encoding and corpus-building technologies (Unicode, XML, TEI, NLP) (Heiden, 2010; Heiden et al., 2010; Pincemin et al., 2010). The word search engine used for TXM is provided by the Open CWB open-source project (http://cwb.sourceforge.net) and syntactic structures can be queried using the TigerSearch engine (provided that the corpus is syntactically annotated in Tiger XML format) (Lezius 2002). Statistical analyses are performed using the R library (http://www.r-project.org). Other search engines and libraries can be plugged in the TXM platform as necessary.

The TXM software is available in the form of a desktop application (for Windows, Mac and Linux) and of a web portal application sharing a common "toolbox" for corpus building, query and statistical analysis. Most of the corpus analysis features are the same in both applications, however a special attention in the poster will be given to the features that are only available in the portal version.

It should be noted that corpus import and annotation features are only available in the desktop version. The portal version allows the administrator to upload previously compiled "binary" TXM corpora.

The major specific feature of the TXM portal is the management of user registration and access rights to the corpora with the possibility to specify access conditions for each individual text of the corpus (e.g. limitation of context size in concordances). This is important for copyrighted texts where owners may wish to prevent users from copying an entire text or a substantial part of it. User accounts and profiles can be edited by the portal administrator through the web interface. Customized web pages ("home", "help" and "contact") can be created for each user profile. Internationalization feature is available for the portal interface and user web pages (the current portal distribution provides English and French interface).

Another feature that is only available in the TXM portal version is the creation of subcorpora by selecting texts with a special interface. It allows the user to choose various criteria to select texts (depending on the metadata available for the corpus), to add or remove texts individually and to visualize the dimensions of the subcorpus in number of words or texts.

The basic tools of textometric analysis are available in all TXM versions. These include creating corpus partitions for contrastive analysis, building indexes and concordances of word or text patterns, display of one or several alternative text edition versions (including facsimile images). One can also search for collocates of a particular word. Statistical analysis tools are available for corpus partitions. These include computing specificity of word or text patterns and correspondence analysis. The results of corpus queries can be downloaded for further analysis in the form of CSV tables.

The TXM portal software is available for free under the GNU GPLv3 license from the sourceforge development site (http://sourceforge.net/projects/txm). A demo TXM portal where the various tools can be tested on sample corpora is accessible at the following address: http://txm.risc.cnrs.fr/demo. TXM portal is currently used in production to provide regular access to the Base de Français Médiéval old

French corpus (http://txm.bfm-corpus.org/bfm) to a user community of 200 medievalists from around the world.

References

Pincemin, B., S. Heiden, M.-H. Lay, J.-M. Leblanc, and J.-M. Viprey (2010). Fonctionnalités textométriques: Proposition de typologie selon un point de vue utilisateur. In S. Bolasco et al. (Eds.), Statistical Analysis of Textual Data — Proceedings of 10th International Conference JADT 2010, Edizioni Universitarie di Lettere Economia Diritto, Rome, 9-11 juin 2010.

Heiden, S. (2010). The TXM Platform: Building Open-Source Textual Analysis Software Compatible with the TEI Encoding Scheme. 24th Pacific Asia Conference on Language, Information and Computation. Éd. Kiyoshi Ishikawa Ryo Otoguro. Institute for Digital Enhancement of Cognitive Development, Waseda University, 2010. 389-398. online.

Heiden S., M. Jean-Philippe, and P. Bénédicte (2010). TXM : Une plateforme logicielle open-source pour la textométrie – conception et développement, in Sergio Bolasco& al (eds), Statistical Analysis of Textual Data - Proceedings of 10th International Conference JADT 2010.

Heiden S., and L. Alexei (2012). The TXM Portal Software Giving Access to Old French Manuscripts Online, Proceedings of the 1st Workshop on Adaptation of Language Resources and Tools for Processing Cultural Heritage Objects, Seventh International Conference on Language Resources and Evaluation, Istanbul, Turkey, 2012.

Lezius, W. (2002). TIGERSearch – Ein Suchwerkzeug für Baumbanken // Proceedings der 6. Konferenz zur Verarbeitung natürlicher Sprache (KONVENS 2002). Saarbrücken. 2002. http://konvens2002.dfki.de.

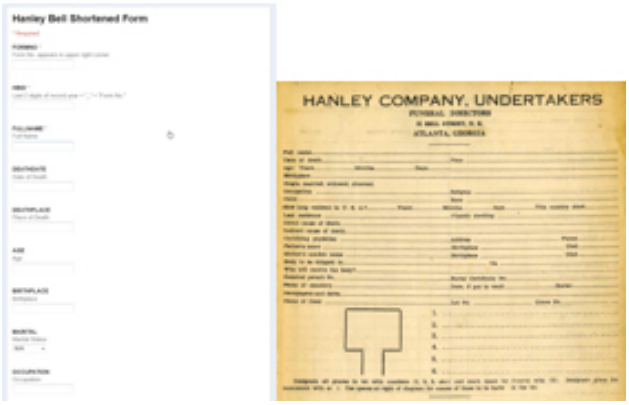
The Atlanta Map project: TEI and GIS collaborate to create a research environment

Hickcox, Alice
alice.hickcox@emory.edu
Emory University, United States of America

Page, Michael C.
michael.page@emory.edu
Emory University, United States of America

Gue, Randy
randy.gue@emory.edu
Emory University, United States of America

Librarians from Emory's Manuscript, Archives, and Rare Book Library and the library's Data Center envisioned building a base of historic maps from library mao collections that would allow scholars to research aspects of the history of Atlanta, especially its racial history. As a first step the librarians digitized and georeferenced a 1928 Atlas of Atlanta. At the same time the library also digitized the 1928 city directory of Atlanta, which was then proofread and marked up in TEI. In addition funeral home records from 1930 were transcribed and marked up to be added to the text collections. The addition of texts encoded in TEI allows flexible searching, giving the public access to the text records. The GIS work allows researchers to visualize and analyze complex patterns of development. This poster will show the work being done in the library by the volunteer labor of a variety of librarians, and some of the connections between the TEI and GIS use of the data.



The TEI development

The text portion of the research is hosted in an eXist database, and can allow queries from public users and researchers to search names and street locations. To prepare the data staff and students proofed the city directory, and marked up the address listings in TEI. Records from the funeral home from the same period have been transcribed and will be encoded and available for searching also. The xml-encoded city directory document allows the production of lists of names and addresses in various forms to automate much of the geocoding. When the geographic references from the geocoder are added to the coding in the city directory, a search by name will be able to bring up map

references showing the locations associated with that name. Web based users can search and to locate names and streets with Google Maps or Google Earth using the georeferenced locations harvested from the geocoder and added back the TEI encoding.

The GIS development

The GIS coding allows researchers with access to GIS software to pursue research via the data encoded in layers of the geographic maps. While the construction of a geodatabase is core to the GIS users, the development of a geocoding reference layer and a specific address locator is what enables us to construct geocoding tools for Atlanta circa 1928. The geocoding reference layer's construction is facilitated by the automation allowed by the TEI encoding of the city directory.

A geocoder transforms data such as addresses into a location on the surface of the earth so the data can be quickly plotted on a map. Using a Geographic Information System (GIS), this project will create an application similar to Google Maps, but it will be a digital research tool for Atlanta from the late 1920s through the early 1950s, not for contemporary Atlanta. The geocoder assigns addresses and maps all of the 250,000 building footprints in Atlanta and its environs in 1930. Students, faculty, and researchers can then add layers and tag attributes to a series of addresses in the historic city. With this research tool, researchers interested in Public Health could explore the correlation between topography, elevation, and incidence of disease in Atlanta, urban historians could map out more complicated, nuanced, and fluid racial residential patterns, political scientists could map the locations of fire hydrants, manhole covers, and sidewalks to provide accurate data about how segregation influenced city services, and historians could map the "watersheds" of churches and discover how far members of a particular congregation lived from a particular church. The goal of the project is to create a digital tool to visualize and analyze historical Atlanta by providing new ways to integrate spatial and non-spatial data in the classroom and in research. This combination of GIS technology and unique datasets will change the way Jim Crow Atlanta is studied.

References

Re-Mapping Segregated Atlanta http://marbl.library.emory.edu/remapping

MARBL Historic Map Collection http://www.digitalgallery.emory.edu/luna/servlet/EMORYUL~3~3

The Digital Orationes Project: Interfacing a Restoration Manuscript

Johnson, Anthony W.
anthony.johnson@abo.fi
Åbo Akademi University, Finland

Juuso, Ilkka
ijuuso@ee.oulu.fi
University of Oulu, Finland

Toljamo, Tuomo
toljatuo@mail.student.oulu.fi
University of Oulu, Finland

Mätäsaho, Timo
timomata@paju.oulu.fi
University of Oulu, Finland

Opas-Hänninen, Lisa Lena
lisa.lena.opas-hanninen@oulu.fi
University of Oulu, Finland

Seppänen, Tapio
tapio.seppanen@oulu.fi
University of Oulu, Finland

Funded by the Academy of Finland (2011-2014), the Digital *Orationes* Project is an interdisciplinary initiative intended to bring an important unpublished Early Modern

manuscript into the scholarly arena. The manuscript, preserved as Lit. MS E41 in the archive of Canterbury Cathedral, was collected, and in part composed, by George Lovejoy (c. 1675), Headmaster of the King's School, Canterbury, after the English Civil War. The texts within it represent one of the most substantial unpublished sources of English School Drama from the period. As well as containing a previously unnoticed adaptation of a pre-war play by a major author (James Shirley), this large volume, comprising 656 folio pages and running to some 230,000 words, includes a number of short plays and dramatized orations written in English, Latin and Greek by the scholars and staff of the King's School. (Amid much else, these works celebrate the Restoration of Charles II to power, re-enact the Gunpowder plot, discuss a wide range of topical issues, and provide a wealth of information about the role of drama in Early Modern schooling.)

The overall aim of the project has been to create a state-of-the-art digital archive which makes the texts in the manuscript available to a wider audience at the same time as it offers new affordances for its scholarly users. In this, we have been responding to, and actively critiquing, best practices within the field of digital editions so succinctly summarized by Pierazzo (2011). The final digital interface for the manuscript will enable the searching of its handwritten text by means of visual recognition of the letter forms as well as the more usual text-based functions relating to the transcribed, translated, and edited manuscript; and simultaneously allow access to a number of higher scholarly functions. Accordingly, the present paper will focus on ongoing developments concerning three elements in the (otherwise, substantially developed) package: a) an image calibration tool (i.e. a tool to flatten pages, scale them consistently and link with the transcript); b) an image search prototype (which can search for graphic features in the manuscript using visual cues); and c) the blueprint for the final *Orationes* user interface. We are presenting these features before they have been packaged into a single polished user-friendly entity – and at the penultimate stage of a larger project involving an international team from four Universities (Oulu, Åbo Akademi, Helsinki [Finland], and Austin, Texas) – in order to make new research available at the same time as we solicit the feedback which will make our final interface as helpful as possible within the DH community.

In its basic form, the *Orationes* manuscript is represented by a rich digital edition that utilizes the high resolution scanned images of the manuscript pages and TEI-compliant transcriptions created by domain experts. As demonstrated in our JADH-2012 presentation (Opas-Hänninen *et al.* 2012), our *modus operandi* has been to approach the work from two opposite directions. First, the team created an uncompromising TEI-XML version of the manuscript (a process which has not only involved rigorous textual transcription but also entailed the translation of the Latin and Greek portions of the material and the identification of features of interest for use in a rich visual interface). Second, it has been working on the production of a reusable software package which, without sacrificing functionality or source integrity, can be used to generate digital editions from similar XML and image source materials in a straightforward manner. In order to combine these two goals successfully, the team navigated the requisite TEI guidelines (tweaking them where necessary as developments within, rather than departures from, the system), constructed an automatic linking mechanism between image and text, and have been creating efficient search mechanisms for TEI data: not to mention an interface that is both intuitive and generic at the same time.

The high resolution scanned images of the manuscript were produced by experts at the Canterbury Cathedral Archives. When beginning the work on image searching, i.e. the recognition of graphical forms such as letters and punctuation, it quickly became evident that in order to be able to carry out any such work, the images of the manuscript would need to be flattened and scaled consistently first, because of slight warping on the inner edges of the pages. Preprocessing will simplify the actual pattern recognition search process, because there will be less variation to account for and thus better results should be achieved; it is also beneficial for creating a clean GUI. Thus we set out to develop a process for calibrating the images, which we think will be very useful for other projects which use similar manuscripts that simply can't be flattened as the scanned images are produced.

Although, since 2012, the Digital *Orationes* interface has been able to allow for full searches of the transcribed text (along with translations from the relevant Latin or Greek passages), links to the apparatus and editorial notes, or a layered comparison (at varying magnifications) between the written text and the transcription, our endeavours in the present move considerably closer to an engagement with the manuscript as an artefact and the sort of operations which a professional palaeographer or historical linguist might require of it. In particular, by developing an optical recognition faculty which is able to identify and search out letters and other graphical forms manifested in the manuscript (such as varieties of dashes and other idiosyncracies of punctuation), the 2013 prototype indicates how our edition might contribute to raising the bar for palaeographers: helping them to search, line up, or compare visual features across the manuscript in an easy, intuitive way. Looking ahead, the presenters of the 2013 poster will also be prepared to engage with the wider perspective of how, adapting a tool developed and presented for other purposes (Juuso *et al.* 2011), our final

interface will also be able to serve historians of language and literature by keying in to etymological dictionaries of English, Latin or Greek, identifying new words and colour-coding lexical items according to the first date of occurrence in the historical corpus.

References

Pierazzo, E. (2011) A Rationale of Digital Documentary Editions, *LLC: The Journal of Digital Scholarship in the Humanities*, 26(4): 463-477.

Opas-Hänninen, L. L., I. Juuso, T. Toljamo, A. W. Johnson, and T. Seppänen (2012). The *Orationes* Project: Bringing a Restoration Manuscript Online. Paper presented at JADH2012, Tokyo, 17 September 2012.

Juuso, I., L. L. Opas-Hänninen, A. W. Johnson, and T. Seppänen (2011). The Time Machine: capturing Worlds Across Time in Texts. Paper presented at DH2011, Stanford 18 June 2011.

Reverse Image Lookup, Paintings, Digitisation, Reuse

Kirton, Isabella
kirton134@googlemail.com
University College London, Information Studies, United Kingdom

Terras, Melissa
m.terras@ucl.ac.uk
University College London, Information Studies, United Kingdom

Once digital images of cultural and heritage material are digitized and placed online, how can we tell if they are copied, disseminated, and reused? This poster explores Reverse Image Lookup (RIL) technologies — usually used to identify unlicensed reuse of commercial photography — to help in assessing the impact of digitized content. We report on a pilot study which tracked a sample of images from The National Gallery, London, to establish where they were reused on other webpages. In doing so, we assessed the current methods available for applying RIL, establishing how useful it can be to the cultural and heritage sectors.

RIL technologies are those which allow you to track and trace image reuse online. The main commercial service, TinEye, available since 2008, finds 'exact and altered copies of the image you submit, including those that have been cropped, colour adjusted, resized, heavily edited or slightly rotated' (TinEye, n.d). Since 2007, Google Image Search has also provided a free service which can find similar images across the Internet. Can these tools provide a useful method for tracking reuse of images of paintings once they are placed online? Kousha et al (2010) published a pioneering study which assessed 'image reuse value' of academic scientific images. We believe ours is the first systematic study to use RIL to look at digitized heritage content.

We choose two samples of paintings from the National Gallery: all paintings held in Room 34 entitled 'Great Britain 1750-1850', containing 26 paintings by 9 artists, just over 1% of their total number holdings (National Gallery, n.d.). We also created a random sample of 6 paintings, from different artistic periods and of varying levels of fame. We analysed the dissemination of these images using TinEye and Google Image Search, using Content Analysis (White and Marsh 2006) to discover the contexts for image reuse.

We then triangulated findings using web access statistics from the National Gallery's Google Analytics account, and from the commercial ISP analysis firm Hitwise. Our results show that the most popular paintings (by access) are the most commonly used elsewhere, but we also uncover a feedback loop which proves dissemination of images online provides direct traffic back to the National Gallery's website. Our content analysis also provides a qualitative analysis of types of image reuse, such as commercial art publishers, blogs, reviews, tourism, image collections, encyclopaedias, other museum websites, DVD cover images, and beyond. We demonstrate that type and volume of image reuse is both subject and artist specific.

This study has allowed us to establish what motivates image reuse in a digital environment. We recommend a framework for data collection that could be used by other organisations. However, we also show that there are limitations to the information that can be gleaned from a study of this kind, due to the problematic implementation of the RIL tools which were not designed for this sector.

Acknowledgments

We thank Charlotte Sexton, Melissa Naylor, and Matt Terrington from the National Gallery, and Mike Tovell from Hitwise.

References

Kousha, K., M. Thelwall, and S. Rezaie (2010). Can the Impact of Scholarly Images be Assessed Online? An Exploratory Study Using Image Identification Technology.

Journal of the American Society for Information Science and Technology, 61(9): 1734

The National Gallery, 'Room 34' http://www.nationalgallery.org.uk/visiting/floorplans/level- 2/room-34 (accessed 12 February 2012.

TinEye, (n.d.) "Frequently Asked Questions". http://www.tineye.com/faq, (accessed 26 September 2011.)

Domas White, M. and E. E. Marsh (2006). Content Analysis: A Flexible Methodology *Library Trends* 55(1): 23-4.

Networking the Belfast Group through the Automated Semantic Enhancement of Existing Digital Content

Koeser, Rebecca Sutton
rebecca.s.koeser@emory.edu
Emory University, United States of America

Croxall, Brian
brian.croxall@emory.edu
Emory University, United States of America

There is increasing work on and interest in social networks in the digital humanities community (Meeks 2011). Analysis is frequently done on digital content —including images (Akdag Salah et al. 2012); email (Hangal et al. 2012); and citation networks (Visconti 2012) — because the data lend themselves to aggregation, conversion, and analysis. Yet despite this flurry of activity, the possibility exists for an exponential jump in network analysis. After all, the holdings and catalogs of galleries, libraries, archives, and museums (GLAMs) include traces of vast paper-based networks, but the data are locked away in forms that don't easily lend themselves to analysis. What if we could open up that content? In this poster, we will report on an attempt to provide tools for archivists to expose the information embedded in the descriptions of their collections as well as a test case for analyzing that data: an examination of the networks of the Irish poets collectively known as "the Belfast Group."

Our goal is to develop software tools and design a workflow to enhance TEI and EAD — documents that are already commonly created and maintained by archivists and text centers — without radically increasing the time and effort involved. The software tools (http://github.com/emory-libraries-disc/name-dropper) consist of a plugin for the Oxygen XML editor and command line scripts that will, first, make use of DBPedia Spotlight to identify and annotate recognized names and other resources within the text and, second, connect to linked-data systems (starting with the Virtual International Authority File [VIAF]) to provide authoritative, scholarly identifiers.[1] The scripts will allow technical users to inspect and tune the results or to automatically tag high-certainty resources, and the plugin will provide a user-friendly interface to review and accept suggested names while editing a document. The enhanced documents should provide significant benefits to GLAMs, allowing them to connect disparate types of content (e.g., digitized texts or photographs from an archival collection) and augment with data from other linked data systems. Furthermore, the enhanced documents will make it possible to expose these data in more machine-readable and research accessible formats. Our tools and workflow could be applied to resources held by different archives (for a different approach, see Blanke et al. 2012). What's more, enhancing these documents helps GLAMs provide a means for researchers to do non-consumptive, social network research on the metadata of collections that might otherwise be closed or problematic in other ways (e.g., restricted correspondence from living authors).

Although our tools are not yet complete, we have already begun preliminary visualization and analysis of network relationships using data that mirrors what we will generate automatically by Summer 2013. The difficulties of defining "the Belfast Group" make for a compelling test case for our attempt to understand networks via data that are newly machine readable. The Group is a contentious network since the label has been variously applied to a weekly writing workshop that ran from 1963-1972, the most famous poets who attended that workshop — including Seamus Heaney, Michael Longley, and Paul Muldoon — or more loosely applied to all of the writers who "put Belfast on the literary map" (Clark 6). The significance of the writing workshop is debated by critics and often rejected by the poets themselves, sometimes vehemently. In contrast to a more formalized group, some scholars identify "an informal community" of poets evidenced by their letters, promotion of each other, and poems dedicated to each other(Drummond 32), connections which are richly documented by archival materials held at Emory University.

Using preliminary data manually generated from a subset of the correspondence EAD, our data suggests a wider set of connections in the Group than traditional scholarly approaches. The latter selectively emphasize the relationships of the most prominent authors and the role of the writing workshop (see fig. 1). Since our data is based on a much larger set of artifacts, as well as their complete

metadata, we find that the locus of poetic activity in Belfast is not so oriented around the workshop (see fig. 2). Once we collect the full dataset via our completed tools and workflow, we will compare it with models generated by traditional scholarly methods, to identify significant gaps and discrepancies in either model.

Providing not only this new analysis of the Belfast Group's network and a report on the development of our tools, our poster presentation at DH 2013 will also include a hands-on demonstration of the software tools and interactive visualizations of network data.

Figure 1.
Graph of relationships inferred from Heather Clark's Ulster Renaissance. Nodes are sized by degree and colored by hub score. The writing workshop is the strongest hub; the trio of large nodes represent Michael Longley, Derek Mahon, and Seamus Heaney.

Figure 2.
Relationship graph based on preliminary correspondence data, sized and colored as in figure 1. Based on this data, the writing workshop does not function as a hub at all, and Paul Muldoon becomes the largest node.

References

Akdag Salah, A. A., et al. (2012). Exploring Originality in User-Generated Content with Network and Image Analysis Tools. *Digital Humanities 2012.* University of Hamburg. 19 July 2012.

Blanke, T., et al. (2012). Information Extraction on Noisy Texts for Historical Research. *Digital Humanities 2012.* University of Hamburg. 19 July 2012.

Clark, H. (2006). *The Ulster Renaissance: Poetry in Belfast, 1962-1972.* Oxford: Oxford University Press.

Drummond, G. (2005). The Difficulty of We: The Epistolary Poems of Michael Longley and Derek Mahon. *The Yearbook of English Studies, Irish Writing since 1950* 35: 31-42

Hangal, S. (2012). Processing Email Archives in Special Collections. *Digital Humanities 2012.* University of Hamburg. 20 July 2012.

Litta Modignani Picozzi, E., J.Norrish, and J. M. Monteiro Vieira (2012). Complex entity management through EATS: the case of the Gascon Rolls Project. *Digital Humanities 2012. University of Hamburg.* 18 July 2012.

Moretti, Franco et al. (2011). Networks, Literature, Culture. *Digital Humanities 2011. Stanford University.* 21 June 2011.

Meeks, E. (2011). More Networks in the Humanities or Did books have DNA? *Digital Humanities Specialist.* 6 December 2011. Web. 1 November 2012. https://dhs.stanford.edu/visualization/more-networks/ .

Mendes, P. N., et al. (2011). DBpedia Spotlight: Shedding Light on the Web of Documents. *Proceedings of the 7th International Conference on Semantic Systems (I-Semantics).* Graz, Austria. 7–9 September 2011. http://www.wiwiss.fu-berlin.de/en/institute/pwo/bizer/research/ publications/Mendes-Jakob-GarciaSilva-Bizer-DBpediaSpotlight-ISEM2011.pdf.

Pitti, D., et al. (2011). The Social Networks and ARchival Context Project. *Digital Humanities 2011* Stanford University. 22 June 2011.

Pitti, D., et al. SNAC: The Social Networks and Archival Context Project. http://socialarchive.iath.virginia.edu/. (accessed 29 October 2012.)

Visconti, A. View DHQ: Citation Network Visualization for Digital Humanities Quarterly. http://digitalliterature.net/viewDHQ/. (accessed 1 November 2012.)

Notes

1. It is in the use of existing systems (DBpedia) and vocabularies (VIAF) that distinguishes this project from the Entity Authority Tool Set (EATS), which involves setting up and maintaining one's own authority server. See Litta Modignani Picozzi, Norrish, and Monteiro Vieira (2012).

Normalisation in Historical Text Collections

Lawless, Séamus
seamus.lawless@scss.tcd.ie
Trinity College Dublin, Ireland

Hampson, Cormac
cormac.hampson@scss.tcd.ie
Trinity College Dublin, Ireland

Mitankin, Petar
pmitankin@fmi.uni-sofia.bg
Sofia University, Bulgaria

Gerdjikov, Stefan
st_gerdjikov@abv.bg
Sofia University, Bulgaria

Improved full-text search, named-entity recognition and relationship extraction are all key research topics across many areas of technology, with emerging applications in the intelligence, healthcare and financial fields amongst many others[1] . In Digital Humanities, there is a growing interest in the application of such Natural Language Processing (NLP) approaches to historical texts[2] with a view to improving how a user can explore and analyse these collections[3] [4] [5] [6] . However, the text contained in handwritten historical manuscript collections can often be 'noisy' in nature — with variation in spelling, punctuation, word form, sentence structure and terminology. This is particularly the case with collections written in archaic language forms, such as Early-Modern English. Multiple studies have concluded that the applicability of modern NLP tooling to such historical texts has been very limited due to this inherent noisiness in the texts. This historical language barrier hinders the accessibility and thus the potential exploration and analysis of many significant historical text collections. This paper will discuss the normalisation of historical texts as a solution to this problem and examine how normalisation can improve the analysis, interpretation and exploration of these collections.

Normalisation is the process of transforming text into a single canonical form, in this case, the modern equivalent of the language. Once this has been completed, the texts can be processed using current NLP techniques and technologies. However, the normalisation of historical texts presents a difficult challenge in itself.

Much research has been undertaken in an attempt to cope with the correction and normalisation of text produced by Optical Character Recognition (OCR), speech recognition, instant messaging etc. which show similar characteristics to those of historical texts. One technique which has been applied is the use of a historical lexicon, supplemented by computational tools and linguistic models of variation. However, because of the absence of language standards, multiple orthographic variations of a given word or expression can be found in a collection of material, even in the same document. As a result, the quality of the results achieved, even after normalisation, has not been satisfactory. Researchers have also noted a general lack of tools and resources specialised to this domain.

This paper will present the normalisation research conducted as part of the CULTURA project, which has developed techniques for the normalisation of a 17th century manuscript collection written in Early Modern English, The 1641 Depositions[7] . CULTURA analyses the artefacts and through the application of novel linguistic models of variation, enables normalisation techniques to

remove issues of inconsistency in spelling, grammar and punctuation. The technologies developed and applied have had to solve issues arising from the need to contend with noisy inputs, the impact noise can have on downstream applications, and the demands that noisy information places on document analysis. The normalisation of texts in Early Modern English can be interpreted as a special (restricted) case of translation. Using this intuition, a methodology was developed based upon statistical machine translation models. The key ingredient of this approach is a new translation module that further develops known OCR correction techniques.

Once the content has been normalised, further analysis is conducted to perform named entity and relationship extraction. This identifies the individuals, events, dates etc. within the collection and the relationships between these entities. It encodes this data in a manner which promotes interoperability with other collections of related cultural heritage material.

The normalisation process allows CULTURA to perform disambiguation on the text. For example, one of the main players mentioned many times in the 1641 Depositions collection is Sir Phelim O'Neill. In addition to huge variations in the spelling of his name (Phelin, felim, ffelim, O'Neill, Neil, Onell etc.), he is also referred to in a number of different ways throughout the depositions (The O'Neill, Phelim MacTurlough MacAodh, 'The rebel leader' etc.). Normalisation is used to address this variation and also identify the context in which he is mentioned – for example, is he accused of being directly involved in an incident or merely mentioned as a known rebel figure? A combination of NLP and SNA techniques are then used to identify relationships between this individual and other people in the community. He can also be associated with specific events and locations over time, all of which would be impossible without the collection having been normalised.

While the normalisation of the collection has facilitated this application of further NLP techniques, it has also supported improved interaction with the collection for novice researchers and members of the general public. People without a deep scholarly knowledge of such collections can find it easier to interact with, and gain an understanding of, the normalised versions of the transcribed text rather than the originals.

The un-normalised text of the collection is still available to "professional" scholars who use the CULTURA portal and every effort is made during normalisation to ensure that changes that implicitly involve interpretation of the text, or that go beyond normalisation, are avoided. The normalisation process should never impact upon the semantic content of the collections.

A number of important advancements have been achieved by CULTURA, including the development of:

- A normalisation algorithm[8] called Regularities Based Embedding of Language Structures (REBELS).
- An integrated REST based web service for the implemented normalisation module.
- A tool for manual annotation which makes the normalisation process as simple as possible, and helps to verify consistency of annotations and to help the resolution of detected conflicts.

These developments all support a more effective text normalisation process and improve the effectiveness of the entity and relationship extraction procedures which follow. Experimental results have demonstrated that the normalisation averages 98% accuracy in the translation of regular words (tokens) in the 1641 Depositions collection.

In order to show the general applicability of the approach, the methodology was applied to another Early Modern English text collection, the Innsbruck Letters Corpus, part of the Innsbruck Computer-Archive of Machine-Readable English Texts (ICAMET) corpus. The results were compared to two state of the art systems, Moses[9] and VARD2[10] . The CULTURA approach achieved an accuracy of 83% with minimal training, when compared with 77% and 65% for Moses and VARD2 respectively.

In summary, this paper will discuss the normalisation of historical texts, and the analysis and feature identification that normalisation facilitates. These are increasingly important processes for "noisy" cultural heritage resources, and provide significant benefits to the analysis, interpretation and exploration of these collections. Together, they can:

- Improve the quality and re-usability of artefacts by normalising spelling, punctuation and nomenclature.
- Facilitate deeper interrogation of the material.
- Identify features of the collection — individuals, locations, dates etc.
- Enable social network analysis on these features, identifying those with the greatest influence.
- Open new pathways for the exploration and interrogation of the resources for both novices and experts.
- Add structure and logic to otherwise featureless material, enabling new forms of engagement, use and enjoyment of the content.

Notes

1. Sunita Sarawagi. Information extraction . FnT Databases, 1(3), 2008.
2. The CHLT Project, funded under EU FP5, aimed to create an infrastructure for pioneering International

Digital Library Technology (IDLT), and a range of IT applications for use within digital collections (with special emphasis on early modern Latin, classical Greek, and Old Norse texts), including generic tools for multi-lingual information retrieval; concept identification and visualisation; vocabulary analysis and syntactic parsing.
3. 3 The IMPACT project http://www.impact-project.eu
4. Annette Gotscharek, Andreas Neumann, Ulrich Reffle, Christoph Ringlstetter, and Klaus U. Schulz. 2009. Enabling information retrieval on historical document collections: the role of matching procedures and special lexica. In *Proceedings of The Third Workshop on Analytics for Noisy Unstructured Text Data* (AND '09). ACM, New York, NY, USA, 69-76.
5. A. Gotscharek, U. Reffle, C. Ringlstetter, and K. U. Schulz. On lexical resources for digitization of historical documents. In DocEng '09: Proceedings of the 9th ACM symposium on Document engineering, pages 193--200, New York, NY, USA, 2009.
6. A. Ernst-Gerlach and N. Fuhr. Retrieval in text collections with historic spelling using linguistic and spelling variants. In JCDL '07: Proceedings of the 7th ACM/IEEE-CS joint conference on Digital libraries, pages 333–341, New York, NY, USA, 2007. ACM.
7. The 1641 Depositions – http://1641.tcd.ie
8. Gerdjikov, S. "Some algebraic properties of alignments of words." In Comptes rendus de l'académie bulgare des science. 2012.
9. Moses is a statistical machine translation system - http://www.statmt.org/moses/
10. A. Baron and P. Rayson. Automatic standardisation of texts containing spelling variation: How much training data do you need? In: Proceedings of the Corpus Linguistics Conference. Lancaster University, Lancaster, 2009.

A Comparative Study of Astronomical Clock towers in Europe and China based on their detailed 3D modeling

Li, Guoqiang
guli@vision.ee.ethz.ch
ETHZ (Swiss Federal Institute of Technology Zurich), Switzerland

Van Gool, Luc
vangol@vision.ee.ethz.ch
ETHZ (Swiss Federal Institute of Technology Zurich), Switzerland

Summary

The rapid development of computer graphics and imaging provides the modern archeologist with several tools to realistically model and visualize archeological sites in 3D. This creates a bridge between humanities and computer science. This project, integrating interdisciplinary research methods, will explore an effective way to reconstruct tangible cultural heritage. In particular, our research will focus on the structure and mechanisms and art values of the astronomical clock towers both in China and Europe, employing modeling tools such as 3DSMAX and SOLIDWORKS to model the astronomical clock-towers, and making use of JavaScript and VRML technology to control the display. Then, we will use the ADAMS software to analyse the kinetic parameters and rotation periods. After extensive consultation of the literature, we will contrast the different astronomical clock towers in terms of mechanics, astronomy, aesthetics, etc. Understanding each other's culture will be one of the main challenges of the next generation of world's citizens. Such a study will not only bring us fancy digital exhibits related to astronomical clock towers from different cultures, but also stimulate cross-fertilization between humanities and technology, raising citizens' awareness towards cultural heritage. Europe has quite a long tradition of building astronomical clock towers, and the candidate's recent research focused on their Chinese ancient counterparts (the water-powered armillary sphere and celestial globe). He invested much effort in modeling and reconstructing the instrument and designing a Virtual exhibition about it. The candidate is now looking forward to get inspired further by scholars from other nations or civilizations. So far, the modeling of mechanisms, which is highly relevant in industrial archaeology in general, has been missing from these projects. The proposal would help to add this aspect to the work that is ongoing in their realm.

Background

In recent years, multidisciplinary approaches combining Virtual Reality, Archaeology, and Cultural Heritage have become increasingly important. The rapid development of 3D capturing technology provides modern archeologists and other scholars with tools to realistically model and visualize archeological sites in 3D. In the same way as the general public is getting used to more and more realistic visualizations of virtual worlds through games, movies and TV, the demand for such 3D models of archeological

sites is growing. These models are not only used for edutainment and site marketing, they also provide a basis for dissemination and scientific discussion about reconstruction hypotheses. In Europe, many research projects were sponsored to support such interdisciplinary research, like EPOCH, 3D-COFORM, ViHAP3D, 3D-Murale, and so on.

This project related to astronomical clocks is an application research integrating computer graphics and 3D acquisition, mechanics, archaeology,arts and history. An astronomical clock is a clock with special mechanisms and dials to display astronomical information, such as the relative positions of the sun, moon, zodiacal constellations, and sometimes major planets. Europe has quite a long tradition of making astronomical clocks and many of them are located in the main cities of Europe, such as Strasbourg, Prague, Copenhagen, but also Olomouc or Lier, and are mostly operating properly.

Similarly, China also has a long tradition of making astronomical clocks. The water-powered armillary sphere and celestial globe were built around the year 1088 AD (Chinese ancient Astronomical Clock-tower built by Su Song and his collaborators in the Northern Song Dynasty). It was undoubtedly one of the pinnacles of this craft and art. In the past two years, the candidate's major research focused on reconstructing the water-powered armillary sphere and celestial globe and on putting forward a novel reconstruction design different from other designs found in historical archives.

In 1964, J.H. Combridge published an article in Nature, the Chinese water-balance escapement, where he defined the Great wheel as an escapement mechanism and made a 1:6 scale model of the escapement. Beside, the book written by Dr. Joseph Needham- the Heavenly Clockwork, also introduced the Chinese ancient Astronomical Clock-tower according to the statement of J.H Combridge. The water-driven astronomical clock tower was a wooden building 12meters in height. On the top platform was an armillary placed in a chamber with a removable roof which resembled the dome of a modern observatory. The armillary was connected through gears with the driving machinery of the whole installation, which enabled it to follow the diurnal motion of the natural celestial sphere. When the observer aimed the sighting tube at the sun, the mechanical motion of the armillary would keep the sun in the visual field for a fairly long time. This device was working based on a water-powered mechanical clock with an escapement regulator. Literary records are available for this invention, but unfortunately surviving hardware is lacking. However, several reconstruction designs were put forward in the past century.

Objectives

This project, through its interdisciplinary research method, will explore an effective way for the reconstruction of tangible cultural heritage. In particular, our research will focus on the structure and mechanisms of the astronomical clock towers in China and Europe. We will employ modeling tools such as 3DSMAX and SOLIDWORKS to model astronomical clock towers, and we will make use of JavaScript and VRML technology to control the display. Then, we will use the ADAMS software to analyze the kinetic parameters and rotation periods. Through the study of the related literature, a report will be produced that highlights the differences and similarities with other astronomical clock towers.

In general, this project has three main purposes:

(1) Exploring a practical and efficient method for the reconstruction of Cultural Heritage where the objects consist of mechanical parts with relative motions
(2) Comparing the difference of Astronomical Clock towers between Europe and China in terms of mechanics, astronomy, arts,and history
(3) Building a virtual imaging space showing the main Astronomical Clock towers from China and Europe, also explaining their structures and operating principles

Research method

The main methods and resources planned for use in this project could be summarized as follows:

1. Historical Archives,
2. Design specifications,
3. Generalized kinematic chains,
4. Specialized chains,
5. 3D detailed Modeling,
6. Comparative study,
7. Programming

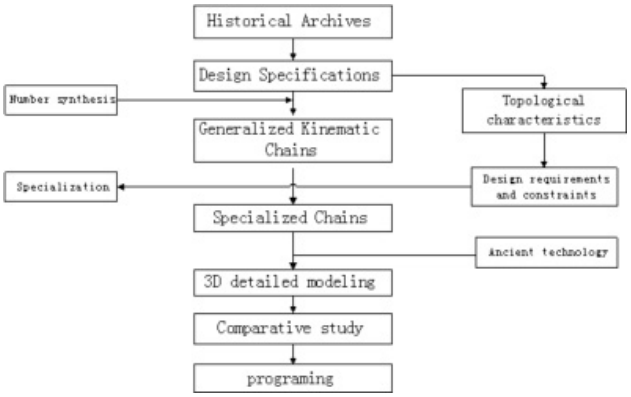


Fig 1. research method

Illustration about Dynamic and Mechanical Analysis

There are a lot of technologies and commercial tools available today to create realistic 3D models of cultural heritage. When producing a dynamics and kinematics model of an Astronomical Clock, the needs go beyond capturing static 3D shapes. We have to produce a geometric model of the parts and to also formulate the possible relative motions, as well as their restrictions. We have already started to make a model with the computer software SOLIDWORKS, including all the main parts. The consistency of the parts will be checked via their assembly into virtual mechanisms, e.g. checking the number of teeth of gears, the diameters of ratchet-wheels, etc. Then, we can also get experimental results by amending the parameters of the simulating model. After we get our 3D model, we can input it to the platform of ADAMS by a kind of Para solid format. Then we can suppose some properties of parts for which the real, original data are lacking, like their masses, original positions, velocities, moving directions and reestablished loads. This will be done through simulations with ADAMS, which can link up the components into complete mechanical systems. This software is also capable of limiting the relative motions between different components by a series of restrictions. Fortunately, a number of parameters for the original parts can still be deduced today. The most important contribution in this however, is that we will endeavor to embed the above analysis within the framework of procedural modeling, which we explain in a bit more detail further below. The aforementioned tools will be used to generate the 3D models of individual parts. Also in a procedural modeling context, several ready-made 3D models (the so-called leaf nodes or `assets') are included directly from libraries, possibly parameterized so that a number of properties can still be modified. The above tools will be used to generate the necessary assets (i.e. some of the solid components of which the entire structure has been composed).

After we get all useful parameters of the gear system, we will perform experiments to calculate several other parameters like moments of inertia, transmitting power, etc. And then we can deduce its motion and effect on the Astronomical Clock design. At last, we will be able to explain the full 3D model. 'Full model' is relative here, as we will discard non-essential parts, such as sculptures, puppets for reporting time, etc. And we predigested some complicated parts as a simple rigid body.

The next pages show some examples of parts that have already been modeled, and also of manual compositions into assemblies. This type of tedious, manual composition should be largely replaced by procedural modeling. As can be seen, some of the parts are quite ornamental. Such pieces definitely have to be modeled separately, as the assets that the procedural models are based upon. The procedural modeling aspect will mainly focus on the underlying mechanisms of these clock towers.

*Several pictures of my recent research: (unpublished)

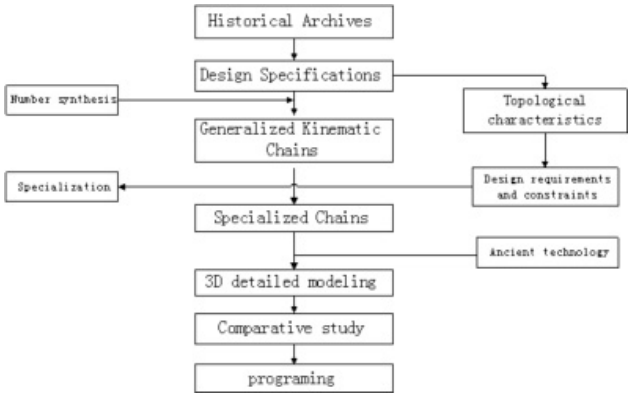


Fig. 1:
research method

Fig. 2:
internal model of the water-powered armillary sphere and celestial globe

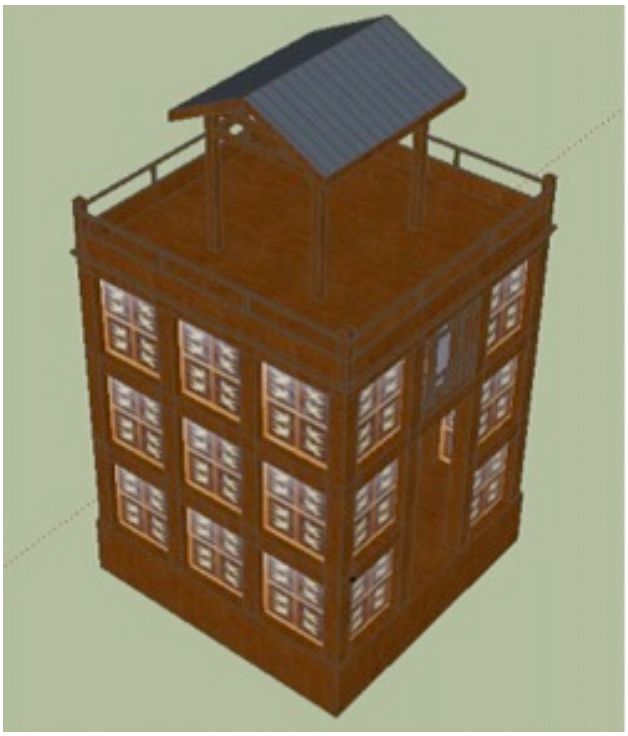


Fig. 3:
outside model of the water-powered armillary sphere and celestial globe

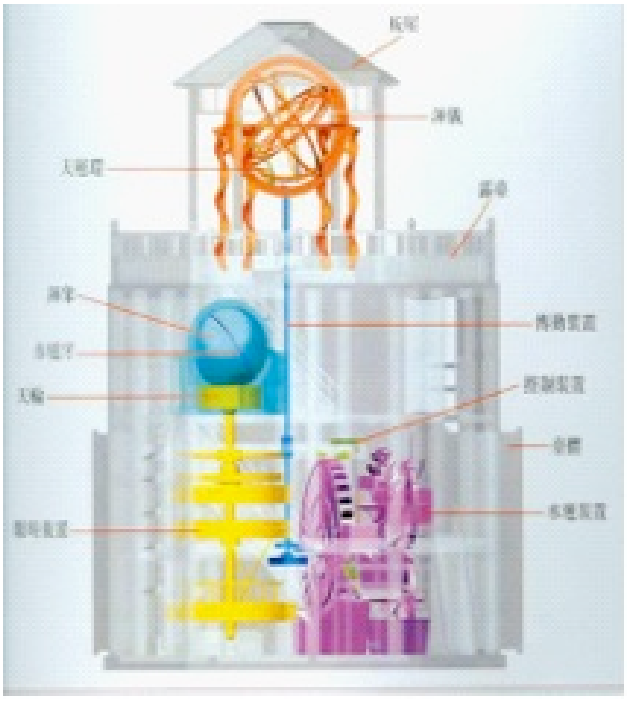


Fig. 4:
internal structure of the water-powered armillary sphere and celestial globe

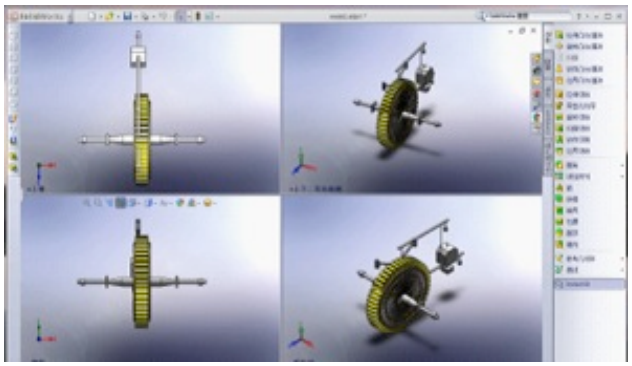


Fig. 5:
model of the Great wheel

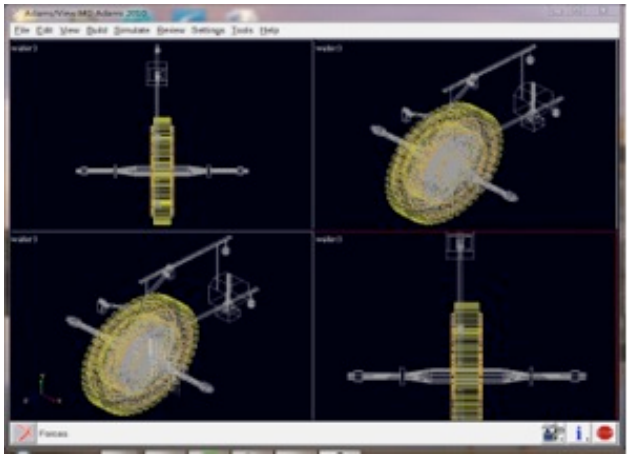


Fig. 6:
Dynamic simulation of the Great wheel based on ADAMS

Creating Port: Research Commons

Lindblad, Purdom
purdom6@vt.edu
Virginia Tech University, United States of America

Pencek, Bruce
bpencek@vt.edu
Virginia Tech University, United States of America

Brooks, Edwin

ebrooks@vt.edu
Virginia Tech University, United States of America

Speer, Julie

jspeer@vt.edu
Virginia Tech University, United States of America

Virginia Tech Libraries recently created a digital research lab and collaboration space, Port: Research Commons. Experimental and crossing library departments, Port offers geospatial, data visualization, text analysis, data management, and graphic design software. Additionally, Port supports digital research for a variety of disciplines at Virginia Tech through one-to-one consultations with library and campus experts, small workshops, and speakers' series.

Within a semester Port has become a hub of intellectual inquiry spanning from graphic design and statistics to data management. This poster highlights the dynamic digital humanities activities and collaborations happening in the space and the process of launching Port: Research Commons.

The poster will detail early library efforts to support digital research and highlight how earlier initiatives laid the foundation for the successful beginnings of Port. It will outline current programming and projects, focusing on the collaborations vital to Port's success.

Situating Port within the growing landscape of library research commons, the poster explores the library, specifically research commons, as hub for the myriad of disciplines working on digital research and scholarship.

In addition to serving as a critical space for a broad range of learning and eResearch related needs, Port is a gathering space for faculty, graduate students, and librarians working within digital humanities. Port has become a space used in unique ways to support digital humanities initiatives stemming from the Library's new Center for Digital Research and Scholarship, while also serving as a space incubating collaborations between other campus groups and library programs, including the library's new learning initiatives. Library outreach efforts to departments around campus are supported by software and inviting collaboration space.

The Center for Digital Research and Scholarship, in collaboration with Center for Applied Technologies in the Humanities, CATH, is working on moving the TEI project, Lord Byron and His Times, to the library for hosting. Bringing together the college librarian for humanities and digital humanities, library ITS, and Dr. David Radcliff, Port is the meeting space for assessing the current TEI information architecture and moving it to a more scalable architecture. In this poster, we'll highlight the important use of this space for technology explorations related to CATH's Lord Bryon project as well as use of the space for other digital humanities projects where the library is involved as collaborator or project lead.

Our poster provides the history, current projects, such as Lord Bryon and His Times and the DH Working Group, and explores library research commons as vital support for digital research on campus.

Rebuilding Civil War Washington

Lorang, Elizabeth M.

liz.lorang@gmail.com
University of Nebraska-Lincoln, United States of America

Dalziel, Karin

kdalziel2@unl.edu
University of Nebraska-Lincoln, United States of America

In April 2012, the Center for Digital Research in the Humanities at the University of Nebraska–Lincoln launched version 2.0 of *Civil War Washington* (civilwardc.org). This poster presentation will document the evolution of *Civil War Washington* (CWW) from version 1.0 to version 2.0, including the technologies used, the changing aims of the project, and the intellectual rationale underlying these decisions. Changes made for version 2.0 included a revamping of the overall design and underlying framework of the site and the development of three major sections, Texts, Data, and Maps. This poster will emphasize the structure and design of the site, as well as the Texts section.

Although the site maintained a similar but updated visual look from version 1.0 to version 2.0, nearly everything on the back end was rewritten, including the HTML and CSS. The technology in the intervening three years had changed so much that we were able to streamline the HTML and employ CSS3 elements for features such as drop shadows and embedded fonts. At the same time, we chose to abandon a CSS framework used in the first version of the site because the framework made it too easy to accidentally break the layout. We are now using heavily commented descriptive classes on HTML elements.

The Texts section of *CWW* changed significantly from the first version of the site to the second. Most visibly, the amount of available material expanded. Originally, *CWW* included six HTML pages of cases from the *Medical and Surgical History of the War of the Rebellion* and a number of issues of a single hospital newspaper. With the relaunch of the site, we added 200 petitions filed in response to the DC Emancipation Act of 1862 (more petitions are added on a regular basis), separated the medical cases into their

own files (more than 1,400 and counting), and posted nearly complete runs of three newspapers. Each of these different types of texts required specific encoding choices that were rooted in our scholarly commitments and were necessarily influenced by more pragmatic concerns.

Throughout the history of the project, all documents represented in the Texts section have been encoded in TEI. Working toward version 2.0 of the site, we determined that the initial encoding for the medical cases—in which all cases from a particular volume of the *Medical and Surgical History* were represented in a single TEI file, transformed to HTML — did not adequately model our understanding of the cases or how we thought they most usefully told the story of Civil War DC. We used XSLT to separate all of the cases into individual files, drawing on existing metadata to generate new TEI headers and keywords for the cases. We also modified our approach to encoding personal, place, and organization names and references to dates. For the petitions, we recognized that the key features to be encoded were form and handwritten text; personal, place, and organizational names; and the structure of the documents. In addition, part of our editorial work was rejoining documents separated into two record groups at the U.S. National Archives. Therefore, our encoding practices needed to be able to accurately describe and reconstruct this quality of the documents.

One goal of the revamped Texts section was to make use of this new deep encoding. For version 2.0 of the site, we made use of the CSS element @font-face, which allowed for considerable typographic choices. This change required further discussion and raised potential problems, however, as many free fonts support only limited character sets, and many of the project documents included Unicode characters such as ligatures. In addition, the encoding and new site infrastructure allowed us to link cases to one another by drawing on the encoded keywords and utilizing SOLR as the search engine. Likewise, we utilized the encoding to visually mark the handwritten text of petitions in the browser. Users have the option to toggle on the highlighting of handwritten text, thereby calling attention to the parts of documents unique across petitions. These examples illustrate just a few the many problems confronted and decisions made during the redesign.

The poster will include examples of the old code and content and the new code and content, noting the differences between them and providing screenshots that demonstrate the evolution of the site. In addition, the poster will document the technical and humanistic reasons for the changes, as part of the larger rationale of *Civil War Washington*.

Elwood Redux: Introducing the Elwood Transcription/Text Encoding Modules as well as a Newly-revised, Browser-independent Version of the Elwood Viewer

Lyman, Eugene W
eugene.lyman@gmail.com
Independent Scholar, United States of America

The ELWOOD VIEWER was created ten years ago as a tool to display the edited text and manuscript images associated with documentary editions produced by the Piers Plowman Electronic Archive (PPEA). As originally written, it contained powerful regular expression search capabilities as well as display formats to enable sophisticated analysis of the texts that it presented. ELWOOD has met with very positive reception during conference presentations and with glowing comment in reviews of individual editions produced by the PPEA, including that of one reviewer (George Shuffelton, *Speculum,* (85) 2010, 285), who urged that the ELWOOD VIEWER represented "a major step forward in the delivery of digital editions" that "ought to be a model for future programs of this kind."

Although originally conceived as a general interface for electronic scholarly editions, to date ELWOOD has been available only as a dedicated viewer packaged in the PPEA's CD-ROM editions of the B-version of *Piers Plowman.* As such, it has not been available for use by other electronic editions . The poster proposed in this abstract highlights recent steps taken to greatly broaden ELWOOD's potential usefulness in the display of scholarly editions available over the Internet. It will also introduce its audience to a new software module developed specifically to enable editors, as well as individuals seeking relatively straightforward means to transcribe and encode manuscript texts for use in their personal scholarship, to have access to the powerful analytical tools contained in the Elwood Viewer. The poster will also introduce a version of Elwood designed to run under all major Internet browsers (the original version could only be run under Microsoft's Internet Explorer)and

enhanced with improvements that draw upon the rich image-handling capacities made possible by HTML 5.

Features that have been supported by the ELWOOD VIEWER for a number of years, including its tight linkage of manuscript image and edited text, are presented (with illustrative figures) in Lyman (2004). The added features to be demonstrated in this poster session include: the enlargement of the software's function from that of the display of single manuscripts to that of serving as an overall archive viewer possessing the capacity to compare and move quickly between edited texts of different witnesses of a work, enhanced text and markup searching capacities along with the capability to export search results for further analysis, frequency counts, organized alphabetically or by descending order of frequency, of searched works or phrases, as well as other enhancements geared to assist readers in their navigation of edited text and manuscript images.

References

Duggan, H. and E. Lyman (2004). A Progress Report on The Piers Plowman Electronic Archive, *Digital Medievalist.* http://www.digitalmedievalist.org/journal/1.1/duggan/#dm.1.1.duggan.0200

Foys, M. (2012). forthcoming. The Piers Plowman Electronic Archive and the Process of Durable Mutation, : (review essay). *Yearbook of LanglandStudies* 26. http://tinyurl.com/kkmn8ap

Langland, W. (2005). The Piers Plowman Electronic Archive, Vol 5: British Library Ms Additional 35287 (M). Woodbridge [England]: Medieval Academy of America and Boydell and Brewer.

Langland, W. (2011). The Piers Plowman Electronic Archive, Vol 7: Bodleian Library Ms Rawlinson Poetry 38 (R). Ed. Robert Adams, Woodbridge [England]: Medieval Academy of America and Boydell and Brewer.

Langland, W. (2004). The Piers Plowman Electronic Archive, Vol. 3: Ms Oriel College, Oxford 79 (0). Ed. Katherine Heinrichs. Woodbridge [England]: Medieval Academy of America and Boydell and Brewer.

Langland, W. (2004). The Piers Plowman Electronic Archive, Vol. 4: Ms Laud Misc. 581 Bodleian Library S.C. 987 (L). Duggan, H. N., and R. Hanna III (eds.). Woodbridge [England]: Medieval Academy of America and Boydell and Brewer.

Langland, W. (2009). The Piers Plowman Electronic Archive, Vol. 6: Huntington Library Ms Hm 128 (Hm). Calabrese, M., H. N. Duggan and T. Turville-Petre (eds.). Woodbridge [England]: Medieval Academy of America and Boydell and Brewer.

Lyman, E. (2009). *Assistive Potencies: Reconfiguring the Scholarly Edition in a Digital Environment,* Dissertation. University of Virginia.

Lyman, E. (2004). In pursuit of radiance: Report on an interface developed for the *Piers Plowman Electronic Archive,* The Face of Text, Computer Assisted Text Analysis in the Humanities, Canadian Symposium on Text Analysis, November 19-21, McMaster University, 65-70. http://tapor1.mcmaster.ca/~faceoftext/FOTfinal.pdf

Lyman, E. (2008). 'May the text rise up to meet you:' New ways of reading old manuscripts. *Digital Humanities Quarterly*. 3 (3) http://www.digitalhumanities.org/dhq/vol/3/3/000058/000058.html http://www.digitalhumanities.org/dhq/vol/3/3/000058/resources/images/figure01.pdf

Shuffelton, G. (2010). Review of Vol 6 of the Piers Plowman Archive, *Speculum*. (85): 984-986.

WordSeer: An Integrated Environment for Literary Text Analysis

Muralidharan, Aditi
aditi@cs.berkeley.edu
UC Berkeley, United States of America

Summary

I will present WordSeer, an environment for literary text analysis. Literature study is a form of sensemaking: a cycle of reading, interpretation, exploration, and understanding. While there is abundant technological support for reading text in new ways through visualizations and algorithms, the other parts of the cycle — exploration and understanding — have been relatively neglected. WordSeer integrates tools for algorithmic text processing with interaction techniques that support the interpretive, exploratory and note-taking aspects of scholarship. Its design has been shaped by individual case studies with literature scholars as well as a semester-long field trial with a class of undergraduate Shakespeare students.

Proposal

To date, text analysis systems for humanities scholars have focused on aiding interpretation (Clement 2008;

Fekete, et al. 2000; J. Guldi, et al. 2012; X. Llorà, et al. 2008; C. Plaisant, et al. 2006; G. Rockwell, et al. 2010; R. Vuillemot, et al. 2009). First, they apply some form of natural language processing to extract aggregate statistics about word usage, topics, named entities, and parts of speech. Second, they display the extracted information with visualizations like word clouds, node-and-link diagrams, and lists of word contexts. Such systems make patterns of style, form, and theme visible, and interpretable by people.

However, literature study is a form of sensemaking (P. Pirolli, et al. 2005): a cycle of reading, interpretation, exploration and understanding. As useful as they are, current digital humanities text analysis systems leave the exploration and understanding part of the cycle unsupported.

The WordSeer project (A. Muralidhara, et al. 2011) is an effort to create a *sensemaking* environment for literature and language study. Like other systems for the humanities, it has search and visualization capabilities, but it also supports sensemaking activities like collecting and reorganizing information, exploring related words, finding frequent phrases and similar passages of text, and annotating, collecting and tagging items. The system has been under development since 2010. Recently, it was used to produce successful analyses of Shakespeare's plays (A. Muralidharan, et al. 2012) and North American slave narratives (A. Muralidharan 2012).

To uncover areas for improvement, WordSeer was field-tested at the University of Calgary in the Spring 2012 semester. Students in the undergraduate Shakespeare class 'Hamlet in the Humanities Lab' (M. Ullyot, et al. 2012) spent a few weeks becoming familiar with WordSeer along with four other computational text analysis tools. Then, during the rest of the semester, they used the tools to analyze a topic of their choice within an act of Hamlet. The students recorded their experiences through weekly posts on the class blog (http://engl203.ucalgaryblogs.ca/).

An analysis of their posts revealed four common ways of using digital tools that were not supported well:

1. Comparing two or more visualizations side-by-side or referring to multiple tools simultaneously
2. Narrowing down analyses by metadata, such as (in Shakespeare) a particular speaker, act, or scene.
3. Investigating a group of words together.
4. Getting ideas for a new search or analysis based on the results of a previous one.

WordSeer had limited, roundabout support for these activities. For example, Activity 1, comparisons: it was technically possible to compare two visualizations — but only by opening a separate browser window, navigating to the WordSeer Shakespeare website, and re-typing the search parameters for the second visualization. Activities 3 and 4: investigating groups of words, or performing new searches based on previous ones, required manually typing in long search queries, and Activity 2 was entirely impossible.

At this conference I will demonstrate a new WordSeer, completely redesigned with the above activities in mind. Instead of a separate web page per visualization, the application now mimics a desktop environment, with different visualizations opening up in "windows". In the following figures (Figure 1-Figure 4), I briefly explain how the new tool supports the above activities.

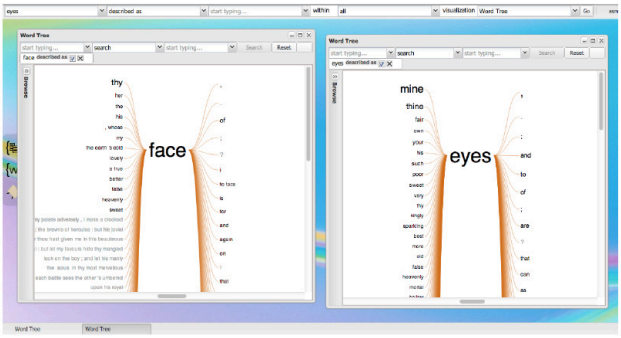


Figure 1
This figure shows WordSeer's new desktop environment, featuring a top bar for queries, a sidebar for collections, and multiple windows. Activity 1, comparison and reference, is much easier because the top bar preserves search parameters. For example, comparing the word tree for "face" with that of "eyes" above only requires changing a single word between queries.

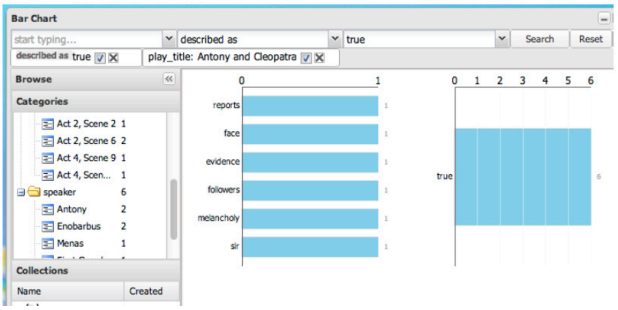


Figure 2
Metadata filters can be used to restrict analyses to relevant subsets, directly supporting Activity 2. For example, this figure shows how, to find all words described as "tru"e within Antony and Cleopatra, or to further restrict the search to individual speakers, users simply have to select the relevant categories from the browsing menu.

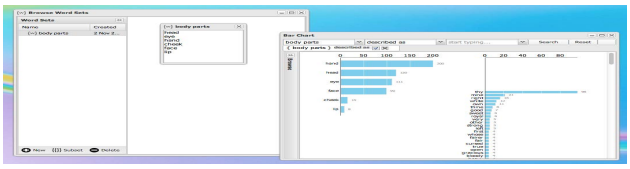


Figure 3
For activity 3 (investigating groups of words together), WordSeer has a new word sets feature for creating groups of words. These sets can be used as search queries, and updating the set with new words automatically reflects itself in the new query. For example, this figure shows how word set called "body parts" containing "head, eye, hand, cheek, face, lip" can be used as a grammatical search query (for "body parts" described as ___).

References

Clement, T. E. (2008). 'A thing not beginning and not ending': using digital tools to distant-read Gertrude Stein's The Making of Americans, *Literary and linguistic computing,* 23(3): 361.

Fekete J.-D., and N. Dufournaud (2000). Compus: visualization and analysis of structured documents for understanding social life in the 16th century, in *Proceedings of the fifth ACM conference on Digital libraries,* New York, NY, USA, 47–55.

Guldi, J., and C. Johnson-Roberson (2012). Paper Machines: A Tool for Analyzing Large-Scale Digital Corpora, College Park, MD, 03-Nov-2012.

Llorà, X., B. Ács, L. S. Auvil, B. Capitanu, M. E. Welge, and D. E. Goldberg (2008). Meandre: Semantic-driven data-intensive flows in the clouds, in *Fourth IEEE International Conference on eScience,* 238–245.

Plaisant, C., J. Rose, B. Yu, L. Auvil, M. G. Kirschenbaum, M. N. Smith, T. Clement, and G. Lord (2006). Exploring erotics in Emily Dickinson's correspondence with text mining and visual interfaces, in *Proceedings of the 6th ACM/IEEE-CS joint conference on Digital libraries,* Chapel Hill, NC, USA, 141–150.

Rockwell, G., S. G. Sinclair, S. Ruecker, and P. Organisciak (2010). Ubiquitous Text Analysis, *Poetess Archive Journal,* 2(1).

Vuillemot, R., T. Clement, C. Plaisant, and A. Kumar (2009). What's being said near 'Martha'? Exploring name entities in literary text collections, in *Visual Analytics Science and Technology* 107–114.

Pirolli, P. and S. Card (2005). The sensemaking process and leverage points for analyst technology as identified through cognitive task analysis, in *Proceedings of International Conference on Intelligence Analysis,* vol. 2005, 2–4.

Muralidharan, A., and M. A. Hearst (2011). Supporting Exploratory Text Analysis Literature Study, *Literary and linguistic computing,* Digital Humanities Conference Issue (forthcoming), Submitted Dec. 2011.

Muralidharan, A., and M. A. Hearst (2012). A sensemaking environment for literature study, in *CHI '12 Extended Abstracts on Human Factors in Computing Systems,* New York, NY, USA, 1955–1960.

Muralidharan, A. (2012). Using Digital Tools to Explore Narrative Conventions in the North American Antebellum Slave Narratives, Seattle, WA, USA, 07-Jan-2012.

Ullyot, M. (2012). Hamlet in the Humanities Lab | English 203, Winter 2012 | Michael Ullyot, *Hamlet in the Humanities Lab English 203, Winter 2012,* 13-Jan-2012. http://ullyot.ucalgaryblogs.ca/teaching/hamlet/. Accessed: 13-Jan-2012.

A Case Study of Integration of Services and Resources on a Web Service

Nagasaki, Kiyonori
nagasaki@dhii.jp
International Institute for Digital Humanities, Japan

Tomabechi, Toru
tomabechi@dhii.jp
International Institute for Digital Humanities, Japan

Muller, A. Charles
acmuller@l.u-tokyo.ac.jp
University of Tokyo

Shimoda, Masahiro
shimoda@l.u-tokyo.ac.jp
University of Tokyo

The SAT Daizōkyō text database committee has released a new version of its integrated Web service that has been offering a series of Buddhist scriptures [1] since June 2012 in order to provide more convenient and powerful digital resources for the field of Buddhist studies. Since August, the number of accesses is over 200,000 per month on an average (not counting accesses by Googlebot and several

other automated indexerss). In this paper, we will describe the integration of various services and resources in this Web service.

First of all, we will explain the function of retrieving and displaying of English translations which has been published as a series of Buddhist scriptures rendered into English by the Bukkyō Dendō Kyō kai (BDK) [2] . This function is an implementation of a result that was produced by 18 young researchers using a web collaboration system described in Nagasaki (2011). The collaboration system allows users to easily link a fragment of a text to another fragment of a resource. As a result, a parallel corpus including over 29,000 pairs of English and classical Chinese Buddhist texts has been published on the web service so that users can easily check the English translation of words or phrases in various contexts by dragging a sequence of the text. Moreover, the parallel texts can be shown on the text which was translated sentence-by-sentence. Thus, not only is the English translation provided, but also an interpretation is given for each division of the classical Chinese texts that are not clearly separated by any kind of punctuation. It also continues to provide the function of retrieving terms contained in the Digital Dictionary of Buddhism (DDB) [3] including over 58,000 entries so that users can easily look up the English renderings of terms. We also continue to provide the function of translating an English term into a Chinese term in the input form of keyword search of the whole text, whereby users can search the Chinese text by using an English term.

The Web service has added a function of retrieving two bibliographical databases: SARDS3 [4] and CiNii [5] . SARDS3 includes over 60,000 bibliographical records of western books and papers published in the field of Indology. The SAT Web service provides automatic translation from selected Chinese words or phrases to English or Sanskrit terms using DDB as a support function of retrieval so that user can easily find related western secondary resources. In the previous version, a function of a Web API of the CiNii (the largest academic bibliographical database in Japan, managed by the National Institute of Informatics) was included via Indian and Buddhist Studies Treatise Database (INBUDS) in order to indicate whether or not a PDF file is available on the Web. However, in the new version, the additional function of retrieving the CiNii itself in order to easily search related resources in other fields such as history and linguistics was added according to user requests.

All digital facsimiles of the Taishō Shinshū Daizō kyō — approximately 80,000 files — are scanned in 600 DPI and made available on the web site. When a user clicks an image button embedded in the lines of the text, an image is displayed inside a window on the left in the page corresponding with the paragraph and lines of the text. The image can be zoomed in and out by use of slider in order to examine its details, which allows users to confirm whether or not the encoded texts are accurate, and whether or not the slight differences in ideographs were unified in the process of text encoding.

In the case of classical Chinese, the distinction and unification of ideographs has been an important topic which often provokes heated discussion in East Asia[6] . As the Taishō Shinshū Daizō kyō is one of the important materials for such discussion, we provide a function to research information about each ideographic character by linking to several character databases such as CHISE, Unihan, and so on.

Finally, we will explain our method of implementation of the above functions. Most of them are implemented by Linux, PostgreSQL, Apache, PHP, jQuery, and JavaScript. Most of the interfaces are implemented by jQuery and jQuery-ui, but partly by plain JavaScript because of insufficient functionality in the former. The above functions are provided in each window of jQuery-ui, which can be arbitrarily opened and closed so that users can use only necessary functions. We hope we can have the opportunity to discuss any aspects of this new Web service with those who visit our booth.

References

Muller, A. C., and K. Nagasaki (2012). Request to add Unencoded Characters in the Taishō Shinshū Daizō kyō (Taishō edited series of Buddhist scriptures). http://appsrv.cse.cuhk.edu.hk/~irg/irg/irg38/IRGN1858_ProposalSATgaiji.pdf (Accessed 28 Oct 2012).

Nagasaki, K., T. Tomabechi, and M. Shimoda (2011). Toward a Digital Research Environment for Buddhist Studies. *Digital Humanities 2011*. 342-343.

Notes

1. http://21dzk.l.u-tokyo.ac.jp/SAT/ddb-bdk-sat2.php Accessed 28 Oct 2012.
2. http://www.bdkamerica.org/default.aspx?MPID=81 Accessed 28 Oct 2012.
3. http://www.buddhism-dict.net/ddb/ Accessed 28 Oct 2012.
4. http://www.indologie.uni-halle.de/sards/ Accessed 28 Oct 2012.
5. http://ci.nii.ac.jp/en Accessed 28 Oct 2012.
6. Regarding unencoded approximately over 6,000 characters in the text, we are addressing to encode them in the Universal Character set. See Muller (2012).

Interfaces for Crowdsourcing Interpretation

Nally, Gwendolyn
egn9b@virginia.edu
University of Virginia Library, United States of America

Peck, Chris
cp3ee@virginia.edu
University of Virginia Library, United States of American

Lin, Shane
ssl2ab@virginia.edu
University of Virginia Library, United States of America

Márquez, Cecilia
cm2ug@virginia.edu
University of Virginia Library, United States of America

Maiers, Claire
cdm6zf@virginia.edu
University of Virginia Library, United States of America

Walsh, Brandon
bmw9t@virginia.edu
University of Virginia Library, United States of America

Boggs, Jeremy
dw04@aub.edu.lb
University of Virginia Library, United States of America

Praxis Program Team
praxis2012@collab.itc.virginia.edu
University of Virginia Library, United States of America

Our research will detail a number of approaches to crowdsourcing interpretation — especially as these approaches relate to the ongoing development and design of *Prism,* a tool that facilitates crowdsourced interpretation of texts. We take up the challenge detailed by Ramsay and Rockwell (2012) that the activity of building provides affordances as rich and informed as writing, and that it is important to to be aware of the nature and quality of the intervention that happens through building. Drucker (2009) and Ruecker et. al. (2011) demonstrate the importance of speculative prototyping as a way to explore humanities questions and make arguments through prototypes. In that spirit, our research will inform the creation of several interfaces that address problems related to the individual and crowdsourced interpretation.

Background

In 2011, the Praxis team at the University of Virginia created *Prism* as digital realization of the "Patacritical Demon" imagined by Drucker (2009), McGann (2004), and Nowviskie (2012). In its current form, *Prism* allows multiple users to highlight a text based on certain predetermined categories. The tool then creates an aggregate visualization of individual responses. In this way *Prism* diverges from current utilizations of crowdsourcing; where crowdsourced projects have traditionally asked users to compile data or do other mechanistic tasks, *Prism* asks individuals to mark up a text with categories of meaning, to discern trends in the way a larger group of users reads the text. Although *Prism* promises to bring individual experience into the fold, Bethany Nowiskie (2012) notes that *Prism* "is not a device for rich, individual exegesis," and the usefulness of *Prism* lies in the overlapping and visualizing of all contributors, "generating spectra of similarity and difference."

Existing Approaches to Crowdsourcing Interpretation

Owens (2012) distinguishes between two approaches that are commonly lumped under the heading of crowdsourcing: "human computation" and "the wisdom of crowds." Human computation projects ask participants to solve problems for which computational solutions are comparatively expensive to develop or perform. Such problems include transcription (Transcribing Bentham, Old Weather, reCAPTCHA), protein folding (fold.it), and image metadata tagging (ESP Game). Crowd-wisdom projects, on the other hand, engage participants in open-ended socially-negotiated tasks that may go further than processing information and actually create new knowledge. (This is the mode of Wikipedia or any website with comment or discussion forum functionality.)

Directions for Research

We have begun to consider other ways to crowdsource interpretation, especially approaches that attend more

closely to individual responses. In particular, we have identified two areas for exploration:

Computation and Crowd Wisdom — Owens (2012) suggests that both existing modes of crowdsourcing are worth designing for and can work in tandem for Digital Humanities projects (Galaxy Zoo). In that spirit, *Prism* presents the user with a task that is in both constrained in such a way that it produces information and somewhat open-ended and socially-negotiated through the user's engagement with a text.

Wisdom of the Individual — Crowdsourcing interpretation might offer an increased focus on individuals as a part of the collaborative process. Preserving the marks of each participant would better respect the human element at the core of crowdsourcing. This feature would also make *Prism* a more powerful pedagogical tool, as an instructor could identify an individual student's remarks for generating discussion. This would also be useful for the social sciences, which are inherently concerned with the individual as a member (and perhaps representative) of a particular group.

Future design and development of *Prism* must account for the many potential roles of individuals in crowdsourcing interpretation. For example, in order to better serve projects concerned with the wisdom of particular individuals *Prism* would need to store user markings and interpretations, in a way that they are extractable and easily viewed in relation to the markings of the crowd. Similarly, in order to better facilitate social science research *Prism* might include a way to separate out user interpretations according to demographic information (such as class, gender, age, etc.) or according to specific user responses. The poster will detail how the the design and development of *Prism* has been and continues to be influenced by the different roles individuals might play within the crowd.

References

Drucker, J. (2009). *SpecLab: Digital Aesthetics and Projects in Speculative Computing.* Chicago: University of Chicago Press, 2009.

Ford, P. (2011). The Web Is a Customer Service Medium. http://www.ftrain.com/wwic.html. (accessed 6 January 2011).

Galey, A., S. Ruecker, and the INKE team (2010). How a Prototype Argues. *Literary and Linguistic Computing* 25(4): 405–24.

McGann, J. (2001). *Radiant Textuality: Literature After the World Wide Web.* New York: Palgrave MacMillan.

McGann, J. (2004). What is a Text? In Schreibman, S., R. Siemens, and J. Unsworth (eds). *A Companion to Digital Humanities.* Oxford: Blackwell. http://digitalhumanities.org/companion/view?docId=blackwell/9781405103213/9781405103213.xml&chunk.id=ss1-3-4

Meister, J. C. (2012). Crowd Sourcing 'True Meaning': A Collaborative Markup Approach to Textual Interpretation. In Deegan, M. and W. McCarthy (eds). *Collaborative Research in the Digital Humanities.* Ashgate. 105–122.

Nowviskie, B. (2012). A Digital Boot Camp for Grad Students in the Humanities. *Chronicle of Higher Education.* http://chronicle.com/article/A-Digital-Boot-Camp-for-Grad/131665/ (29 April 2012).

Owens, T. Human Computation and Wisdom of Crowds in Cultural Heritage. http://www.trevorowens.org/2012/06/human-computation-and-wisdom-of-crow ds-in-cultural-heritage/

Ramsay, S., and G. Rockwell (2012). Developing Things: Notes toward an Epistemology of Building in the Digital Humanities. In Gold, M. K. (ed). *Debates in the Digital Humanities.* Minnesota: University of Minnesota Press.

Ruecker, S., M. Radzikowska, and S. Sinclair (2011). *Visual Interface Design for Digital Cultural Heritage: A Guide to Rich-Prospect Browsing.* Burlington, VT: Ashgate.

Surowiecki, J. (2004). *The Wisdom of Crowds: Why the Many Are Smarter Than the Few and How Collective Wisdom Shapes Business, Economies, Societies, and Nations.* New York: Doubleday.

The Textual Communities Transcription workspace: a poster and demonstration

Nelson, Brent
brent.nelson@usask.ca
University of Saskatchewan, Canada

Klaassen, Frank
frank.klaassen@usask.ca
University of Saskatchewan, Canada

Robinson, Peter
peter.robinson@usask.ca

University of Saskatchewan, Canada

This poster will present the beta version of the Textual Communities transcription tool, describing its underlying principles, its innovative structure, and its functionality.

Study of literary works that exist in many different forms is one of the most important and difficult tasks in the humanities. The number of forms a work may have —eighty-four fifteenth-century manuscripts and printed texts of Chaucer's *Canterbury Tales,* more than eight hundred manuscripts of Dante's *Comedia,* and five thousand manuscripts of the Greek New Testament — is both testimony to their significance and a challenge to scholars. In order to understand these texts and how they relate, we have to discover as much as we can of how they came to be written and disseminated. Only then can we seek to establish how they might best be read and prepare texts (in the form of scholarly editions) for scholars to use.

The work of building archives of primary materials for this kind of work is daunting and prohibitive in the old lone scholar method. At the same time that scholars working on large editing projects have opted for a team approach, several projects internationally have demonstrated the tantalizing possibilities of crowd-sourcing for processing large amounts of textual data. The challenge it to coordinate this work in a way that produces high-quality, useful results.

The work of building archives of primary materials for this kind of work is daunting and prohibitive in the old lone scholar method. At the same time that scholars working on large editing projects have opted for a team approach, several projects internationally have demonstrated the tantalizing possibilities of crowd-sourcing for processing large amounts of textual data. The challenge it to coordinate this work in a way that produces high-quality, useful results.

The Textual Communities workspace is a tool, in the first instance, for defining, transcribing, and linking textual materials for a digital archive or edition and for marshaling and managing a community of participants with an array of community building tools.

There are many transcription tools under development and a few that are already functional. There are three defining features that make this tool different than the rest: its integrated participant and document management systems, and its mapping of fundamental document-entity structure. These features correspond to two underlying principles: that the work of amassing large corpora of textual materials is best accomplished by a well-managed community of interested participants from within, but also potential from outside the academy; and that for the resulting materials to be useful, their relationships must be clearly articulated.

As its name suggests, the Textual Communities tool is designed for gathering and organizing multiple participants around a common editorial project. It supports a wide range of relational structures, from a carefully crafted team to ad hoc community built on crowd-sourcing. Crucially, it enables definition of roles in the project with varying degrees of access to project materials, and authority to do the work of pressing these materials, and oversight over other participants. It is also built on a data structure that uses RDF files built on the FRBR ontology to identify and relate the produced transcriptions ("texts"), the exemplars they derived from ("documents," usually in the form of a digital image of a particular witness), and the intellectual construct they instantiate (the "work," or our preferred term, "entity"). Thus anyone interested in John Donne's poem "The Good Morrow" will find various "texts" (transcriptions) of this work as found in the extant "documents" (the poem as it is found in each of the manuscripts and printed books that contain it).

The tool itself enables uploading of digital images of primary documents in jpeg, tiff, or pdf, and linkage of these images with a transcription space. The user supplies information for each document, which produces an RDF file that defines the text that is to be transcribed and its relationship to the source *document* and *entity.* The user also defines the structure of the document, which is rendered behind the visible transcription in TEI conformant XML. The transcription area, which is automatically linked to the source image, can also support any XML markup that is desired or required for intelligent transcription of the source document.

This open-source tool will be freely available free of charge for use and adaptation by anyone anywhere. Development of this tool is funded by a generous grant from the Canadian Foundation for Innovation with the support of the Digital Research Centre at the University of Saskatchewan.

This poster will be accompanied by a live demonstration of the transcription workspace.

The AIDS Quilt Touch Mobile Web App

NeuCollins, Mark
mark-neucollins@uiowa.edu
University of Iowa Digital Studio for Public Humanities, United States of America

Thompson, Kelly J.
kelly-j-thompson@uiowa.edu

University of Iowa Digital Studio for Public Humanities and School of Library and Information Science, University of Iowa, United States of America

Dudley, Nikki J.

nicole-dudley@uiowa.edu
University of Iowa Digital Studio for Public Humanities, United States of America

Haldeman, Lauren

lauren-haldeman@uiowa.edu,
University of Iowa Digital Studio for Public Humanities and Virtual Writing University, University of Iowa, United States of America

Winet, Jon

jon-winet@uiowa.edu
University of Iowa Digital Studio for Public Humanities, United States of America

Haar, Kayla

kayla-haar@uiowa.edu
University of Iowa Digital Studio for Public Humanities, United States of America

Q: "What weighs 54 tons and can be held in the palm of your hand?"
A: "The AIDS Memorial Quilt."

AIDS Quilt Mobile Web App Development Team:

Mark NeuCollins (Lead Developer, Database, PHP, Drupal, CSS)
Lauren Haldeman (Drupal, CSS, jQuery Mobile)
Nikki Dudley (PHP, Drupal, Database, Import Scripting)
Kelly Thompson (Drupal, Mapping, Experimental Module Research, User Experience)
Kayla Haar (Graphic Design)
Jon Winet (Project Director, NAMES Foundation Liaison, Editor, Publicity Information)

The AIDS Memorial Quilt is the largest living monument in the world. Composed of over 48,000 three foot by six foot individual panels, it pays tribute to the lives of more than 98,000 individuals who have died during the AIDS pandemic. Maintained by the NAMES Foundation in Atlanta, each panel is painstakingly hand-crafted by those who knew and loved these individuals. Each panel carries the emotional weight of a life lived, of loving relationships, and of heartfelt loss.

Our interdisciplinary team —drawn from backgrounds in information technology, the humanities, and public art, and including undergraduate and graduate students, staff, and faculty — would like to share our experiences building a public digital humanities project that provides access to a virtual experience of the Quilt. Provided with data stored in a structure created in the 1980s, we were tasked with transforming this arcane information repository into something robust, agile, and modern. We aimed to design a system that would be usable by the community surrounding the quilt, comprised of people from all walks of life. The process of creating this public digital humanities project, with its goal of preserving the culture and purpose of the original Quilt, presented many learning opportunities we believe could be of value to others seeking to undertake similar projects. If accepted, we propose a discussion around the cultural, technological, artistic, and community-driven aspects of developing this project. We would also like to discuss the outpouring of community-sourced stories, memorials, and heart-felt comments contributed in the space of this digital memorial by those who visited the Quilt this summer, both physically and virtually.

Working in concert with project director Anne Balsamo at the USC Annenberg Innovation Lab, the University of Iowa Digital Studio for Public Humanities (DSPH) developed "AIDS Quilt Touch." A mobile web application for mobile devices and laptop | desktop computers, AIDS Quilt Touch is a digital extension of the Quilt, which celebrated its twenty-fifth anniversary this year.

DSPH was approached last spring to create this digital version of the Quilt in conjunction with a display of all 48,000 panels on the National Mall in Washington DC in July 2012. The immediate purposes of the app were to allow visitors to the Mall to find the panel of their loved one, and to leave digital remembrances. The timeline for the project was formidable, with just two months to produce a functional beta version of the app.

With records dating from the 1987, the NAMES database is a cobbled-together collection. Our first task was to parse the flat spreadsheet document from the NAMES foundation into a relational database format that we could use. Simultaneously we needed to learn enough about the Drupal Content Management System to deliver a complete and stable mobile web app based on these data. This web app was to be used by thousands of people trying to find the panel of their loved ones during the quilt's display in Washington DC. The stakes were high, the challenges were great, but in an incredibly satisfying and successful collaborative effort, we pulled it off. You can see the results of our efforts here: http://www.aidsquilttouch.org/

To the question of "When will the quilt be on display?" we can now answer, "It is always on display." To the question of "Where is it being displayed?" we can now answer, "Everywhere." There is much that we can add to

the quilt application, and plan to continue development with future displays of this living monument.

The AIDS Quilt Touch mobile web app allows people to leave comments, to extend the narrative that the quilt has begun to tell, and to create virtual celebrations of the lives lived. We plan to expand the types of media users are able to add in the near future: photographs, audio, video, and information and metadata about the quilt panels, those who constructed the panels, and those whose lives are memorialized by the panels.

The celebrations of life that users have left on the mobile app convey a deep resonance with the heart of the human experience. It is these comments that begin to show the promise of the technology. This is the good stuff—the material of human culture. It is not the mobile web app that is important, but the possibility that this technology can facilitate a deep conversation, can create a well of experience from which we can all draw. The mobile web app points to the possibility that these devices we carry in our pockets hold the potential to be portals to a larger and more inclusive cultural realm.

Programming with Arduino for Digital Humanities

Ohya, Kazushi
ohyakazushi@gmail.com
Tsurumi University, Japan

Background

The department of library, archive and information studies at Tsurumi University provides courses on computer science for humanities students. The courses include one introductory and two intermediate programming courses. There are project-based courses on Digital Humanities instead of an advanced programming course. In the introductory course, Scratch was used, and in the intermediate courses, Java has been adopted.

Difficulties in teaching programming in an introductory course to humanities students have been to let students find interest or enjoyment in (1) symbol manipulation and (2) grouping tasks with functions or other similar units.

Humanities students tend to expect big results in programming. They are disappointed at small results from source codes in exercises. It is usually difficult to expect them to find interest in a small result from symbol manipulation in introductory programming.

As far as our experience goes, humanities students seem to struggle to envision the existence of a computational world in their minds. The formats and abstract behaviors of typical programming patterns appearing in structured programming such as assignment, iteration, condition and flow control themselves are not difficult for humanities students to understand. The grammar is easy to comprehend. The problem they face is to understand the ways to use them as substantial components making up the whole code in actual programming (Winslow 1996).

Embedded System: Arduino

In order to respond to aforementioned difficulties or problems, we set two requirements to the programming environment:

(1) letting students be interested in something moving, and
(2) letting students be satisfied with small results.

Then, we decided to introduce embedded systems to an introductory programming course. We expected embedded systems to bring the real world into a scene of learning programming.

Current students have been immersed in digital or virtual environment from an early age. They are not easily satisfied with computational results on screens. On the other hand, interestingly, they show their interest to physical phenomena even if it is slight (Ishii 2006). They are very sensitive to physical stimuli. Embedded systems can be expected to let students be interested in something moving that may be small. And, it can contribute to facilitate students to envision a computational world in their minds.

As an embedded system, we adopted Arduino which provides a good developing environment:

(1) there is no need to prepare a device to install native codes to ICs,
(2) there is an easy IDE based on Processing , and
(3) the price of Arduino is affordable.

A concern we noticed about introducing Arduino was that students have to learn about electricity to some extent. It is an unfamiliar subject to humanities students.

The IDE for Arduino on Processing provides simple descriptive rules reminiscent of BASIC, and easy to view and write grouping tasks in methods. We expected this feature would make it easy for students to concentrate on finding categories of processes and making groups with functions.

Course Design

One semester consists of 15 classes, which are lectures on and practices of programming. The practices consist of four themes: LED handling, variable registor, sound handling, and binary display.

In all practices, we used only three circuit patterns based on a voltage divider. The typical four programming patterns in procedural languages were learned with eight source codes using one circuit pattern of an LED. We used a variable resistor as an input device and a game controller in learning structural programming. An array is not a difficult topic in programming, but it is not easy for students to understand the usefulness of the array. Arrays working as music scores to handle sound seemed to be a good example. We think that even though a topic of binary digit could not be useful in Digital Humanities, students should learn it to feel the philosophy behind symbol manipulation. We used bit operation to control LEDs of a binary digit display, and provided a chance to learn binary digit.

Observation and Future

Fortunately, many students seemed to find enjoyment from a small result in physical programming and grouping tasks. However, as we expected before this experiment, learning electricity seemed to be difficult for students. For example, Ohm's law was difficult for students who had learned it in junior high school. It might be possible to teach electronic circuit like LEGO Block without any explanation about a theory or background knowledge. However, learning theories in nature is inevitable in science education. We are planning to devise course materials to reduce the offset of learning electricity for next year.

References

Banzi, M. (2011). *Getting Started with Arduino.* 2nd edn. O'Reilly Media.

Gibbs, N. E. and A. B. Tucker (1986). A Model Curriculum for A Liberal Arts Degree in Computer Science. *Communications of the ACM* 29(3), ACM.

Ishii, H. (2006). *Tangible User Interfaces.* CHI 2006 Workshop Proceedings.

Kobayashi, S. (2010). *Prototyping Lab in Japanese.* O'Reilly Japan.

Liberal Arts Computer Science Consortium (LACS) (2007). A 2007 Model Curriculum for a Liberal Arts Degree in Computer Science. *ACM Journal on Educational Resources in Computing* 7(2). ACM.

Marshall, P. (2007). *Do tangible interfaces enhance learning?* TEI'07. ACM.

Monk, S. (2012). *Programming Arduino*. McGraw-Hill Companies.

Schmidt, M. (2011). *Arduino, The Pragmatic Programmers*.

Winslow, L. E. (1996). Programming Pedagogy — A Psychological Overview. *SIGCSE Bulletin* 28(3). ACM.

Arduino, http://http://www.arduino.cc/

Processing, http://http://processing.org/

Scratch, http://http://scratch.mit.edu

Building a Digital Curation Workstation with BitCurator

Olsen, Porter
polsen@umd.edu
MITH Maryland Institute for Technology in the Humanities, University of Maryland, United States of America

Kirschenbaum, Matthew
mkirschenbaum@gmail.com
MITH Maryland Institute for Technology in the Humanities, University of Maryland, United States of America

This poster builds on the recent report from the Online Computer Library Center (OCLC) titled “You’ve Got to Walk Before You Can Run: First Steps for Managing Born-Digital Content Received on Physical Media” (Erway, 2012). The report identifies eleven specific steps archivist can follow to safely and effectively process born-digital content. This poster considers the hardware needs of archivists and scholars as they look to implement the OCLC’s recommendations by offering a model born-digital curation workstation using readily available PC hardware and a suite of free and open source tools being developed and extended by the BitCurator project. It also offers a use-case demonstrating why such a workstation is a valuable asset for a working digital humanities center.

While a number of commercial solutions are available to aid in processing and curating born-digital collections, the cost of these solutions are frequently beyond the means of libraries and archives at smaller institutions (Kirschenbaum et al, 2010). BitCurator helps ameliorate these obstacles by making available a suite of open source digital forensics tools that run on industry standard PC hardware. The use of standard hardware and open source tools lowers the cost of entry to born-digital curation, thereby enabling

more institutions to begin processing their born-digital collections.

The poster offers a best practices example of how to build a digital curation workstation. In conjunction with basic system hardware, a digital curation workstation requires a wide range of media access devices. Some of these devices, such as flash drive readers and DVD/CD-ROM drives, are readily available and come standard on most desktop PCs. However, older media devices, such as 3.5" and 5.25" disk drives, can be more difficult to find and integrate into a present-day computer. In addition, older media may not be formatted for a file system that is recognized by current operating systems, so part of this poster will include instructions on how to use the USB based FC5025 5.25" drive controller to access data across a range of file formats.

While this poster presentation offers practical solutions to some of the challenges of digital curation, it is also intended to promote serious inquiry into the need for the preservation of our digital heritage. In fact, much early work in the digital humanities may be in need of reclamation through tools such as those described above, as was the case recently at MITH with the Shelley-Godwin Archive. That project extends textual analysis work done by Neil Fraistat in the late 1980s on Percy Shelley's poems and manuscripts. However, that early work was done on a PC running MS-DOS and saved as WordPerfect 4.2 files on 5.25" disks. Using the digital curation workstation described in this poster, we were able to access the floppy disks, download the original files onto a flash drive, and then convert them into the current Word format, all while preserving the original notations. This example illustrates the need for effective and accessible tools for the preservation and recovery of our digital past, both in terms of scholarship and born-digital culture. This poster will enable other digital humanists to take advantage of the work that is being done in the field of digital curation to help preserve and extend their own research for future generations of scholars.

References

Erway, R. (2012). *You've Got to Walk Before You Can Run: First Steps for Managing Born-Digital Content Received on Physical Media.* OCLC Research.

Kirschenbaum, M. G., R. Ovenden, and G. Redwine (2010). Digital Forensics and Born-Digital Content in Cultural Heritage Collections. *Council on Library and Information Resources.*

Lee, C. A., M. Kirschenbaum, A. Chassanoff, P. Olsen, and K. Woods (2012). BitCurator: Tools and Techniques for Digital Forensics in Collecting Institutions. *D-Lib Magazine* 18: 5-6.

The INKE NewRadial Prototype: Evolving the Space and Nature of Digital Scholarly Editions

Saklofske, Jon
jon.saklofske@acadiau.ca
Acadia University, Canada

How can digital scholarly editions take full advantage of environmentally-generated opportunities to focus on process, collaboration, and distributed control without losing the traditional affordances that make an edition "scholarly?" The Modeling and Prototyping Team of the Implementing New Knowledge Environments (INKE) project is currently exploring ways in which the scholarly edition can be re-imagined within digital settings. Our prototypes function as virtual environments that encourage play within their designed frames, and as Galey and Ruecker (2010) have argued, the trajectory of prototype iterations establishes a valuable record of critical enquiry. In this spirit, we wonder whether the digital scholarly edition, in addition to being perceived as an environment which is a trace record of the theoretical and argumentative motivations that inform the editorial processes of selection, organization and design, could actively and dynamically host the formation of multiple, simultaneous, and community generated editions. Material print editions are records, artefacts that efface the process of their formation, version-objects that assert an argument and establish a historical position through the printed finality of their collation and production. If digital editions are to take full advantage of their environments (rather than simply emulating print traditions) they need to visibly include both process and product, and offer opportunities for editorial diligence, contribution, perspective, control and debate to their users. Top-down forms of authoritative and exclusive editorial selectivity become ironic and anachronistic in dynamic digital environments which privilege "a new kind of scholarly discourse network that eschews traditional institutionally-reinforced hierarchical structures" (Siemens 2011). We are exploring modelling the digital scholarly edition as a social edition workspace in which a community of users can contribute content and emerge from the debilitating condition that William Blake described as "single vision."

In this spirit, and to provide essential opportunities for user-based contributions and scholarship within digital edition environments, the INKE Modeling and Prototyping team is currently developing a software environment called NewRadial that significantly reinvents an earlier prototype designed by Saklofske and Giffin (2009) and builds on earlier conceptual work by Nowviskie (2007) and Saklofske (2010, 2011). NewRadial's collaborative space is a reimagining of the digital scholarly edition as a transparent workspace layer in which established primary objects from existing databases can be gathered, organized, correlated, annotated, and augmented by multiple users in a dynamic environment that also features centralised margins for secondary scholarship and debate. The INKE NewRadial prototype—consisting of an HTML5 frontend and server-based backend—is a workspace that uses simple adapters to query databases for specific results, and then uses those results to harvest representations of the objects (i.e. thumbnails) to populate its workspace. Linked data and annotations produced by a community of users in relation to these objects within NewRadial's environment then become available to other applications. Our focus on relational data models, scalable data browsing, and crowd-sourced descriptive frameworks means that the INKE NewRadial prototype is being designed as an effective means for working with all types of media objects, for aggregating search results from multiple databases using meta-adapters, and for making its RDF-based secondary scholarship and annotation data available over HTTP for use in other tools and workspaces. Currently, our prototype installation has successfully used adapters to import NINES/ARC data, the Archbook image repository, Google images and other scholarly database holdings.

NewRadial's affordances introduce a dynamic multiplicity of vision into what has traditionally been a reductive, oppositional and snail's pace process of inter-edition debate and evolution. The development of this digital edition environment prototype is the first step towards creating inclusive editorial workspaces which draw from broad data foundations and which encourage knowledge-building communities to actively reimagine edition-building processes. This poster/demonstration session is designed to offer conference participants hand-on experience with the INKE NewRadial prototype and to demonstrate the ways that the unique affordances and flexibility of this workspace impacts the nature of scholarly editing, the scholarly edition itself, and the secondary scholarship that such editions generate.

References

Galey, A., and S. Ruecker (2010). How a Prototype Argues. *LLC*, 25(4): 405-424.

Nowviskie, B. (2007). Collex: Collections and Exhibits for the Remixable Web. *Electronic Book Review*. 1-17.

Saklofske, J. (2011). Remediating William Blake: Unbinding the Narrative Architectures of Blake's Songs. *European Romantic Review* 22(3): 381-88.

Saklofske, J. (2010). NewRadial: Re-visualizing the Blake Archive. In *Poetess Archive Journal*. 2(1).

Saklofske, J., and J. M. Giffin (2009). NewRadial. http://sourceforge.net/projects/newradial/

Siemens, R., et al. Toward Modeling the Social Edition: An Approach to Understanding the Electronic Scholarly Edition in the Context of New and Emerging Social Media, Accepted for publication in *Literary and Linguistic Computing.* 70.

DARIAH-EU's Virtual Competency Center on Research and Education

Schöch, Christof
c.schoech@gmail.com
University of Würzburg

Costis, Dallas
c.dallas@dcu.gr
Digital Curation Unit — IMIS, Athena Research Centre

Munson, Matt
mmunson@gcdh.de
Göttingen Centre For Digital Humanities (Gcdh)

Tasovac, Toma
ttasovac@humanistika.org
Belgrade Center for Digital Humanities

Champion, Erik Malcolm
echampion@hum.au.dk
DIGHUMLAB Denmark

Schreibman, Susan
susan.schreibman@gmail.com
Trinity College Dublin

Benardou, Agiatis

agiati.benardou@gmail.com
Digital Curation Unit — IMIS, Athena Research Centre

Huang, Marianne Ping
mph@adm.au.dk
Aarhus University, Denmark

Links, Petra
p.Links@niod.knaw.nl
NIOD Institute for War, Holocaust and Genocide Studies

DARIAH (Digital Research Infrastructure for the Arts in Humanities — http://dariah.eu) is a large-scale, long-term, pan-European endeavor aiming to enhance and support digitally-enabled research across the arts and humanities. DARIAH aims to develop and maintain an infrastructure in support of ICT-based research practices. It will explore and apply ICT-based methods and tools to enable new research, improve research opportunities through linking distributed digital source materials and tools, and exchange expertise, methodologies and practices across domains and disciplines.

The aim of this poster is to present one of the four primary contact points of DARIAH, the *Virtual Competency Center Research and Education Liaison* (VCC2). We would like to inform the DH community of our aims and encourage researchers to contact us and explore cooperation opportunities. The VCC is led by Susan Schreibman (Trinity College Dublin, Ireland) and Erik Champion (DIGHUMLAB, Denmark). The VCC's activities fall into four areas, each of which has a coordinator who serves as the primary contact point.

Understanding Research Practices

VCC2 aims to develop an evidence-based foundation to ensure fitness for purpose of planned digital infrastructures for scholarly research.

To achieve this objective, VCC2 will develop a research protocol and manual for transnational longitudinal research on scholarly practices and for digital tools and services user requirements elicitation. It will conduct baseline research on scholarly research practices, digital tools and services requirements for the arts and humanities across Europe. It will develop a knowledge base of empirically attested research practices in the arts and humanities, by tracking a) information access, use and curation processes, and their relations with disciplines, methods, research objects, tools and services, and b) researcher feedback and case studies on digital tools and services use.

Work will build on earlier interview-based and questionnaire survey research in the "Preparing DARIAH" and EHRI projects, and on synergies with projects such as eCloud, ARIADNE, and NeDiMAH.

Training and Education Program

VCC2 aims to provide a training program for researchers in the methods, tools, and approaches needed to engage with the digital environment, including DARIAH services, tools, and content.

The Training and Education Program will focus in particular on the development and delivery of international summer school programs; development of collaborative, consortium-wide online training materials; and activities that foster a better understanding of teaching DH across disciplinary, institutional, linguistic and cultural borders. In addition, the program will collaborate with institutions providing undergraduate and postgraduate training in the digital humanities to embed DARIAH tools and services in their courses and provide structured feedback on their use.

As part of its effort to support academic mobility and international opportunities in DH education, the Training and Education Program will also create a registry of undergraduate and postgraduate DH courses in Europe.

Community Engagement

Engagement with digital tools, methods and content is an emerging practice for many arts and humanities researchers. VCC2 will seek to engage with these scholars about digital humanities, explain the added value digital data and methods can provide whilst seeking to foster exchange and cooperation within scholarly communities.

To promote engagement with the increasingly large volume of research data being digitized, and in turn the research infrastructures designed to support the use of this data, we support communities of researchers in coming together to learn from each other and express their requirements.

Services to this end include: a series of workshops aimed at younger researchers wishing to make the move towards digital research methods; expert meetings in which confirmed researchers focus on the further development of specific methods and tools; a publication series documenting results from DARIAH's activities.

Virtual Research Environments

VREs are changing. What previously were closed systems where everything needed to be developed for that system are now becoming more modular systems where

the appropriate outside tools can be plugged into a VRE platform.

With this in mind, it seems counterproductive to attempt to create a one-size-fits-all VRE. Instead, VCC2 expects to produce a VRE Blueprint: a document that can be used to guide the conversations between humanities scholars who wish to develop a VRE and the computer developers who will help them to realize this goal.

The blueprint will consist of a list of technical requirements that every VRE will have to meet, questions to guide the conversations between developers and scholars, an annotated bibliography, a list of common pitfalls in VRE development, and an end-user survey template to help developers get feedback from the VRE's users.

Framework for Testing Text Analysis and Mining Tools

Simpson, John Edward
john.simpson@ualberta.ca
University of Alberta

Rockwell, Geoffrey
grockwel@ualberta.ca
University of Alberta

Sinclair, Stéfan
stefan.sinclair@mcgill.ca
McGill University

Uszkalo, Kirsten C.
circe@ufies.org
University of Alberta

Brown, Susan
susanirenebrown@gmail.com
University of Alberta; University of Guelph

Dyrbye, Amy
amy.dyrbye@gmail.com
University of Alberta

Chartier, Ryan
recharti@ualberta.ca
University of Alberta

The most extensive compendium of text mining tools to date included 71 tools and summarized each based on ten criteria (van Gemert 2007). While extensive, this listing of tools and their properties is general in its review criteria and does not offer any testing-based observations to help users assess actual usability. Humanists looking to try text analysis, visualization and mining tools for research need better information that is relevant to their needs and reviews of tools that help them make choices. This poster presents the testing framework developed for the TAPoR 2.0 portal reviews. The poster will cover:

1. The need for tool reviews
2. The information gathered about tools
3. The testing and reviewing process
4. Conclusions about the state of text tools

The poster will be accompanied by a demonstration of TAPoR 2.0 so that users can see the reviews in context.

1. The Need for Tool Reviews

A humanities researcher new to computing methods looking for reviews of text tools on the internet by peers is going to be disappointed. There is nothing like the *New York Review of Books,* though in the early days of humanities computing you could find short announcements about tools in journals like *Computing in the Humanities.* We, however, believe that certain text tools are intellectual contributions to the field (Ramsay 2012) that should be reviewed not just to help people choose what tools to use, but also as a way of engaging these tools in a dialogue around computer-assisted interpretation. While there are individual blog entries about tools scattered across the web, each is from the perspective of a single user with an entirely different dataset, making comparison difficult. If we want to make computing methods accessible and encourage colleagues to use tools we need a more systematic approach. This is especially true of text mining tools that can't simply be tried with a text at hand.

2. Information Gathered About Tools

TAPoR 2.0 (www.tapor.ca) is a portal for text analysis, visualization mining tool discovery and review. TAPoR 2.0 is a complete redevelopment of the original TAPoR portal (Rockwell 2009) that has refocused the portal on discovery and review instead of trying to provide access only to web services. As part of the redevelopment of TAPoR 2.0 we used a persona/scenario usability design approach (Cooper 2004) to identify attributes that users

might want to discover tools. Further we built TAPoR 2.0 so that editors can add new attributes without the database having to be reprogrammed. Some of the attributes we currently record for tools include the author(s), ease of use, type of analysis, type of license and so on. We also have links to related tools and tools people also used. Our poster will be accompanied by a demonstration of TAPoR 2.0 so that visitors can explore what we have and how we represent it.

Figure 1:
TAPoR 2.0 Home Screen

3. The Testing and Reviewing Process

Recording basic information about tools alas, is not enough, especially with sophisticated text mining tools like Mallet (mallet.cs.umass.edu/) that take time to learn and that can be used in different ways. With text mining tools users need longer narrative reviews. For this reason we developed processes for testing and reviewing tools. For simpler text analysis and visualization tools this involved developing a set of different texts with which to test tools so we could compare their use. For text mining we had to go further and are working with the CWRC project (Canadian Writing Research Collaboratory, www.cwrc.ca) developing a number of literary corpora with experts we can draw on to help assess the value of results. As of writing we have three corpora drawn from the Orlando project (www.ualberta.ca/ORLANDO) and one of Victorian children's literature. We expect to have two more by the time or presentation. The poster will discuss the criteria used to develop these open test corpora.

The reviews take the form of comments that have been pinned to the top of the list of comments available. This allows others to leave comments, though we haven't seen much activity by people not connected to the project (with the exception of spammers who seem to feel there is a connection between text analysis tools and various stimulants.) We have developed guidelines for reviews so as to make them accessible and comparable. The poster will outline our guidelines.

4. Conclusions from Testing and Reviewing

Having tested and reviewed a variety of tools and text mining systems we see some common barriers to access. Most of these tools have been developed for use by the developers and are poorly documented for people not involved in the development. Further, many tools, including those we are involved in, are in continuous development so what documentation there is, is out of date. We will therefore end this poster with lessons learned while testing and reviewing text mining tools, with particular attention to removing usability barriers for novice users.

References

Cooper, A. (2004). *The Inmates Are Running the Asylum.* Indianapolis, Indiana: SAMS.

Ramsay, S. and G. Rockwell (2012). Developing Things: Notes toward an Epistemology of Building in the Digital Humanities. In Gold, M. K. (ed.), *Debates in the Digital Humanities*. Minneapolis, Minnesota: University of Minnesota Press, 75-84.

Rockwell, G. (2006). TAPoR: Building a Portal for Text Analysis. *Mind Technologies: Humanities Computing and the Canadian Academic Community.* edited by Raymond Siemens and David Moorman. Calgary: University of Calgary Press, 285-299.

van Gemert, J. (2000). *Text Mining Tools on the Internet.* ISIS Technical Report Series. http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=4312886.

Voyant Notebooks: Literate Programming and Programming Literacy

Sinclair, Stéfan
sgsinclair@gmail.com
McGill University, Canada

Rockwell, Geoffrey
grockwel@ualberta.ca
University of Alberta, Canada

Some thirty years ago Donald Knuth, a computer scientist, proposed literate programming as a better way of organizing narrative and code (1984). Knuth argued that more emphasis should be placed on explaining to humans what computers are meant to do, rather than simply instructing computers what to do. Knuth was especially interested in weaving together macrostyle code snippets with prose that provided a larger narrative context, not merely functional comments of specific lines of code that are the distilled remnants of an intellectual process.

Literate programming has been more influential in theory than in practice (Nørmark), despite several utilities and environments including Mathematica, Knuth's (C)WEB, Sweave for R, and Marginalia for Clojure. Perhaps the exigencies of programming in the real world correspond poorly with the vision of Knuth of the programmer as author: "the practitioner of literate programming can be regarded as an essayist, whose main concern is with exposition and excellence of style" (1992, 1). However, that balance of essayist and coder strikes us as perfectly appropriate for the digital humanities, a natural blend of the expression of intellectual process with the exposition of technical methodologies. The prose can gloss the code, or viceversa, in a symbiotic relationship that serves to strengthen an argument and demonstrate its own workings.

One of the most significant potential benefits of the literate programming paradigm is pedagogical: these works can both explain an interpretive insight and present the methodology for reproducing the data or results that were part of the process. Many widely-read digital humanities blogs already present these characteristics of exploration, explanation, interpretation and step-by-step instructions (see for example blogs by Ted Underwood, Benjamin Schmidt, Lisa Rhody and Scott Weingart). Literate programming can be more self-contained and more useful for those learning new methodologies and new programming techniques. This is about the principles of literate programming, but also about the potential for increasing programming literacy.

This poster will introduce Voyant Notebooks, a web-based literate programming environment designed for the digital humanities (see Appendix A). There is already a working prototype and we anticipate having a more feature-rich version available by July 2013. Voyant Notebooks inherits many of the characteristics of the Voyant Tools environment, including a concern for usability and flexibility (researchers and students should be able to use it with minimal or no training and with their own texts of interest). Voyant Notebooks also addresses one of the main weaknesses of Voyant Tools: the fact that most tools are constrained by assumptions about how they would be most commonly used. For instance, the Wordle-like (word cloud) Cirrus tool is designed to show the top frequency terms from a corpus or document; but what if the user instead wants to visualize the top frequency nouns, or people, or repeating phrases? All of that functionality could be built into the tool, but possibly at the cost of usability (endless menus and options), and it could still never address all of the possible use cases. Voyant Notebooks, by contrast, empowers the user to customize some of the functionality by leveraging the analytic capabilities of the Voyant back-end and the visualization interfaces in the front-end (like Cirrus). Our poster will have two parts, a) a usable demonstration on one or more laptops and b) a poster that illustrates how Voyant Notebooks implements Knuth's concept of literate programming. In addition to these conceptual aspects, the poster will outline technical details about the Voyant Notebooks prototype for those interested, including the technologies used for both client-side (browser) and server-side components. Some of the technical challenges that will be described include:

managing the flow of code execution in an asynchronous architecture,
using web workers to avoid browser freezes during longer executions,
mitigating the security risks of user-defined and persistent Javascript code,
code variable scoping across editor instances and window components,
embedding of Voyant tool panels (visualizations) and other services,
developing a flexible API for different programming levels and styles,
developing an API that includes both client-side and server-side operations, and
ensuring efficiency of repeated code snippets during writing and viewing.

And of course, visitors to the poster session will be warmly encouraged to play with Voyant Notebooks.

Appendix A: Mockup of Voyant Notebooks (previously called Voyeur Notebooks).

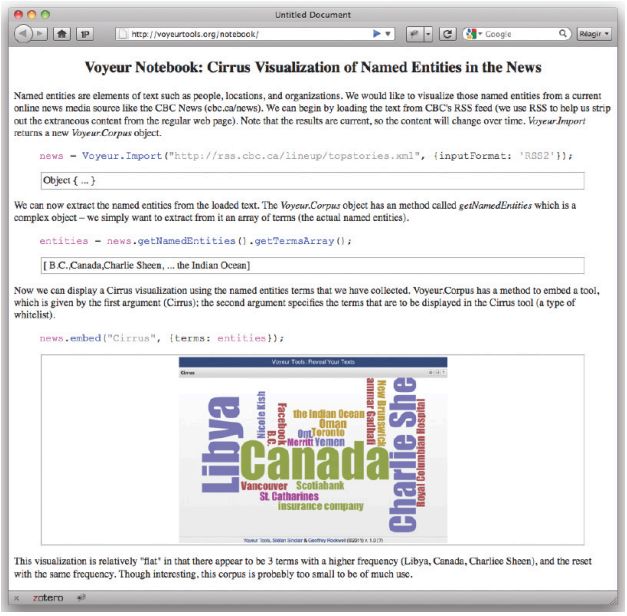


Figure 1:
Mockup of Voyant Notebooks

References

Knuth, D. (1984). Literate Programming. *The Computer Journal* 27(2): 97-111, 1.

Knuth, D. (1992). *Literate Programming.* Stanford University Center for the Study of Language and Information.

Nørmark, K. (1998). Literate Programming: Issues and Problems. http://www.cs.aau.dk/~normark/litpro/issues-and-problems.html.

Sinclair, S. and G. Rockwell (2012). Teaching Computer-Assisted Text Analysis: Approaches to Learning New Methodologies. in *Digital Humanities Pedagogy.* Open Book Publishers.

Reliable Citation as a Foundation for Preservable Web-Based Digital Humanities Projects

Smith, James
jgsmith@gmail.com
University of Maryland, United States of America

Preservation of digital humanities (DH) projects is an emerging problem. As languages, frameworks, platforms, libraries, and databases evolve, the effort required to maintain meaningful access to existing projects is a challenge that competes with efforts to produce new work. Creators of DH projects need to be able to continue to iterate and innovate but also demonstrate good stewardship of digital materials.

The field recognizes this problem, from the 2004 Sustaining Digital Scholarship (SDS) final report and continuing through such projects as Memento, SiteStory, and TAPAS, and is actively researching ways to enable sustainable, long-term stewardship for certain project components, but these efforts do not provide a comprehensive platform for the full DH project. All of these efforts consider a project to be a set of static pages or resources that seldom change. None of them would be sufficient for hosting or describing ongoing projects such as *World Shakespeare Bibliography Online.*

Many new DH projects build their web presence with open source platforms such as Wordpress, Drupal, or Omeka, extending them through customizations. The long-term sustainability of such projects is an issue, involving the cost of hosting as well as migration of customizations as platforms and languages evolve. SalahEldeen and Nelson show evidence that after only a year, an average of eleven percent of cited on-line resources are lost. It is not surprising then that none of the common systems used to build DH projects allows reliable citation since the same platforms used in DH host a share of the resources lost each year. It is surprising that reliable citation is not seen as a greater problem by the scholarly community. Reliable citation is necessary for long-term preservation and therefore expresses a need for a temporal content and data management system that allows for reliable citation of scholarly narratives and resources.

We might base the value of a scholarly work on its place in the larger scholarly conversation. If it is not part

of the conversation, then it has little effect on the field and thus has little value. Scholarly work can only be part of the conversation if it can be referenced. This is well developed for traditional publications (e.g., citing a particular edition of a printed work), but remains a problem for web-based scholarly work, not because particular pages can not be addressed, but because the information presented as part of that page is not stable. Without being able to reference the particular version of the page, scholars can not make reliable arguments about the work. What a reader sees might differ from what the author saw when researching and writing the work referencing the webbased project. These problems increase when referencing a dynamic, algorithmic project.

We see three fundamental requirements for such a system to enable reliable citation of scholarly works and resources: temporal citation, reproducible citation, and sustainable citation. Any platform meeting these requirements should be able to provide level four preservation as described in the SDS final report (11).

A temporal citation of a web-based scholarly work must be able to address the view within the context of the project's history. Not only must a scholar be able to point a reader to a particular resource in a project, but the scholar must be able to point to a particular resource at a particular date and time.

A reproducible citation of a webbased scholarly work must show the same content over time. Fetching the cited resource year after year should show no significant changes in the scholarly content of the resource.

A sustainable citation of a web-based scholarly work allows a scholar to cite a project and know that their readers will be able to see the same information they saw by following the citation, for as long as their citation exists. Sustainability is a social issue as much as a technical one. We are not trying to address the social issues involved in sustainability in this poster.

The poster consists of diagrams and text explaining how reliable citation works with respect to resource versioning and project timelines. In addition, a demonstration of a temporal content and data management system hosted at http://alpha.ookook.net/ providing reliable citation will accompany the poster so that attendees can interact with the system and see the platform affordances in action. The demonstration will also provide an opportunity for attendees to interact with the developer.

Reliable citation does not require any unique data model or software architecture. The poster outlines both the data model and the architecture as they are developed in the demonstration software, principally by segmenting a project's history into discrete editions that aggregate changes to the project.

Discussion will quickly muddy if we don't establish some nomenclature for dates and times. A resource date and time is the date and time for which the resource should be rendered. For example, if I specify a resource date and time of noon on January 1st, 2012, then I expect the see a rendering of the resource as it appeared at noon on January 1st, 2012. A request date and time is the date and time at which the request is made, even if the request is for a resource with a resource date and time different than the request date and time.

The data model partitions the project into two classes of objects: editioned objects, such as a project or theme, and versioned objects, such as pages or stylesheets. Editions are published for a span of time during which no public changes are made to the pages. Any changes made to a page require the creation of a new page version which will be aggregated with other versions when a new edition is created and published. Only one project edition is active for a resource date and time. By tracking the time spans for which an edition is active, we can reproduce the project as it existed at a particular date and time.

The demonstration software separates information into two editioned resources: Projects (web sites) and Themes (collections of style information). Editions of projects and themes are independent of each other, with each managing their own history.

References between different editioned resources are done by naming the referenced resource as well as the referenced resource date and time. This allows a project to select a theme in a reproducible fashion. References to a versioned resource within an editioned resource (e.g., a page within a project) may reference the page without referencing a particular version. The appropriate version will be retrieved based on the edition selected by the resource date and time of the request.

This poster will be of interest to anyone wishing to see how a platform supporting reliable citation might be designed.

References

Drupal. http://www.drupal.org/

Memento. http://www.mementoweb.org/

Omeka. http://www.omeka.org/

SalahEldeen, M. Hany, and M. L. Nelson Losing My Revolution: How Many Resources Shared on Social Media Have Been Lost? (2012) arXiv:1209.3026.

Sustaining Digital Scholarship Final Report. (2004) http://www2.iath.virginia.edu/sds/SDS_AR_2003.pdf

SiteStory. http://mementoweb.github.com/SiteStory/

TEI Archiving Publishing and Access Service (TAPAS) Project. http://www.tapasproject.org/

Wordpress. http://www.wordpress.org/

World Shakespeare Bibliography Online. http://www.worldshakesbib.org/

Visual Historiography: Visualizing "The Literature of a Field"

Staley, David J.
staley.3@asc.ohio-state.edu
Ohio State University, United States of America

French, Scot A.
scotfrench09@gmail.com
University of Central Florida, United States of America

Ferster, Bill
bferster@virginia.edu
Research Professor, University of Virginia

The call for visualizing "Big Data" has generated a groundswell of interest among historians and humanities scholars, as demonstrated by the international response to the National Endowment for the Humanities' 2010 and 2011 Digging into Data challenges (Williford and Henry, 2012). Exemplary efforts from the first two rounds of projects, such as Stanford University's "Mapping the Republic of Letters," and the University of Nebraska at Lincoln's "Railroads and the Making of Modern America," suggest the great potential for visualizing large repositories of primary sources for historical insight.

Our project treats the published work of historians in a peer-reviewed scholarly journal — *Florida Historical Quarterly,* housed at the University of Central Florida and accessible in digital form through J-STOR — as a primary source dataset to be analyzed and visualized. In applying macro-level reading and text-mining tools to the secondary literature of a scholarly field (History) and its subfields (American History/Florida History), we propose to make "visible" patterns of topical "coverage," changing conceptual/analytical/theoretical frames of reference, and patterns of scholarly influence. We know, for instance, that the Civil Rights Movement, feminism, and the opening of the academy in the 1970s had a profound impact on historical interpretation at the national level (Novick 1988). We expect our data visualization of *FHQ* journal articles to reveal the impact of these and other "turns" within the scholarly subfield of Florida History, both confirming and perhaps challenging assumptions based on more traditional reading practices.

"Knowledge mapping" is more commonplace in the sciences than in the humanities. Chaomei Chen asks, "Why do scientists not have a viewfinder to their own fields? Why cannot scientists videotape the evolution of their own fields, their paradigm shifts, the rises of their own stars and the expansion of their own galaxies of intellectual contributions? … How can we visualize the process of a paradigm shift?" (Chen 2006) We ask the same questions about history: can we use the tools of knowledge domain visualization to map out shifting paradigms and historiographic "turns" in history? What happens when we apply the techniques of "distant reading" (Moretti 2005) and human-assisted "machine reading" (Hayles 2012) to the secondary literature in history? What insights might visualization of the data produce for historiography? (Staley 2003) Our reading of the literature at this scale forces us to rethink our assumptions about how we understand "seminal" articles/works. More importantly, new questions emerge when we "read" the secondary literature of a field at this macro scale, such as determining the influence of journal editors on the topical/historiographic orientation of the journal.

This poster will present the results of our case study, which examines an 85-year run of the *Florida Historical Quarterly,* (1924-2009) made accessible and searchable via Data For Research J-STOR. We will display our findings through interactive visualizations generated in consultation with Ohio State and UCF graduate researchers, *FHQ* editorial staff, and visualization specialist Bill Ferster (Ferster, 2012) at the University of Virginia. Visual patterns, the result of a reading of a large textual corpus at a distance, can in and of themselves result in interpretive insights. Stephen Ramsay's "algorithmic criticism" (criticism derived from algorithmic manipulations of text) relies on the generation of visual patterns as the product of these computer-enabled interpretive acts (Ramsay 2011). Rather than being an intermediate step toward a written piece of scholarship — the preferred approach of humanists — the visualization itself becomes the hermeneutical/scholarly performance, the visualization is the hermeneutic object. Such a practice asserts "the primacy of pattern as the basic hermeneutical function [which would unite] art, science, and criticism." Our poster visualizes the "primacy of topical/historiographic patterns" in the *Florida Historical Quarterly*.

Our poster will outline our research methodology — a hybrid of qualitative and quantitative analysis, enabled by human-assisted machine-reading. We extend David Mimno's "computational historiography" (Mimno 2012) beyond a two-topic, p- or not-p approach to topic modeling. For analytic purposes, we will compare the results of our human-assisted "distant reading" with the perceptions of editors about how the field of Florida history, in its discourse with history in general, has changed over time. We will also problematize these results, pointing to the

potential "false positives" that can result from such machine reading and visualization. We will explore the relationship of necessity between machine-and human-reading.

References

Williford, C. and C. Henry (2012). *One Culture: Computationally Intensive Research in the Humanities and Social Sciences: A Report on the First Respondents to the Digging into Data Challenges.* Washington, D.C.: Council on Library and Information Resources.

Mimno, D. (2012). Computational Historiography: Data Mining in a Century of Classics Journals. *Journal on Computing and Cultural Heritage,* 5(1):1-19.

Hayles, N. K. (2012). *How We Think: Digital Media and Contemporary Technogenesis.* Chicago: University of Chicago Press.

Ferster, B. (2012) *Interactive Visualization: Insight Through Inquiry.* Boston: MIT Press.

Ramsay, S. (2011). *Reading Machines: Toward an Algorithmic Criticism.* Champaign, Il.: University of Illinois Press.

Chen, C. (2006). *Information Visualization: Beyond the Horizon.* New York: Springer.

Staley, D. (2003). *Computers, Visualization and History: How New Technology Will Transform Our Understanding of the Past.* Armonk, N.Y.: M.E. Sharpe.

Novick, P. (1988). *That Noble Dream: The 'Objectivity Question' and the American Historical Profession.* Cambridge: Cambridge University Press.

Not Exactly Prima Facie: Understanding the Representation of the Human Through the Analysis of Faces in World Painting

Suárez, Juan Luis
jsuarez@uwo.ca
The CulturePlex Lab, University of Western Ontario, Canada

de la Rosa Pérez, Javier
jdelaro@uwo.ca
The CulturePlex Lab, University of Western Ontario, Canada

Ulloa, Roberto
rulloa@uwo.ca
The CulturePlex Lab, University of Western Ontario, Canada

In his 1872 book *The Expression of Emotions in Man and Animals* (Darwin 1872), Charles Darwin drew our attention to the relationship between human expressions, movements and emotional states, and tried to frame his conclusions by highlighting the similarities between humans and animals. The light he shed on cultural differences with respect to the appearance of the face and variations in expressions perceived across different groups to communicate the same emotions is also important.

More recent scientific evidence highlights the importance that the human beings give to the face. The brain has a specialized amygdala to discriminate scenes in favor of facial expressions, a primitive mechanism to detect potentially dangerous situations (Hariri et al. 2002), that explains our impulse to immediately look at people faces and try to read their facial expression. Finally, the recent discovery of "mirror neurons" (Rizzolatti and Craighero 2004) and their connection with the imitative ability of several primates offers a glimpse about the social construction of emotions, helps to explain the spread of behaviors within human groups, and opens up the possibility of a phenomenology of human expressions.

Nevertheless, the fascination with the human face is not something new in humanities. It has been present since the beginnings of art history; artists have always sought to relate the face to the human body (Chase 2005) and, especially, to the different ways in which how faces reflect the human condition. In this sense, human facial representations contain a human expressions and emotions archive that can help us understand, through a science of the face (Cleese and Ekman 2001), various human condition traits that evolved through time and space. We aim to answer questions about periods in art history, such as the Baroque significance as a culture derived from human expansion. Even if we cannot say that Baroque is just as a historical period, before the discovery of America all faces in art were mostly european; the presence of indigenous faces in paintings is a big disrupt. Another example would be the cultural meaning of the progressive human face disappearing from modern painting. Our methodology analyzes this through facial recognition techniques, data mining, graph theory and visualization and cultural history.

Quantitative analysis of huge amounts of data has provided answers to new and different questions that otherwise couldn't have been considered (Michel et al.

2011). The study borrows some ideas from the Culturomics (Michel et al. 2011) concept by creating a set of more than 123,500 paintings from all periods of art history, and applying a face recognition algorithm used in Facebook's photo-tagging system#. The result is a set of over 26,000 faces ready to be analyzed according to several features extracted by the algorithm.

The extracted information accuracy may fluctuate depending on several factors, such as the reproduction quality and size; the thematic content, a portrait is not the same than a hunting scene; the pictorial style, e.g., Cubism versus Figurative art; the lighting, dark and hellish overtones as opposed to daytime or celestial images; or even the contrast between the background and skin tone. At the same time, these elements also provide information about how human faces are depicted by authors, styles and cultures across time. Once a face has been recognized, it provides data about the position of eyes, mouth, chin and ears — what we call the face *basic features*. If the confidence measure we have reached during the recognition process is satisfactory, we can consider gender, mood, position of lips, range of age or even if the person is wearing glasses — we call these *extended features* of facial recognition.

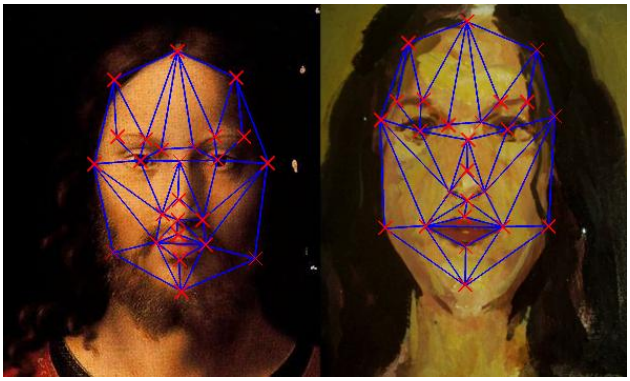


Figure 1:
Example of face graph for two randomly selected paintings. The red crosses indicate what we called face basic features. The blue lines show the different distances in between the features.

The methodology tackles the study of this huge amount of features in three steps. First, we build a graph with the set of *basic features*. Second, we look for clusters in the *extended features*. Third, we compare the graphs and the clusters, corresponding to the *basic* and *extended features* respectively, using time intervals and space as factors of the comparison. The selection of the procedures in each of the steps considers different strategies to address the fluctuations in accuracy. What follows is a more detailed explanation of the main mathematical and computational tools used for the processing of the data in each step.

For the *basic features* graph (Figure 1), we apply the Elastic Bunch Graph Matching (Wiskott et al. 1997) to the extracted data, and then we calculate the resulting weighted graph after applying graph similarity functions like Euclidean Geometry Similarity and Least Square Geometry Similarity. Once the graphs that represent the 26,000 faces are done, we normalize the weights and prune those relationships with weights below the first quartile. Then, a combination of YiFan Hu Multilevel (Hu 2005) and ForceAtlas (Bastian et al. 2009) layouts algorithms are run against the graph to see how the faces are clustered according to their modularity class.

For the *extended features* clustering, we create a vector for each face and run the K-Means method (Sculley 2010) that is able to cluster unlabeled data, i.e there is no need of a pre-training process but the classification emerges naturally from the data similarities. We check the clusters obtained according to their homogeneity and completeness, according to the definitions given by Rosenberg and Hirschberg (Rosenberg and Hirschberg 2007), in order to establish a good value for V-measure, a conditional entropy-based external cluster evaluation measure that indicates the success of a clustering solution.

Finally, we compare the two sets: the *basic features* set using graphs and the *extended features* set using clustering by K-Means method (Sculley 2010). At this point, we are at the perfect position to analyze and characterize each of the groups according to different historical perspectives and cultural questions, for instance, the distinction among styles by giving a minimum set of features that determines its membership. A similar set of features is obtained for a particular interval of time or a specific geographical region. Extending that analysis we are able to study the evolution of these sets of features over time.

In this study, we tackle three important issues that also show the potential of this kind of work to develop new approaches to cultural history and to establish some yardsticks in order to complement the incipient methodology for a Big History (Christian 2004) approach to human culture. First, we show how the analysis of the representation of human faces — both the internal features and in its relative position to the rest of the composition — offers important data to determine periods and borders in the history of art beyond the generalizations supported by the notions of "style", "genre" and "national history". Second, we study the correlations between the European expansion overseas from the 16th Century onwards, and the introduction of new human "types" in world paintings, focusing on concepts of identity and gender (with special emphasis on the size and form of the forehead), and relating the results to notions of Baroque, hybridization

and globalization (Suarez 2007; Suarez 2008). Finally, we move to the 20th Century and study the disappearance of the human face from art in relation to Ortega y Gasset's concept of the "dehumanization of art" (Ortega y Gasset 1968) and the artistic and political movements of the first half of the century.

References

Bastian M., S. Heymann, and M. Jacomy (2009). Gephi: an open source software for exploring and manipulating networks. International AAAI Conference on Weblogs and Social Media.

Chase, P. G. (2005). *The Emergence of Culture: The Evolution of a Uniquely Human Way of Life.* Birkhäuser.

Christian, D. (2004). *Maps of time: an introduction to big history.* Heldref Publications.

Cleese, J., and P. Ekman. (2001) *The Human Face.* [BBC] John Cleese, Paul Ekman. http://www.bbc.co.uk/programmes/b00pfqy6

Darwin, C. (1872). *The expression of the emotions in man and animals.* London: John Murray.

Hariri, A. R., et al. (2002). The Amygdala Response to Emotional Stimuli: A Comparison of Faces and Scenes. *NeuroImage* 17: 317–323.

Hu, Y. (2005). Efficient, high-quality force-directed graph drawing. *Mathematica Journal*. mathematica-journal.com.

Michel, J. B., et al. (2011). Quantitative Analysis of Culture Using Millions of Digitized Books. *Science* 331(6014): 176-182.

Ortega y Gasset, J. (1968). *The Dehumanization of Art and Other Essays on Art, Culture, and Literature.* Princeton Paperbacks. 128.

Press notice: https://developers.face.com/signup/?g and http://face.com/blog/facebook-acquires-face-com/ Accessed October, 31st , 2012.

Rizzolatti, G., and L. Craighero, (2004). The mirror-neuron system. *Annual Revision Neuroscience.* annualreviews.org

Rosenberg, A., and J. Hirschberg. (2007). V-measure: A conditional entropy-based external cluster evaluation measure'. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning* held June 2007 in Prague. 410–420.

Sculley, D. (2010). 'Web Scale K-Means clustering'. *Proceedings of the 19th international conference on World wide web.*

Suarez, J. L. (2007). Hispanic Baroque: A Model for the Study of Cultural Complexity in the Atlantic World. *South Atlantic Review* 72(1): 31-47.

Suarez, J. L. (2008). Complejidad y Barroco. *Revista de Occidente.* **323**. 58-74.

Wiskott, L., et al. (1997). 'Face recognition by elastic bunch graph matching. Pattern Analysis and Machine Intelligence'. *IEEE Transactions.* 19(7): 775-779.

Architecture to enable large-scale computational analysis of millions of volumes

Sun, Yiming
yimsun@umail.iu.edu
Indiana University, United States of America

Kowalczyk, Stacy
skowalcz@indiana.edu
Indiana University, United States of America

Plale, Beth
plale@indiana.edu
Indiana University, United States of America

Downie, J. Stephen
jdownie@illinois.edu
University of Illinois at Urbana-Champaign

Auvil, Loretta
lauvil@illinois.edu
University of Illinois at Urbana-Champaign

Capitanu, Boris
capitanu@illinois.edu
University of Illinois at Urbana-Champaign

Hess, Kirk
kirkhess@illinois.edu
University of Illinois at Urbana-Champaign

Peng, Zong
zongpeng@umail.iu.edu
Indiana University, United States of America

Ruan, Guangchen

gruan@umail.iu.edu
Indiana University, United States of America

Todd, Aaron
toddaaro@indiana.edu
Indiana University, United States of America

Zeng, Jiaan
jiaazeng@umail.iu.edu
Indiana University, United States of America

The HathiTrust Research Center (HTRC) is a collaborative research center to provide Digital Humanities researchers access to not only millions of volumes from the HathiTrust (HT) digital library but also cutting-edge software tools and cyberinfrastructure to perform advanced computational analysis over the corpus at an unprecedented scale.

The corpus at the HTRC currently consists of over 3 million public domain volumes, and anticipates access to an additional 6 million in-copyright volumes. In their raw form at the HathiTrust, these volumes are stored as files on special hardware using an internal Pairtree structure. The internal HathiTrust structure is optimal for its primary function of the digital page image delivery to digital library patrons for viewing, however, it does not support well the large-scale computational analysis which is the primary function of the HTRC; navigating the Pairtree and uncompressing the text data would encounter major performance and scalability issues. While researchers from other scientific communities have been addressing aspects of the "Big Data" problem with success, the large corpus that HTRC hosts to support computational analysis presents a unique setting in that it consists of a massive number of small text-based files whereas most solutions from the scientific communities are tailored towards large files and non-text-based content. In this poster, we will present the approach the HTRC takes to solve this problem — the HTRC keeps the Pairtree only for the purpose of synchronization with the HT, and processes and pushes the volume data from the local Pairtree to a NoSQL storage cluster using Apache Cassandra hosted on conventional hardware during the ingest process. In order to balance the data store and ingest workload, the developers at the HTRC and the HT also devised a very simple yet effective way to parallelize the rsync of the single source Pairtree at the HT on all Cassandra nodes by starting rsync at lower branches instead of at the root.

The use of a NoSQL cluster adds more complexity to the architecture than traditional file systems, but such complexity is transparent to the Digital Humanities researchers as most of the HTRC components with which user algorithms have interaction are RESTful web services, such as the Data API for accessing the data. The HTRC uses Blacklight, an open source bibliographic search and display interface, backed by a Solr index, to let users search for volumes for analysis and create collections. To apply analytical techniques to the data, a user may choose from a number of provided algorithms from the web portal, including SEASR/Meandre flows. In addition, the HTRC is actively researching and developing a secure computation environment (dubbed the Sloan Cloud) to support large-scale non-consumptive research over copyrighted volumes, and an experimental release is scheduled for end of March. This Sloan Cloud will allow researchers to deploy their own analysis algorithms against a corpus like the HT data, and to save intermediate data for later reuse, as well as to include custom worksets for the computation. We will present our early findings of the experimental Sloan Cloud and hope to get feedback from the digital humanities research community.

KORA: A Digital Repository and Publishing Platform

Tegtmeyer, Rebecca
REBECCA@RLTDESIGN.NET
Michigan State University, United States of America

Rehberger, Dean
rehberge@msu.edu
Michigan State University, United States of America

MATRIX has developed an open source application that cultural and educational institutions can use to preserve digital materials and display them online. The application, KORA (http://kora.matrix.msu.edu), is particularly well-suited for working with digital objects of all media types and for easily creating displays of these objects in multiple ways that enhance their educational and research value. MATRIX, a humanities computing research center at Michigan State University, has built and enhanced this application in the course of seven years of research with support from the National Science Foundation.

Designed for long-term preservation and access, KORA includes unique features that meet two important needs of institutions that have limited technological resources: (1) simple design of the digital repository and the ingesting of data, and (2) the ability to display digital materials

online in diverse ways, such as image galleries, multimedia educational activities, or story chapters.

The KORA architecture is unique in that it can accommodate any set of metadata schemes (or tables) in individualized digital libraries. Users can easily create metadata elements (database fields) using a simple point-and-click interface, select the type of form control for each element (e.g., required formats for date, URL, file upload, etc.), and then determine whether the element is required for each record, whether it should appear in search returns, and other features. KORA then automatically generates storage structures, ingestion (data entry) forms, and validation requirements for each metadata scheme.

Because the back-end of projects can be created in minutes by people without technical training, the overhead for getting projects started is reduced immeasurably compared to beginning with a blank SQL or other database. And because KORA is an online application, multiple users can develop a collection from separate locations at the same time. Also, KORA can ingest materials from any standardized repository and can output XML that can be harvested by these repositories.

KORA also includes an easy-to-use "associator" tool for creating relationships that combine objects of various media types. As demonstrated by diverse websites built with KORA, many creative displays are possible using this open source application.

In keeping with the need to ensure authenticity and integrity of files ingested into KORA, as described in the International Research on Permanent Authentic Records in Electronic Systems (InterPARES) guidelines, automatic fixity checking has been built into KORA to verify that data has been kept free of tampering and corruption. Long-term access to digital material can be assured by storing this preservation information in the digital repository, as described by the ISO Reference Model for an Open Archival Information System (OAIS) model and Preservation Metadata: Implementation Strategies (PREMIS).

New Release: KORA 3.0

KORA 3.0 will be released this fall with a host of new features, including many major changes:

- all new user experience design;
- independence from MYSQL so it can be used with other database management systems;
- enhanced Multilanguage capabilities;
- rebuilt on Symphony, KORA will have enhanced plugin capabilities;
- and many more features.

Textal: a text analysis smartphone app for Digital Humanities

Terras, Melissa
m.terras@ucl.ac.uk
University College London, Information Studies, United Kingdom

Gray, Steven
steven.gray@ucl.ac.uk
University College London, Centre for Advanced Spatial Analysis

Rudolf, Ammann
ammann@gmail.com
University College London, Information Studies, United Kingdom

This poster introduces Textal, a text analysis application for iOS, and text analysis service infrastructure, which is currently in development at UCLDH and UCLCASA and will be freely available from Summer 2013. This poster will present findings evaluating the development, launch, and reception of the app, indicating how smartphone technology can increase the potential for public engagement within the Digital Humanities.

Textal (soon to be launched at www.textal.org, currently on twitter at @textal) will be a freely available smartphone application which allows users to create, share, and explore word clouds of a document, website, or tweet stream. Those in visualization and Digital Humanities have tended to sneer at the popular use of word clouds (Harris 2011, Meeks 2012), given we are used to applying robust text analysis tools (such as http://voyant-tools.org/). However, Textal turns word clouds into an intuitive, visually-oriented interface: once a Textal of a chosen text is generated, users can click on words to access underlying statistics, such as frequency and collocates, and so we believe that the pinch, stretch, and click potential in smartphones, along with our judicious design, can fix the elements of word cloud visualization which are currently held to be problematic and act as a bridge between those who have never encountered text analysis techniques, and the more detailed approaches undertaken by researchers in Digital Humanities. All Textal visualizations, including word-clouds, graphs, charts, and

word lists, can be shared via social media such as Twitter and Facebook, and the resulting interface word clouds will also be available online. Textal is powered by server-side processing of linguistic data (users can submit any text material they want, by URL, or copy or paste). The resulting sever architecture will also serve as an API for those wishing to carry out on-the-fly generation of text analysis statistics, which can be used in conjunction with other web services.

We envision the Textal iPhone, and iPad, app as a fun text-analysis-in-your-pocket product, which can raise the profile of this technique. We have built Textal with the general audience in mind, to bring Digital Humanities approaches to as wide an international audience as possible (we will be translating the interface into many languages). With an increasing move towards smartphone rather than desktop technologies (Tofel 2012) there is a need to understand how mobile technologies fit within the Digital Humanities remit. We believe we are one of the first teams to build, from scratch, a stand-alone app that brings Digital Humanities techniques to a wider, mobile based, audience. (Previous apps, do exist, such as the DH2012 conference app (https://itunes.apple.com/app/dh2012/id536290090?mt=8), which is an app based version of the conference programme. Geostoryteller is a platform for history walking tours that allow smartphone users to interact with multimedia historical information as they move around a neighbourhood (http://www.geostoryteller.org/index.php , see Rabina and Cocciolo 2012.). Others have used augmented reality viewers for historical and archaeological sites (see http://www.dead-mens-eyes.org/), often built on existing commercial platforms. We don't believe, however, that others have built smartphone apps that allow the user to do much data analysis or processing in the way we describe).

We are building Textal from the ground up using our own server infrastructure, with the app programmed in house in Objective-C. Textal will be available for iOS only, with a plan to build a stand-alone application for use with Apple laptop and desktops. Depending on reception, we may then build an app for other operating systems. Given that we own the infrastructure, we will be able to view and analyse how, why and when people are using text analysis: we will be tracking use and users, including geo-locating text analysis, to ascertain the potential audience for this type of service and to understand more about the kind of texts people want to analyse, allowing us to undertake a reception study into Textal's uptake, which will be of great interest to the wider Digital Humanities audience.

Although the app will not be launched until Summer 2013 this is not a promissory abstract: most of the development, including both technical infrastructure, server architecture, and design-work on the app, is now complete and at time of submission we are moving into alpha-testing with a core group of users interested in text analysis. This poster will be an up-to-the-minute account of a very recent development in Digital Humanities: what ramifications do apps hold for Digital Humanities as a discipline or a field of practice? We will report using up-to-date statistics generated from Textal as a case study, and demonstrate Textal at the poster session.

References

Harris, J. (2011). Word Clouds Considered Harmful. *Nieman Journalism Lab.* http://www.niemanlab.org/2011/10/word-clouds-considered-harmful/ (accessed 13 October 2011.)

Meeks, E. (2012). Using Word Clouds for Topic Modeling Results. *Digital Humanities Specialist blog* https://dhs.stanford.edu/algorithmic-literacy/using-word-clouds-for-topic-modeling-results/. (accessed 15 August 2012).

Rabina, D. L., and A. Cocciolo (2012). *Uncovering lost histories through GeoStoryteller: A digital GeoHumanities project.* Digital Humanities 2012, Hamburg. http://www.dh2012.uni-hamburg.de/conference/programme/abstracts/uncovering-lost-histories-through-geostoryteller-a-digital-geohumanities-project/

Tofel, K. C. (2012). Uh-Oh, PC. Half of Computing Device Sales are Mobile. *Gigaom blog,* http://gigaom.com/mobile/uh-oh-pc-half-of-computing-device-sales-are-mobile/ (accessed 16 January 2012.)

Innovations in Finding Aids and Digital Archives

Thornton, Trevor
trevorthornton@nypl.org
New York Public Library, United States of America

Reside, Doug
dougreside@gmail.com
New York Public Library, United States of America

Despite the ubiquity of digital programs at most research libraries, many archives still document and describe collections in much the same way as they did in the pre-digital era. The development and adoption of the Encoded Archival Description (EAD) XML schema serves as the basis of an infrastructure for dynamic research tools

capable of seamlessly connecting archival materials to the digital information ecosystem. However, many institutions implement EAD as little more than a formatting guide for displays designed to resemble, as closely as possible, the digital finding aids' paper predecessors. If we instead think of EAD as a means of encoding archival description as machine-readable data, we open new possibilities for how our finding aids can be displayed and for exposing them to the ever-growing linked open data environment of the semantic web, ensuring that our content can be found and effectively used by those we hope to serve.

XML documents have several advantages over PDF or Word documents as a vehicle for recording archival description: they "future proof" the finding aid data against changes in display technology; they allow for multiple presentations of the same data; and they allow descriptive data to be harvested by automated agents for purposes other than display. Despite these advantages, EAD is frequently used simply to produce Web documents designed to resemble PDFs of Word Documents, often in such a way as to make the original data unavailable for other purposes. This situation is not due to any inherent shortcoming of the standard, but rather to a failure to make full use of its potential.

If we forget, for a moment, our preconceptions about finding aid design and instead ask ourselves what our researchers want to know about our collections, we will likely find that the 'right' way to present a finding aid depends very much on the needs and intentions of the user. Imagine, for example, if instead of a text-heavy display we generate an Excel-style table in which the specific components displayed could be narrowed as the user types a query string in a search box. On the other hand, with the right styles and visual themes, we could use the descriptive data to construct the sort of "featured collection" sites so many donors support. What if switching between views was as easy as clicking an iTunes-style button that flipped from one option to another to best suit the users' needs and preferences?

In addition to enabling innovative displays, a data-centric approach to EAD facilitates the transformation of archival description into forms that can participate more effectively in the Web environment, particularly in the area of Linked Open Data. The same technologies that enable Google to suggest resources related to searches and Facebook to suggest new friends can be employed to recommend related resources in our own or other institutions that may be of interest to researchers. In addition to bibliographic and archival resources, data from other data sets such as Wikipedia and the Internet Movie Database can be incorporated to provide additional context for our collections and to point readers to more sources of information.

Of course the structure of our data is only half of the story — in order to facilitate these new uses the data itself must be sufficient to effectively establish links to other resources. The philosophy of "more product, less process" has allowed us to make more archival materials available for public use despite shrinking resources, but the primary cost of this efficiency has been the depth of descriptive data produced. Here is another area where we may benefit from allowing our data to participate in the larger Web environment, by taking advantage of the Web's capacity for enabling interaction between users and providers. Consider a finding aid that, in addition to presenting descriptive information, provides access to digital surrogates of the archival materials — a feature that is increasingly common. Among the features of the new generation of finding aids could be tools that allow researchers to provide annotations and access terms associated with the collection as they read and become familiar with the materials. These tools could integrate dynamically with open data sources in order to aid the user in selecting standardized identifiers for names, subjects and titles relevant to the collection. This user-contributed metadata could increase the depth and quantity of our descriptive data with very little investment on the part of the institution, providing a clear benefit to all.

There is very little reason why we can't do most of this right now. Indeed, the New York Public Library is currently experimenting with finding aid interfaces such as those described above as part of several collection-specific projects. In this presentation, Doug Reside (Digital Curator for the Performing Arts) and Trevor Thornton (Senior Applications Developer at NYPL Labs) will demonstrate prototype interfaces for archival collections and reflect on the future of finding aids and digital archives.

Encoding Historical Financial Records

Tomasek, Kathryn
ktomasek@wheatonma.edu
Wheaton College, Norton, Massachusetts, United States of America

Bauman, Syd
Syd_Bauman@brown.edu
Brown University, Providence, Rhode Island, United States of America

A significant number of scholars in Europe and North America are now involved in projects utilizing or encoding historical financial and tabular records. Many of them hope

that it will be possible to develop guidelines that account for both the idiosyncrasies of such manuscripts and the semantic information embedded in them.

A genre of primary sources that includes such materials as bills, receipts, cashbooks, journals, and account ledgers, historical financial records (HFRs) are abundant in traditional archives. Most current digitization projects do not capture the full range of financial information, and if they do, they have yet to develop a common method for fully expressing this range.

HFRs share certain structural characteristics with such other genres of historical records as plague bills, theatre returns, and probate records. Documents from such genres are generally represented as lists or tables, and in many cases they include numerical sums. The apparent regularity of these documents presents perhaps the most significant challenge for those who seek to encode them, as it often collapses in use. Thus such tabular records tend to include information that cannot be represented through simple transcription of tabular layout. In fact, they tend to contain significant variations and idiosyncrasies, often within the same document or collection.

In the subgenre of double entry accounts, the impulse to keep regular records produced a set of standards for recording financial information. Through the centuries, various influential texts offered ordinary businessmen opportunities to learn how to keep regular accounts. But the popularity of these texts did not guarantee perfect adherence to their principles.

HFRs tend to include three levels of data to consider: layout, textual expression, and a third, more abstract level of financial semantics that are not as yet easily captured through TEI conformant markup. Attention to layout may or may not be necessary. In cases where page images are included in online publication, for example, some projects may choose to omit digital representation of layout. Similarly, different projects place varying emphasis on particular textual features.

At the more abstract level, double entry bookkeeping uses a specialized vocabulary, a professional jargon that requires data modeling with attention to the special meanings of the terms "debtor" and "creditor," as well as the relationships between transactions recorded in the journal and accounts kept in a separate ledger. We are developing a TEI customization for conveying such meanings and their expressions within double entry account books through a "transactionography" that will represent the relationships among such records in abstracted form.

As currently conceived a "transactionography," like a "personography," provides information about the financial information within each transaction separately from the transcribed text. "Transactionographies" follow the principles of double entry accounting to model *transactions* as a sequence of one or more *transfers* of anything of value from one *account* to another. Thus, the simple purchase of a candy bar from a convenience store is represented as two `<transfer>`s: one of a candy bar from the vendor's stock account to the buyer, and one of $1.25 from the buyer's cash account to the vendor's cash account.

We believe that this model will be sufficient to represent double entry bookkeeping, though we have not yet tested it thoroughly. We are presenting a (working) ODD file for a first cut at such a "transactionography" at the TEI meeting in fall 2012, and we hope to have a more refined version for presentation at DH2013.

This abstract only begins to suggest the research opportunities that might eventually be available should large numbers of HFRs be digitally accessible in machine processable form. As editors of the Alcalá Account Book Project have noted with regard to their digital edition of the account books of the Royal Irish College of Saint George the Martyr in Alcalá, such records promise "insight into the day-to-day running of the college with valuable information on diet, discipline, and domestic matters." Standardized digitization of HFRs, a rich yet currently inaccessible genre of texts, has the potential to produce harvestable data that could open significant new lines of inquiry about economic, social, and cultural history.

References

The Alcalá Account Book Project http://archives.forasfeasa.ie/index.shtml.

Burnard, L., and S. Bauman (eds.) *TEI P5: Guidelines for Electronic Text Encoding and Interchange.* 2.0.2. 2012-02-02T17:24:24Z. http://www.tei-c.org/P5/.

Gleeson-White, J. (2011). *Double Entry: How the Merchants of Venice Created Modern Finance* New York: Norton.

Mair, J. (1765). *Book-keeping Methodiz'd; or, A methodical treatise of Merchant-accompts, according to the Italian form. Wherein the theory of the art is fully explained,... To which is added, a large appendix. ...* Edinburgh: W. Sands, A. Murray, and J. Cochran.

McCusker, J. J. (2001). *How Much Is That in Real Money?: A Historical Price Index for Use as a Deflator of Money Values in the Economy of the United States* Worcester, MA: American Antiquarian Society.

Pacioli, L. *The Rules of Double-Entry Bookkeeping: Particularis de computis et scripturis.* Michael Schemmann, ed. Orig. pub. 1494. International Institute of Certified Public Accountants, 2012.

Poovey, M. *Genres of the Credit Economy: Mediating Value in Eighteenth- and Nineteenth-Century Britain.* Chicago, Ill.: University of Chicago Press, 2008.

Visible Prices: A Collection of Literary and Historical Economic Information. http://staff.washington.edu/paigecm/vp.

Evaluating Natural Light in Historic Structures through Digital Simulation

VanZee, Lisa
lvanzee@purdue.edu
Purdue University, United States of America

The search for sustainable solutions to current design problems has been an ongoing focus within the architectural and interior design communities. Returning to historical uses of natural resources is one way to solve these problems. Natural light, which can be defined as light from a natural source (sun, moon, atmosphere) is a resource that has advantages and disadvantages, both aesthetically and physically, in terms of use as a valuable light source to illuminate spaces. The goal of this research was to address how past and present structures have utilized natural light as an effective lighting solution, and then apply these methods to future structures by means of digital simulation and quantifiable data. Through the analysis of exterior façade systems, fenestration design, interior spaces, and building orientation, past and present structures can help provide a complete picture of both successful and unsuccessful daylighting solutions.

Current methods to evaluate daylight prior to the design phase of a building project are wide-ranging and not standardized nor regulated. To effectively study daylighting methods, computer simulation models needed to be developed of past and present structures. Measuring the amount of light from an electric or artificial source can often be straightforward, but daylight is highly variable with many factors to consider and there is not one universal method to calculating quality or quantity. Lighting calculation tools must be able to account for these variables in order to provide an effective lighting solution that can both reduce energy and ensure adequate light levels. The methodology and procedures for this research was conducted as follows: structure selection; three-dimensional computer models of selected structures; physical site visits where applicable; daylight simulation software selection; exterior and interior lighting analysis. Autodesk Ecotect was the software chosen for this research study, which is billed as sustainable design analysis software that offers a variety of simulation and building functionalities.

Data extracted from the digital simulations focused on the amount of interior illuminance in footcandles (one footcandle is equal to the illumination produced by one candle at a distance of one foot), as well as the amount of solar radiation (sunlight) available at each site, average temperatures, and solar geometry (path of travel of the sun). The software data shows that the natural daylight within the buildings as a sole light source can provide adequate illumination levels for tasks completed within the space during daylight hours under varying sky conditions. Site visits were conducted with physical light meters to determine the accuracy of the digital simulations. The techniques employed in the historic structures to control natural light (oculus windows, clerestory windows, site orientation, roof overhangs) were then applied to a current building environment, which was additionally analyzed on daylighting implementation.

Analyzing the success of daylight as a light source before physical construction is difficult in the design industry because there are various units of measurement when dealing with lighting that pertains to different factors (such as climate, weather conditions, time of year). These types of variables can produce a large amount of data in digital simulations, which can alter the perception of the data analysis, where the lighting levels might be extremely low during a winter storm or extremely high at noon in the summer. Determining which data to use and analyze was a large portion of the research in this project. Although standards exist today to attempt to measure the quality and quantity of light within a space, such as illuminance level (footcandle) recommendations and daylight factor ratios (the amount of light available indoors verses outdoors), there is still a great deal of trial and error when designing for daylight. The results of the study were used to verify that the use of current technology, in the format of digital simulation and analysis, on existing historic structures can be reliable indicator of the success of natural lighting solutions for future buildings. From the perspective of energy savings, environmental benefit and occupant comfort, the need for additional studies and research in daylighting metrics in buildings is needed to for natural light to become a viable lighting source.

References

Bhavani, R. G., and M. A. Khan (2011). Advanced lighting simulation tools for daylighting purposes: Powerful features and related issues. *Trends in Applied Sciences Research* 6 345-363.

Galasiu, A. D., and M. R. Atif (2002). Applicability of daylighting computer modeling in real case studies: Comparison between measured and simulated daylight availability and lighting consumption. *Building Environ.,* 37. 363-377.

Reinhart, C. F., and A. Fitz (2006). Findings from a survey on the current use of daylight simulations in building design. *Energy and Buildings*, 38. 824-835.

Seward, A. (2011). Light meter. *Eco-structure.* 9. 21-24.

Webb, A. (2006). Considerations for lighting in the built environment: Non-visual effects of light. *Energy and Buildings,* 38. 721-727.

"Making the Digital Humanities More Open": Modeling Digital Humanities for a Wider Audience

Visconti, Amanda

amandavisconti@gmail.com
Maryland Institute for Technology in the Humanities, University of Maryland; Department of English, University of Maryland

Guiliano, Jennifer

jenguiliano@gmail.com
Maryland Institute for Technology in the Humanities, University of Maryland

Smith, James

jgsmith@gmail.com
Maryland Institute for Technology in the Humanities, University of Maryland

Williams, George

georgehwilliams@gmail.com
Department of Languages, Literature, and Composition, University of South Carolina Upstate

Bohon, Cory

corybohon@gmail.com
Maryland Institute for Technology in the Humanities, University of Maryland

"Making the Digital Humanities More Open", a NEH ODH Digital Humanities Start-Up Grant Project, is creating a free and easy-to-use tool that enables end-users with a variety of disabilities and abilities to access online humanities resources, allowing digital humanities projects to share the products of their text-based projects with this often-neglected audience of readers. During the year previous to DH 2013, our team will design and deploy a WordPress-based accessibility tool that will create braille content for end-users who are blind or low vision. Specifically, we plan to extend the use of Anthologize — a free and open source plug-in for WordPress that currently translates any RSS text into PDF, ePub, HTML, or TEI — to include the conversion of text to braille. As a result, we will not only make it easy for digital research content creators to convert a text into braille, thereby extending humanities content to hundreds of thousands of visually disabled readers, but we will also experiment with making braille available visually through the WordPress interface. This tool will also make it possible to translate the textual content of an Omeka archive into braille provided the site — like most that use Omeka — publishes an RSS feed; we'll have conducted initial tests of the tool by translating the existing *BrailleSC* oral histories into braille, and then we will reach out to other Omeka- and WordPress-based humanities computing projects to ask for their cooperation and collaboration in translating their content.

Over the last several decades, scholars have developed standards for how best to create, organize, present, and preserve digital information so that future generations of teachers, students, scholars, and librarians may still use it. What has remained neglected for the most part, however, are the needs of people with disabilities (Abou-Zahra). As a result, many of the otherwise most valuable digital resources are useless for people who are blind or have low vision; the barriers to participation are varied and include such obstacles as the high price of specialized software and hardware, the advanced expertise that such software and hardware often requires, and design choices that can prevent end-users with sensory disabilities from taking full advantage of online resources (Fox).

For public humanities practitioners, including the many local museums, small organizations, and individual scholars, our project provides an entrance into digital humanities communities that might otherwise be obfuscated by the high costs of technology adoption, customization, and deployment. Accessibility is important because disabled users need to be able to participate fully in humanities research and teaching. In providing accessibility tools to disabled communities, we enrich their individual research and learning efforts beyond the formal educational process. As the insights of scholars working in disability studies in the humanities have shown, creating tools for individuals

with disabilities improves digital environments for all users (Williams). Our work aims to increase participation by all people in experiencing and creating scholarly digital projects.

The deliverables of the project, all completed by May 2013, will be a public GitHub repository of the code, a documentation guide, a pilot test of using a local LibLouis library with WordPress, a pilot test of using a remote LibLouis library with WordPress and Anthologize, and a white paper explaining what we've learned about various options for creating online braille documents. A demo of the technical components of our work will be part of our poster presentation, but we will emphasize the theoretical takeaways of the project, modeling the ways in which digital humanities projects should be designed and implemented with the needs of disabled users in mind. We will discuss what we have learned about accessibility for both digital humanities projects and digital writing in general, and explore both the direct impact of this tool and the significance of opening digital humanities work to a wider audience of participants. Our poster will be of interest to anyone interested in the theory behind how we code and design digital humanities tools and site, designing for universal and accessible use, and accessing, preserving, and working with the cultural histories of people interacting with the braille form of reading and writing.

References

Abou-Zahra, S. (ed). (2011). Evaluating Web Sites for Accessibility: Overview. Web Accessibility Initiative. World Wide Web Consortium. http://www.w3.org/WAI .

Fox, S. Americans Living with Disability and Their Technology Profile. Pew Research Center's Internet & American Life Project. http://pewinternet.org/Reports/2011/Disability.aspx. (accessed 21 January 2011).

Williams, G. H. (2012). Disability, Design, and the Digital Humanities. In **Gold, M. K.** (ed.), *Debates in the Digital Humanities*. Minneapolis: University of Minnesota Press. 202-212.

TEI Boilerplate

Walsh, John
jawalsh@indiana.edu
Indiana University, United States of America

Simpson, Grant Leyton
glsimpso@indiana.edu
Indiana University, United States of America

Introduction

TEI Boilerplate (http://teiboilerplate.org/) is a lightweight, HTML5 compliant framework for publishing TEI documents. TEI Boilerplate (TEIBP) is designed to bridge the gap between the browser-friendly features of HTML and the semantic richness of native TEI documents (Walsh, Simpson, & Moaddeli, 2012).

Although TEI provides mechanisms for describing the design, presentational, and material features of the source document, projects and individual scholars that use TEI are responsible for developing their own methods, or implementing existing solutions, for converting the TEI to a presentation-ready state for the web or print (Rahtz, 2006). Two potential paths to reach this goal are:

1. Transforming TEI to HTML using XSLT and styling the HTML output with CSS.
2. Styling the TEI directly with CSS by referencing a CSS stylesheet from within the TEI document.

Both approaches have advantages and disadvantages. Although HTML is the language of the web and is well supported by browsers, HTML's descriptive capabilities are less expressive than TEI's. When TEI is transformed to HTML, much of the richness of the TEI is lost in the resulting HTML. However, the browser understands HTML very well and knows, for example, when to initiate retrieval of a document based on certain user events, such as clicking a link. The second option, CSS-styled TEI, delivers the TEI document directly] to the browser. However, while the browser may apply CSS to format and style a TEI document, the browser does not understand the semantics of TEI. For instance, the browser does not understand that TEI's `<ptr>` and `<ref>` elements are linking elements.

TEIBP bridges the gap between these two approaches by making use of the built-in XSLT (1.0) capabilities of browsers to embed the TEI XML, with minimal modifications, within an HTML5 shell document. Features expected of web documents, such as clickable links and display of linked images, are enabled through selective transformation of a very small number of TEI elements and attributes. Both the HTML5 shell and the embedded TEI are styled using CSS.

TEIBP gives HTML/CSS/JavaScript documents direct access to original TEI content, and it gives TEI documents direct access to the substantial capabilities of HTML, CSS, and JavaScript — the dominant document format, styling language, and (client-side) programming language of the web. TEIBP aims for simplicity and elegance, but it also facilitates complexity and innovation by exposing TEI

content directly to the capabilities of JavaScript, the many powerful JavaScript frameworks, and CSS.

Theoretical Motivations

The power of TEI lies in the richness of its vocabulary. But much of that richness and expressiveness is lost in the translation to HTML. TEIBP largely preserves the integrity of the TEI document. Because the TEI document is delivered directly to the browser, that source TEI document — unchanged by any XSLT transformation — can be easily accessed and saved.

Scholars labor over the intricate encoding of TEI documents, encoding that may represent sophisticated readings and analysis. But with the typical XSLT publishing solution, much or all of the richness of the TEI content is lost. Furthermore, the presentation of the document is targeted at the HTML surrogate rather than the encoded TEI document. This results in a conceptual disconnect between the design of the document and the original TEI encoding. By exposing the TEI itself to the browser, one may format the TEI directly, applying intentional design to a sophisticated document model.

TEIBP respects the integrity of the TEI document, and keeps the TEI document central throughout the publication process. TEIBP takes advantage of the separation of form and content inherent in XML, XSLT, CSS frameworks. However, like Liu (2004), Galey (2010), and others, the authors of TEIBP view that separation with suspicion. TEIBP attempts to weaken that separation of form and content in the typical TEI-to-web design and delivery model by largely removing the HTML layer, exposing the TEI-encoded text directly to the browser, and providing scholars with more immediate access to the readings, models, and analysis embedded in the TEIencoded document.

Our proposed poster will provide an overview of the TEIBP system and explore in more detail the theoretical motivations behind the project.

References

Galey, A. (2010). The human presence in digital artifacts. In W. McCarty (ed), *Text and genre in reconstruction: effects of digitization on ideas, behaviours, products, and institutions* 93–117. Oxford: Open Book.

Liu, A. (2004). Transcendental data: Towards a cultural history and aesthetics of the new encoded discourse. *Critical Inquiry,* 31: 49–84.

Omeka (2012). Omeka: Serious web publishing. http://omeka.org/about/ .

Rahtz, S. (2006). Storage, retrieval, and rendering. In Burnard, L., K. O'Brien O'Keeffe, and J. Unsworth (eds.), *Electronic textual editing* 310–333. New York, NY: Modern Language Association of America.

Walsh, J., G. Simpson, and S. Moaddeli (2012). TEI Boilerplate. http://teiboilerplate.org .

Juxta Commons

Wheeles, Dana
dana@nines.org
University of Virginia, United States of America

Jensen, Kristin
kristin@performantsoftware.com
Performant Software, United States of America

Free and open source, Juxta Commons (juxtacommons.org) is an online workspace for comparing multiple witnesses to a single textual work, privately storing collations, and sharing visualizations.

Originally offered as a downloadable Java-based application developed by the Applied Research in 'Patacriticism group at the University of Virginia in 2005, Juxta was taken up by NINES (http://nines.org) and transformed into a web service. This open-source API modularizes the sequence of steps required for digital collation and offers more options for working with XML documents. In addition, the NINES R&D team created the interface for Juxta Commons, a destination site for collation on the web.

Juxta Commons offers three visualizations of the differences between a group of texts: the heat map (an overlay of the texts with differences highlighted by color), the side-by-side view (two texts visualized with lines connecting the sites of difference) and the histogram (a global view of the text illustrating the portions with the most change across versions). It is also compatible with TEI Parallel Segmentation, so that users can upload their encoded files and make use of Juxta's visualizations, or those with text or XML files can export their collations as a digital critical apparatus.

The interface streamlines the workflow of the original desktop application, and allows the user to edit their sources, filter the XML content included in a given collation, and share their results in a number of different ways. Even though the tool provides sophisticated options for working with texts (XML and TXT), our goal in designing Juxta Commons was to allow teachers, scholars, programmers and any other users curious about variants between documents the ability to create, collate and share

their findings with others without requiring any extra software downloads or logins.

In taking Juxta to the web, NINES has transformed the tool into a rhetorical as well as analytical tool, enabling initial discoveries to be interrogated, explored, and ultimately offered as a new kind of evidence for scholarly arguments. Instead of relying upon footnotes or links in a digital edition, a scholar can embed their collations within the argument itself, targeting sites of interest, and encouraging readers to explore the full visualizations on their own.

Even as it presents some exciting opportunities for students of book history and textual criticism, Juxta Commons also broadens the scope of interest in collation to the many different kinds of texts on the web. Speeches and transcriptions, Wikipedia article versions, news releases, Google Books and HATHI Trust texts: all these require the kind of authentication and analysis that Juxta Commons makes possible.

Please visit the site juxtasoftware.org to learn more, and to find sample sets to spark your creativity.

Surveying a Corpus with Alignment Visualization and Topic Modeling

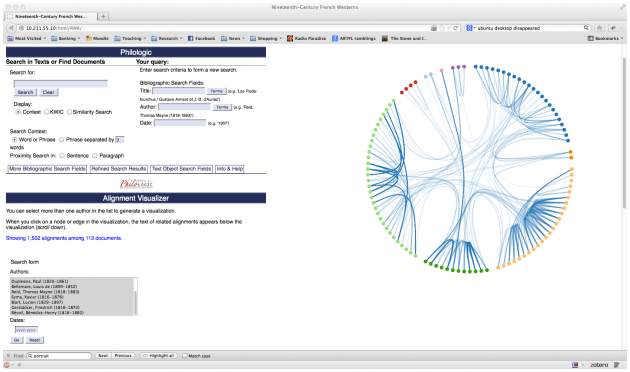
Wolff, Mark
wolffm0@hartwick.edu
Hartwick College, United States of America

With the advent of sophisticated computational tools for analyzing large corpora of literary texts, scholars have proposed a variety of terms to describe what they do when they use these tools. Text data mining (Hearst), distant reading (Moretti), algorithmic criticism (Ramsay), assisted reading (Schmidt), hyper reading (Hayes), faceted search (Whaling), scalable reading (Mueller): these seek to capture a shift in how digital technology allows readers to observe patterns within texts and to use the findings to inform a closer reading. N. Katherine Hayes summarizes the affordances of computational tools as methods of scanning, which quickly seek specific words within texts, and methods of skimming, which attempt to get a rapid sense of what texts are about (61). Scanning is best served by text retrieval, whereas skimming requires techniques like corpus queries and topic modeling.

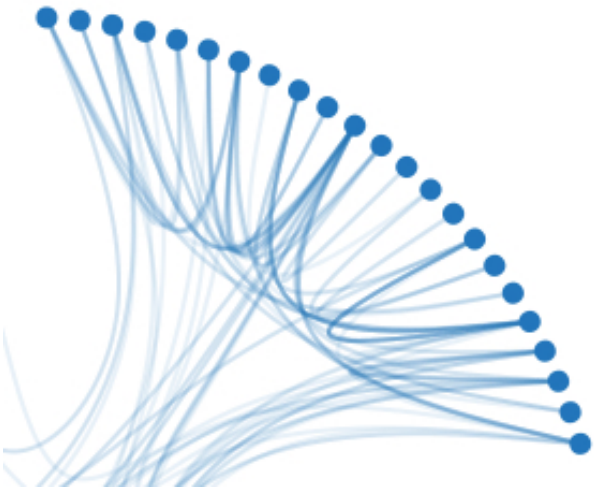
To add to the lexicon of digital text analysis, I propose the idea of surveying a corpus using a combination of text retrieval for scanning and data visualization for skimming. Stan Rucker, Milena Radzikowska and Stéfan Sinclair have articulated a theory of data visualization based on the notion of prospect, "a view of the world where enough information is available for the perceiver to understand the terrain and have a sense of what it affords, without necessarily seeing all the details" (26). Surveying a corpus relies on sufficient prospect to enable the reader to see textual information from a distance and to perceive patterns on a scale larger than a single text. The ability to skim a corpus for the purpose of surveying it would ideally allow the reader to easily zoom from a comprehensive view of patterns to specific instances within texts where scanning for specific words may be desired. Conversely, the results of text retrieval searches should be available for constructing corpus queries: if a particular string appears in one or more texts, a reader surveying a corpus may want to get a broader sense of how the string is used without skimming through a long KWIC listing.

To show what I mean by surveying a corpus, I will offer a demonstration of a database built with ARTFL's Philologic that includes a visualizer for sequence alignments and a topic model browser for ordered lists of documents, words and algorithmically derived topics (see http://bumppo.hartwick.edu/rw/). The corpus is comprised of 181 adventure novels published in French during the nineteenth century. The digital texts were created recently by the Bibliothèque de France using OCR and they contain no markup. Using plaintext documents with minimal metadata, I am able to survey the corpus to get a sense of its themes, recurrent rhetorical devices, innovators, and imitators. This approach is proving particularly helpful for studying a corpus that has received relatively little critical attention. It does not require significant resources for text preparation, and it allows for exploring an obscure corpus to determine if further research is warranted.

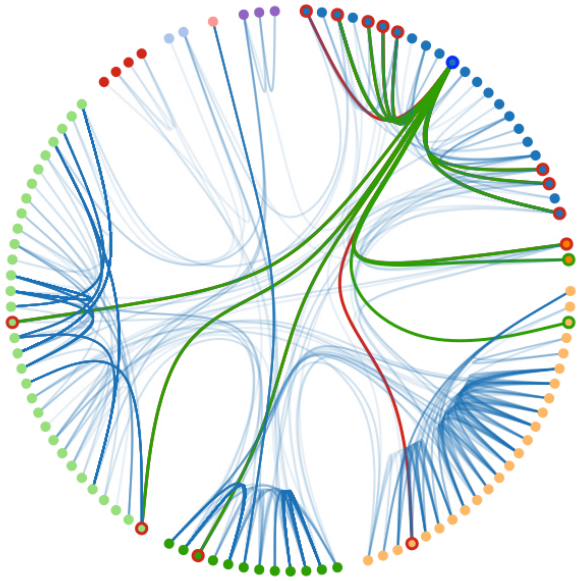
I first identified sequence alignments among texts in the corpus using ARTFL's Philoline module for Philologic. Sequence alignment is the identification of common n-grams shared between two or more documents (Horton et al). The technique, developed for DNA analysis and plagiarism detection, identifies segments of text that have been "borrowed" or reproduced from one document to another. The alignment visualizer, based on an implementation of Danny Holton's hierarchical edge bundling algorithm and built with Michael Bostock's D3 JavaScript library, displays texts as a circle of nodes with edges indicating which texts share common passages (Fig. 1).



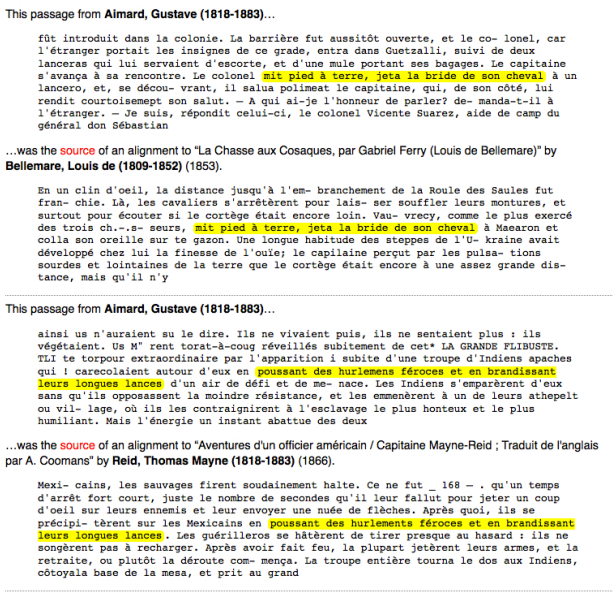
Nodes are grouped by author, and authors are distinguished by colors. The number and thickness of the edges reflect the number of alignments between texts: a node with many edges represents a text that shares alignments with many other texts, and thicker edges mean more alignments between two texts. A "fan" among nodes of a similar color shows numerous alignments among texts by the same author (Fig. 2).



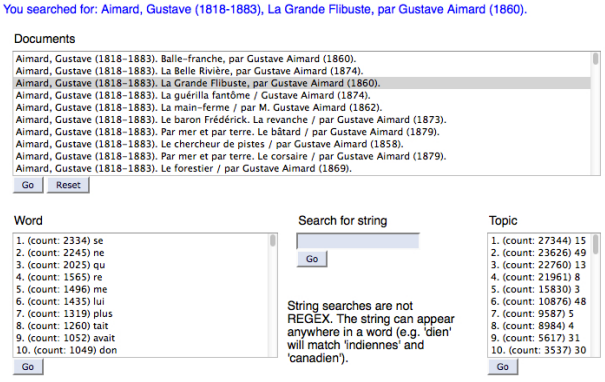
The CSS stroke attribute of edges and nodes changes color when the reader moves the pointer over them (red strokes designate source nodes, green strokes target nodes, blue strokes nodes that are selected by the pointer) (Fig. 3).



The reader can click on an edge or node and immediately examine the associated aligned passages. The selected node in Figure 3 represents Gustave Aimard's *La Grande Flibuste* (1860), a text that shares many alignments with other texts by Aimard and five other authors. The graphic suggests that the text is an important one in the corpus, both as a source and target of aligned passages. The visualizer allows the reader to skim the corpus to get a distant view of what it contains. Figure 4 shows two particular alignments between Aimard's novel and novels by Louis de Bellemare and Mayne Reid. In examining alignments like these, one can contextualize them and get a sense of how *La Grande Flibuste* is representative of the corpus.



The topic model browser is a JavaScript interface that allows the reader to query the frequencies of documents, words and topics against each other in a corpus. The topics were generated using Andrew McCallum's Mallet toolkit and then migrated to a MySQL database where each word in each text is assigned to one of fifty modeled topics (Fig. 5).



Beginning with a list of documents, the reader can select a text and see lists of words and topics in the text in descending order. From these query results the reader can select any item in any list as the parameter for a subsequent query, which will produce new, ordered lists of documents, words and topics. A basic string search feature (without regular expressions) will also generate ordered lists. In scanning the corpus for words and topics related to *La Grande Flibuste*, the reader can see how the text relates to other texts and explore the relationships between words and topics.

As prototypes, both the alignment visualizer and the topic model browser are currently limited in their capacity to survey a large corpus. Numerous texts are difficult to represent as bundled nodes in a web browser, and the reader may only be able to skim a subset of the corpus instead of its entirety. My implementation of topic modeling uses the default algorithm for Mallet and does not afford the reader any control over or explanation of the algorithm's implementation. It also takes a long time to execute queries on a MySQL database of 7,500,000 tokens and their assigned topics. Alternative data structures such as key-value stores may increase speed.

Despite these limitations, the database interface I present here enables the reader to survey a corpus with enhanced prospect. Alignment visualization and topic model browsing offer two methods for skimming a corpus for patterns, which may lead to further insights for interpretation.

References

Bostock, M., V. Ogievetsky, and J. Heer (2011). D3: Data-Driven Documents. *IEEE Transactions on Visualization and Computer Graphics* (October 2011).

Hayes, N. K. (2012). *How We Think: Digital Media and Contemporary Technopoesis*. University of Chicago Press.

Hearst, M. A. (1999). Untangling Text Data Mining. In *Proceedings of ACL '99: the 37th Annual Meeting of the Association for Computational Linguistics* held June 20-26, 1999 at the University of Maryland.

Holton, D. (2006). Hierarchical Edge Bundles: Visualization of Adjacency Relations in Hierarchical Data. *IEEE Transactions on Visualization and Computer Graphics* 12(5). (September/October 2006).

Horton, R., M. Olsen, and G. Roe (2010). Something Borrowed: Sequence Alignment and the Identification of Similar Passages in Large Text Collections. *Digital Studies/ Le champ numérique* 2(1).

McCallum, A. K. (2002). MALLET: A Machine Learning for Language Toolkit.

Moretti, F. (2000). Conjectures on World Literature. *New Left Review.* 1: 54-68.

Mueller, M. (2012). Scalable Reading. *Scalable Reading*. 29 May 2012. https://scalablereading.northwestern.edu/scalable-reading/

Ramsay, S. (2011). *Reading Machines: Toward an Algorithmic Criticism*. Univeristy of Illinois Press.

Ruecker, S., M. Radzikowska, and S. Sinclair (2011). *Visual Interface Design for Digital Cultural Heritage: A Guide to Rich-Prospect Browsing*. Surrey: Ashgate.

Schmidt, B. (2010). Assisted Reading vs. Data Mining. *Sapping Attention*. 30 December 2010. http://sappingattention.blogspot.com/

Whaling, R. (2011). Faceted Search for a Corpus Query System. *2011 Chicago Colloquium on Digital Humanities and Computer Science*. 21 November 2011. http://chicagocolloquium.org/dhcs-2011-program/

Author Index