Digital Humanities 2008

University of Oulu, June 24-29

Book of Abstracts

The Association for Literary and Lingustic Computing

The Association for Computers and the Humanities

Society for Digital Humanities — Société pour l'étude des médias interactifs

Digital Humanities 2008

The 20th Joint International Conference of the Association for Literary and Linguistic Computing, and the Association for Computers and the Humanities

and

The 1st Joint International Conference of the Association for Literary and Linguistic Computing, the Association for Computers and the Humanities, and the Society for Digital Humanities — Société pour l'étude des médias interactifs

University of Oulu, Finland

24 – 29 June, 2008

Conference Abstracts

UNIVERSITY of OULU

OULUN YLIOPISTO

International Programme Committee

- Espen Ore, National Library of Norway, Chair
- Jean Anderson, University of Glasgow, UK
- John Nerbonne, University of Groningen, The Netherlands
- Stephen Ramsay, University of Nebraska, USA
- Thomas Rommel, International Univ. Bremen, Germany
- Susan Schreibman, University of Maryland, USA
- Paul Spence, King's College London, UK
- Melissa Terras, University College London, UK
- Claire Warwick, University College London, UK, Vice Chair

Local organizers

- Lisa Lena Opas-Hänninen, English Philology
- Riikka Mikkola, English Philology
- Mikko Jokelainen, English Philology
- Ilkka Juuso, Electrical and Information Engineering
- Toni Saranpää, English Philology
- Tapio Seppänen, Electrical and Information Engineering
- Raili Saarela, Congress Services

Edited by

- Lisa Lena Opas-Hänninen
- Mikko Jokelainen
- Ilkka Juuso
- Tapio Seppänen

ISBN: 978-951-42-8838-8

Published by

English Philology

University of Oulu

Cover design: Ilkka Juuso, University of Oulu

© 2008 University of Oulu and the authors.

Introduction

On behalf of the local organizers I am delighted to welcome you to the 25th Joint International Conference of the Association for Literary and Linguistic Computing (ALLC) and the Association for Computers and the Humanities (ACH) at the University of Oulu.

This is a very special year because this is also the 1st Joint International Conference of the ALLC, ACH and the Society for Digital Humanities – Société pour l'étude des medias interactifs, as this year the SDH-SEMI joined the Alliance of Digital Humanities Organizations (ADHO), and I wish to extend a special welcome to them.

With over 16 000 students and 3000 staff members, the University of Oulu is one of the largest universities in Finland. Linnanmaa, the campus area where the conference is being held, not only houses the University, but also Technopolis, the largest technology centre in Finland, and VTT, the State Technical Research Centre. We hope that this strong influence of technology and the opportunities it affords for multidisciplinary research will be of particular interest to the conference delegates.

Many people have contributed to putting together the conference programme. I wish to thank the International Programme Committee and particularly its Chair, Espen Ore, for all the hard work that they have put into making this conference a varied and interesting one. I also wish to thank John Unsworth and Ray Siemens, who passed on their experiences of the 2007 conference and helped us set up the conference management site.

I wish to thank Riikka Mikkola, Mikko Jokelainen, Ilkka Juuso and Toni Saranpää, who have all contributed to making this conference happen. I also thank Tapio Seppänen of the Engineering Faculty for all his help and support.

I hope you will find the conference stimulating and useful and that you enjoy meeting the fellow humanities computing scholars from all over the world. If this is your first time in northern Finland, I hope you like it and that it entices you to come and visit us again.

Tervetuloa! — Welcome!

Lisa Lena Opas-Hänninen

Local Organizer

A Richness of Research and Knowledge

One of the bad things one may experience in the Programme Committee for a conference under the ADHO umbrella, is that there are more good submissions than there is room for. As we read through the submissions for this year's conference, it became clear that Humanities Computing or Digital Humanities as a field is growing up. The field is also growing, period. Ten years ago I was also Chair of the Programme Committee for the ALLC/ACH conference in Debrecen, Hungary. I am happy that there are now so many new names among the submittors. And there were many very good submissions for papers, panel and paper sessions and posters. The hosting institution, the University of Oulu, could luckily supply the rooms needed if we were to run up to four parallel sessions. With four parallel sessions it may seem that there is a larger possibility for a conflict of interests: one might simultaneously wish to attend two presentations - or more. But the only way to remove this problem would be to have a single strand conference with a duration of a couple of weeks. So since this, as usual, is a multiple strand conference, I hope the Book of Abstracts may be used as a guide for which presentations to listen to and, alternatively, which authors one might seek out for an informal talk during one of the breaks.

I would like to thank all the members of the PC, and especially the Vice Chair, Claire Warwick, for all their work, which has made the work on the Programme much easier than it could have been. I would also like to thank both Sara Schmidt at the University of Illinois and the local administration group at the University of Oulu who have done a tremendous amount of work in preparing for the conference programme. And last but definitely least, my thanks go to the Local Organizer, Lisa Lena Opas-Hänninen, who has been a great aid in the programme work and who has done everything I can imagine to help us get the rooms, people and more sorted out. But there are also those without whom there would be no conference. First, all of you who have prepared submissions, and then all of you who are attending this conference: thank you very much, and welcome to Oulu!

Espen S. Ore

Chair, International Programme Committee

Table of Contents

Panels

Papers

Posters

Service Oriented Computing in the Humanities 3 (SOCH3)

One-Day Workshop at Digital Humanities 2008

Tuesday, 24 June, 2008 9:30 – 17:30

Organized jointly by King's College London, and the University of Oulu

The workshop is organized by Stuart Dunn (Centre for e-Research, King's College), Nicolas Gold (Computing Science, King's College), Lorna Hughes (Centre for e-Research, King's College), Lisa Lena Opas-Hänninen (English Philology, University of Oulu) and Tapio Seppänen (Information Engineering, University of Oulu).

Speakers

John Bradley
john.bradley@kcl.ac.uk
King's College London, UK

Elpiniki Fragkouli
elpiniki.fragkouli@kcl.ac.uk
King's College London, UK

Allen Renear
renear@uiuc.edu
University of Illinois, USA

Monica Schraefel
mc@ecs.soton.ac.uk
University of Southampton, UK

Tapio Seppänen
tapio.seppanen@oulu.fi
University of Oulu, Finland

Since software services, and the innovative software architectures that they require, are becoming widespread in their use in the Digital Humanities, it is important to facilitate and encourage problem and solution sharing among different disciplines to avoid reinventing the wheel. This workshop will build on two previous Service Oriented Computing in the Humanities events held in 2006 and 2007(under the auspices of SOSERNET and the AHRC ICT Methods Network). The workshop is structured around four invited presentations from different humanities disciplines. These disciplines are concerned with (e.g.) archaeological data, textual data, the visual arts and historical information. The presentations will approach humanities data, and the various uses of it, from different perspectives at different stages in the research lifecycle. There will reflection on the issues that arise at the conception of a research idea, through to data gathering, analysis, collaboration and publication and dissemination. A further presentation from Computer Science will act as a 'technical response' to these papers, showcasing different tool types and how they can be applied to the kinds of data and methods discussed.

The presentations will be interspersed with scheduled discussion sessions and The emphasis throughout is on collaboration, and what the humanities and computer science communities can learn from one another: do we have a common agenda, and how can we take this forward? The day will conclude with a moderated discussion that will seek to identify and frame that agenda, and form the basis of a report.

Panels

The Homer Multitext Project

Casey Dué
casey@chs.harvard.edu
University of Houston, USA

Mary Ebbott
ebbott@chs.harvard.edu
College of the Holy Cross, USA

A. Ross Scaife
scaife@gmail.com
University of Kentucky, USA

W. Brent Seales
seales@netlab.uky.edu
University of Kentucky, USA

Christopher Blackwell
cwblackwell@gmail.com
Furman University, USA

Neel Smith
dnsmith.neel@gmail.com
College of the Holy Cross, USA

Dorothy Carr Porter
dporter@uky.edu
University of Kentucky, USA

Ryan Baumann
rfbaumann@gmail.com
University of Kentucky, USA

The Homer Multitext Project (HMT) is a new edition that presents the *Iliad* and *Odyssey* within the historical framework of its oral and textual transmission.

The project begins from the position that these works, although passed along to us as written sources, were originally composed *orally* over a long period of time. What one would usually call "variants" from a base text are in fact evidence of the system of oral composition in performance and exhibit the diversity to be expected from an oral composition. These variants are not well reflected in traditional modes of editing, which focus on the reconstruction of an original text. In the case of the works of Homer there is no "original text". All textual variants need to be understood in the historical context, or contexts, in which they first came to be, and it is the intention of the HMT to make these changes visible both synchronically and diachronically. The first paper in our session is an introduction to these and other aspects of the HMT, presented by the editors of the project, Casey Dué and Mary Ebbott. Dué and Ebbott will discuss the need for a digital Multitext of the works of Homer, the potential uses of such an edition as we envision it, and the challenges in building the Multitext.

The works of Homer are known to us through a variety of primary source materials. Although most students and scholars of the classics access the texts through traditional editions, those editions are only representations of existing material. A major aim of the HMT is to make available texts from a variety of sources, including high-resolution digital images of those sources. Any material which attests to a reading of Homer – papyrus fragment, manuscript, or inscription in stone – can and will be included in the ultimate edition. The HMT has begun incorporating images of sources starting with three important manuscripts: the tenth-century Marcianus Graecus Z. 454 (= 822), the eleventh-century Marcianus Graecus Z. 453 (= 821), and the twelfth/thirteenth-century Marcianus Graecus Z. 458 (= 841). Marcianus Graecus Z. 454, commonly called the "Venetus A," is arguably the most important surviving copy of the *Iliad*, containing not only the full text of the poem but also several layers of commentary, or *scholia*, that include many variant readings of the text. The second paper in our session, presented by project collaborators W. Brent Seales and A. Ross Scaife, is a report on work carried out in Spring 2007 at the Biblioteca Nazionale Marciana to digitize these manuscripts, focusing on 3D imaging and virtual flattening in an effort to render the text (especially the *scholia*) more legible.

Once text and images are compiled, they need to be published and made available for scholarly use, and to the wider public. The third paper, presented by the technical editors Christopher Blackwell and Neel Smith, will outline the protocols and software developed and under development to support the publication of the Multitext.

Using technology that takes advantage of the best available practices and open source standards that have been developed for digital publications in a variety of fields, the Homer Multitext will offer free access to a library of texts and images, a machine-interface to that library and its indices, and tools to allow readers to discover and engage with the Homeric tradition.

References:

The Homer Multitext Project, description at the Center for Hellenic Studies website: http://chs.harvard.edu/chs/homer_multitext

Homer and the Papyri, http://chs.harvard.edu/chs/homer___the_papyri_introduction

Haslam, M. "Homeric Papyri and Transmission of the Text" in I. Morris and B. Powell, eds., *A New Companion to Homer.* Leiden, 1997.

West, M. L. *Studies in the text and transmission of the Iliad.* München: K.G. Saur 2001

Digital Images of Iliad Manuscripts from the Marciana Library, *First Drafts @ Classics@*, October 26, 2007: http://zeus.chsdc.org/chs/manuscript_images

Paper 1: The Homer Multitext Project: An Introduction

Casey Dué and Mary Ebbott

The Homer Multitext (HMT) of Harvard University's Center for Hellenic Studies (CHS) in Washington, D.C., seeks to take advantage of the digital medium to give a more accurate visual representation of the textual tradition of Homeric epic than is possible on the printed page. Most significantly, we intend to reveal more readily the oral performance tradition in which the epics were composed, a tradition in which variation from performance to performance was natural and expected. The Homeric epics were composed again and again in performance: the digital medium, which can more readily handle multiple texts, is therefore eminently suitable for a critical edition of Homeric poetry—indeed, the fullest realization of a critical edition of Homer may require a digital medium. In other words, the careful construction of a digital library of Homeric texts, plus tools and accompanying material to assist in interpreting those texts, can help us to recover and better realize evidence of what preceded the written record of these epics. In this presentation we will explain how the oral, traditional poetry of these epics and their textual history call for a different editorial practice from what has previously been done and therefore require a new means of representation. We will also demonstrate how the Homer Multitext shares features with but also has different needs and goals from other digital editions of literary works that were composed in writing. Finally, we will discuss our goal of making these texts openly available so that anyone can use them for any research purpose, including purposes that we can't ourselves anticipate.

The oral traditional nature of Homeric poetry makes these texts that have survived different in important ways from texts that were composed in writing, even those that also survive in multiple and differing manuscripts. We have learned from the comparative fieldwork of Milman Parry and Albert Lord, who studied and recorded a living oral tradition during the 1930s and again in the 1950s in what was then Yugoslavia, that the Homeric epics were composed in performance during a long oral tradition that preceded any written version (Parry 1971, Lord 1960, Lord 1991, and Lord 1995; see also Nagy 1996a and Nagy 2002). In this tradition, the singer did not memorize a static text prior to performance, but would compose the song as he sang it. One of the most important revelations of the fieldwork of Parry and Lord is that every time the song is performed in an oral composition-in-performance tradition, it is composed anew. The singers themselves do not strive to innovate, but they nevertheless compose a new song each time (Dué 2002, 83-89). The mood of the audience or occasion of performance are just two factors that can influence the length of a song or a singer's choices between competing, but still traditional, elements of plot. The term "variant," as employed by textual critics when evaluating witnesses to a text, is not appropriate for such a compositional process. Lord explained the difference this way: "the word *multiform* is more accurate than 'variant,' because it does not give preference or precedence to any one word or set of words to express an idea; instead it acknowledges that the idea may exist in several forms" (Lord 1995, 23). Our textual criticism of Homeric epic, then, needs to distinguish what may genuinely be copying mistakes and what are performance multiforms: that is, what variations we see are very likely to be part of the system and the tradition in which these epics were composed (Dué 2001).

Once we begin to think about the variations as parts of the system rather than as mistakes or corruptions, textual criticism of the Homeric texts can then address fresh questions. Some of the variations we see in the written record, for example, reveal the flexibility of this system. Where different written versions record different words, but each phrase or line is metrically and contextually sound, we must not necessarily consider one "correct" or "Homer" and the other a "mistake" or an "interpolation." Rather, each could represent a different performance possibility, a choice that the singer could make, and would be making rapidly without reference to a set text (in any sense of that word).

Yet it is difficult to indicate the parity of these multiforms in a standard critical edition on the printed page. One version must be chosen for the text on the upper portion of the page, and the other recorded variations must be placed in an apparatus below, often in smaller text, a placement that necessarily gives the impression that these variations are incorrect or at least less important. Within a digital medium, however, the Homer Multitext will be able to show where such variations occur, indicate clearly which witnesses record them, and allow users to see them in an arrangement that more intuitively distinguishes them as performance multiforms. Thus a digital criticism—one that can more readily present parallel texts—enables a more comprehensive understanding of these epics. An approach to editing Homer that embraces the multiformity of both the performative and textual phases of the tradition—that is to say, a multitextual approach—can convey the complexity of the transmission of Homeric epic in a way that is simply impossible on the printed page.

We know that the circumstances of the composition of the Homeric epics demand a new kind of scholarly edition, a new kind of representation. We believe that a digital criticism of the witnesses to the poetry provides a means to construct one. Now the mission is to envision and create the Multitext so that it is true to these standards. The framework and tools that will connect these texts and make them dynamically useful are, as Peter Robinson has frankly stated (Robinson 2004 and Robinson 2005), the challenge that lies before us, as it does for other digital scholarly editions. How shall we highlight multiforms so as to make them easy to find and compare? We cannot simply have a list of variants, as some digital editions have done, for this would take them once again out of their context within the textual transmission and in many ways repeat the false impression that a printed apparatus gives. Breaking away from a print model is, as others have observed (Dahlström 2000, Robinson 2004), not as easy as it might seem, and digital editions have generally not succeeded yet in doing

so. How do we guide users through these many resources in a structured but not overly pre-determined or static way? Some working toward digital editions or other digital collections promote the idea of the reader or user as editor (Robinson 2004, Ulman 2006, the Vergil Project at http://vergil.classics.upenn.edu/project.html). Yet Dahlström (2000, section 4) warned that a "hypermedia database exhibiting all versions of a work, enabling the user to choose freely between them and to construct his or her 'own' version or edition, presupposes a most highly competent user, and puts a rather heavy burden on him or her." This type of digital edition "threatens to bury the user deep among the mass of potential virtuality." Especially because we do not want to limit the usefulness of this project to specialized Homeric scholars only, a balance between freedom of choice and structured guidance is important. We are not alone in grappling with the questions, and the need to recognize and represent the oral, traditional nature of Homeric poetry provides a special challenge in our pursuit of the answers.

References

Dahlström, M. "Drowning by Versions." *Human IT* 4 (2000). Available on-line at http://hb.se/bhs/ith/4-00/md.htm

Dué, C. "Achilles' Golden Amphora in Aeschines' Against Timarchus and the Afterlife of Oral Tradition." *Classical Philology 96* (2001): 33-47.

——. *Homeric Variations on a Lament by Briseis*. Lanham, Md.: Rowman and Littlefield Press, 2002: http://chs.harvard.edu/publications.sec/online_print_books.ssp/casey_du_homeric_variations/briseis_toc.tei.xml_1

Dué, C. and Ebbott, M. "'As Many Homers As You Please': an On-line Multitext of Homer, Classics@ 2 (2004), C. Blackwell, R. Scaife, edd., http://chs.harvard.edu/classicsat/issue_2/due-ebbott_2004_all.html

Foley, J. *The Singer of Tales in Performance*. Bloomington, 1995.

——. *Homer's Traditional Art.* University Park, 1999.

Greg, W.W. *The Shakespeare First Folio: Its Bibliographical and Textual History*. London, 1955.

Kiernan, K. "Digital Fascimilies in Editing: Some Guidelines for Editors of Image-based Scholarly Editions." *Electronic Textual Editing*, ed. Burnard, O'Keeffe, and Unsworth. New York, 2005. Preprint at: http://www.tei-c.org/Activities/ETE/Preview/kiernan.xml.

Lord, A. B. *The Singer of Tales*. Cambridge, Mass., 1960. 2nd rev. edition, 2000.

——. *Epic Singers and Oral Tradition*. Ithaca, N.Y., 1991.

——. *The Singer Resumes the Tale*. Ithaca, N.Y., 1995.

Monroy, C., Kochumman, R., Furuta, R., Urbina, E., Melgoza, E., and Goenka, A. "Visualization of Variants in Textual Collations to Anaylze the Evolution of Literary Works in the The Cervantes Project." *Proceedings of the 6th European Conference on Research and Advanced Technology for Digital Libraries, 2002*, pp. 638-653. http://www.csdl.tamu.edu/cervantes/pubs/ecdl2002.pdf

Nagy, G. *Poetry as Performance*. Cambridge, 1996.

——. *Homeric Questions*. Austin, TX, 1996.

——. *Plato's Rhapsody and Homer's Music: The Poetics of the Panathenaic Festival in Classical Athens*. Cambridge, Mass., 2002.

Parry, A. ed., *The Making of Homeric Verse*. Oxford, 1971.

Porter, D. "Examples of Images in Text Editing." *Proceedings of the 19th Joint International Conference of the Association for Computers and the Humanities, and the Association for Literary and Linguistic Computing, at the University of Illinois, Urbana-Champaign, 2007*. http://www.digitalhumanities.org/dh2007/abstracts/xhtml.xq?id=250

Robinson, P. "Where We Are with Electronic Scholarly Editions, and Where We Want to Be." *Jahrbuch für Computerphilologie Online* 1.1 (2005) at http://computerphilologie.uni-muenchen.de/jg03/robinson.html January 2004. In print in *Jahrbuch für Computerphilologie* 2004, 123-143.

——. "Current Issues in Making Digital Editions of Medieval texts—or, Do Electronic Scholarly Editions have a Future?" *Digital Medievalist* 1.1 (2005). http://www.digitalmedievalist.org/article.cfm?RecID=6.

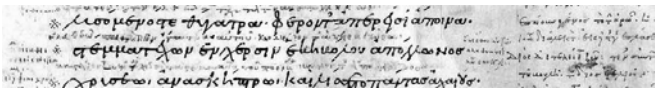
Stringer, G. "An Introduction to the Donne Variorum and the John Donne Society." *Anglistik* 10.1 (March 1999): 85–95. Available on-line at http://donnevariorum.tamu.edu/anglist/anglist.pdf.

Ulman, H. L. "Will the Real Edition Please Stand Out?: Negotiating Multi-Linear Narratives encoded in Electronic Textual Editions." *Proceedings of the <Code> Conference, 2006*. http://www.units.muohio.edu/codeconference/papers/papers/Ulman-Code03.pdf

Paper 2: Imaging the Venetus A Manuscript for the Homer Multitext

Ryan Baumann, W. Brent Seales and A. Ross Scaife

In May, 2007, a collaborative group of librarians, scholars and technologists under the aegis of the Homer Multitext project (HMT) traveled to Venice, Italy to undertake the photography and 3D scanning of the Venetus A manuscript at the Biblioteca Nazionale Marciana. Manoscritto marciano Gr.Z.454(=822): Homer I. Ilias, or Venetus A, is a 10th century Byzantine manuscript, perhaps the most important surviving copy of the *Iliad*, and the one on which modern editions are primarily based. Its inclusion will have the utmost significance for the HMT edition. In addition to the Iliad text, Venetus A contains numerous layers of commentary, called *scholia*, which serve to describe particular aspects of the main text. The scholia are written in very small script (mostly illegible in the only available print facsimile (Comparetti 1901)), and their organization on the page appears as a patchwork of tiny interlinking texts (see figure 1). The scholia are what truly set this manuscript apart, and their inclusion is of the greatest importance for the HMT. For this reason, we included 3D scanning in our digitization plan, with the aim to provide digital flattening and restoration that may help render the text (especially the *scholia*) more legible than digital images on their own. In this presentation we will describe how we obtained both digital photographs and 3D information, and will show how we combined them to create new views of this manuscript.



[Figure 1: This slice from Venetus A shows the main text, along with several layers of scholia: interlinear, marginal (both left and right), and intermarginal.]

This particular manuscript presents an almost perfect case for the application of digital flattening in the context of an extremely important, highly visible work that is the basis for substantial and ongoing scholarship in the classics. In the thousand-plus years since its creation, the environment has taken its toll on the substrate, the parchment upon which the texts are written. Several different types of damage, including staining, water damage, and gelitination (degradation of the fiber structure of the parchment) affect different areas of the manuscript. Much more widespread, affecting each folio in the manuscript and interfering with the reading of the text, is the planarity distortion, or *cockling*, that permeates the manuscript. Caused by moisture, cockling causes the folios to wrinkle and buckle, rather than to lie flat (see figure 2). This damage has the potential to present difficulties for the scholars by obstructing the text, especially the *scholia*. Virtual flattening provides us with a method for flattening the pages, making the text easier to read, without physically forcing the parchment and potentially damaging the manuscript.

[Figure 2: Cockling is visible along the fore edge of the Venetus A]

Digital Flattening

The procedure for digital flattening involves obtaining a 3D mesh of the object (a model of the shape of the folio), in addition to the 2D photograph. The photographic texture information can then be mapped on to the 3D mesh, and the 3D mesh can be deformed (or unwarped) to its original flat 2D shape. As the associated texture is deformed along with it, the original 2D texture can be recovered. For the digital flattening of the Venetus A, the project used the physical-based modeling approaches established in (Brown 2001, Brown 2004). The largest previous published application of this digital flattening technique was to BL Cotton Otho B. x, a significantly more damaged 11th century manuscript consisting of 67 folios (Kiernan 2002, Brown 2004). By contrast, Venetus A consists of 327 folios, a nearly five-fold increase. Thus, a system was needed which could apply the necessary algorithms for flattening without manual intervention, in an automated fashion. We also required a reliable system for imaging the document, acquiring 3D scans, and obtaining calibration data.

Imaging: 2D and 3D

Digital photographs were taken using a Hasselblad H1 with a Phase One P45 digital back, capable of a resolution of 5428x7230 (39.2 megapixels). The manuscript was held by a custom cradle built by Manfred Mayer from the University of Graz in Austria. The cradle enabled the photography team to hold the book open at a fixed angle and minimize stress on the binding, and consisted of a motorized cradle and camera gantry perpendicular to the page being imaged. The camera was operated via FireWire, to reduce the need for manual adjustments.

To acquire data for building the 3D mesh, we used a FARO 3D Laser ScanArm. Since we were unlikely to be able to use a flat surface near the cradle to mount the arm, we set the arm on a tripod with a custom mount. The ScanArm consists of a series of articulating joints, a point probe, and a laser line probe device attachment. The system operates by calibrating the point probe to the arm, such that for any given orientation of joints, the arm can reliably report the probe's position in 3D space. Another calibration procedure calibrates the point

probe to the laser line probe. The result is a non-contact 3D data acquisition system which can obtain an accurate, high-resolution untextured point cloud.

Imaging Procedure

As the document required a minimum of 654 data sets to be fully imaged (verso and recto of each folio), data acquisition was split into a number of sessions. Each session typically represented a 4 hour block of work imaging a sequence of either recto or verso, with two sessions per day. Fifteen sessions were required to image the entire manuscript. Due to the size and fragility of the codex, it was moved as seldom as possible. Thus, each session was bookended by a set of calibration procedures before and after placing the manuscript in the cradle. In order for the extrinsic calibration to be accurate, neither the FARO nor the cameras could be moved during a session.

The process for a typical page was thus:

- Position the page and color calibration strips, then turn on the vacuum to suction the page down
- Open the barn door lights on both sides of the cradle
- Photograph the page, via USB control
- Close lights to minimize damage to the manuscript
- Scan the page with the laser line probe on the ScanArm (due to the small width of the line, this last step usually took 2-5 minutes)

Periodic spot-checks of the calibration data were carried out to ensure that good data was being obtained. Test flattenings were also performed to check that everything worked as expected.

Relationship between 2D and 3D

As the 2D and 3D data acquisition systems are completely separate (unlike the tightly coupled systems of (Brown 2001, Brown 2004, Brown 2007)), a calibration procedure was required to determine the correspondence between 3D points in the arm and 2D points in the camera. By calibrating the camera and ScanArm in relation to one another, we were able to determine the location of any point in the three dimensional space within view of the camera and to flatten the folios within that space. As we were unsure of the amount of control we would have over the primary photography, we used our own camera (a Canon EOS 20D, which was attached to the camera gantry immediately beside the Hasselblad) to obtain a set of control images for each page and calibration data for each session. The Canon was set to manual focus and a fixed focus and zoom, in order to have intrinsic camera calibration which would remain relatively stable. As the goal of the Hasselblad photography was to obtain the highest quality photographs of each page, it did not retain the same focal length for every page and would instead be focus on the current page depth. As a result, intrinsics (including focal length, distortion matrix, and center of radial distortion) were acquired using the average focus that the camera would be at during a session. Once all this information was obtained – photographs, 3D information, and calibration – we were able to map the photographs to the 3D mesh and unwrap it to its 2D shape, helping to make legible texts that may otherwise be difficult to read.

The publication of legible, high-resolution digital photographs of the Venetus A manuscript is the first step towards an electronic edition of the text, *scholia*, and translation which will constitute a resource with far-reaching implications. Students and scholars of Greek literature, mythology, and palaeography could study in minute detail the most important manuscript of the *Iliad*. This codex has been seen by only a handful of people since its discovery and publication by Villoison in 1788 (Villoison 1788). Although Comparetti published a facsimile edition in 1901 (Comparetti 1901), very few copies of even this edition are available, even in research libraries, and the smallest writing in Comparetti's facsimile is illegible. The Homer MultiText, with its high-resolution digital images linked to full text and translation, will make this famous manuscript freely accessible to interested scholars (and others) worldwide.

Bibliography:

Baumann, R. *Imaging and Flattening of the Venetus* A. Master's Thesis, University of Kentucky, 2007.

Brown, M. S. and Seales, W. B. "Document Restoration Using 3D Shape: A General Deskewing Algorithm for Arbitrarily Warped Documents." ICCV 02 (2001): 367-375.

Brown, M. S. and Seales, W. B. "Image Restoration of Arbitrarily Warped Documents." *IEEE Transactions on Pattern Analysis and Machine Intelligence 26*: 10 (2004): 1295-1306.

Brown, M. S., Mingxuan, S., Yang R., Yun L., and Seales, W. B. "Restoring 2D Content from Distorted Documents." *IEEE Transactions on Pattern Analysis and Machine Intelligence 29*: 11 (2007): 1904-1916.

Comparetti, D. *Homeri Ilias Cum Scholiis. Codex Venetus A, Marcianus 454. Phototypice editus, Praefatus est Dominicus Comparettix.* Lugduni Batavorum: A. W. Sijthoff, 1901.

Kiernan, K., Seales, W. B., and Griffioen, J. "The Reappearances of St. Basil the Great in British Library MS Cotton Otho B. x." *Computers and the Humanities 36*: 1 (2002): 7-26.

Lin, Yun. *Physically-based Digital Restoration Using Volumetric Scanning.* PhD Dissertation, University of Kentucky, 2007.

De Villoison, J. B. G., ed. *Homeri Ilias ad veteris codicis Veneti fidem recensita*. Venice, 1788.

Paper 3: The Homer Multitext: Infrastructure and Applications

Christopher Blackwell and Neel Smith

The Homer Multitext (HMT) seeks to create a library of materials documenting the history of the Homeric tradition, all of which will be freely available online. To support this aim, Neel Smith and Christopher Blackwell, in collaboration with other scholars, librarians, and technologists, have spent five years defining a set of protocols for a distributed digital library of "scholarly primitives". The protocols aim to be as technologically agnostic as possible and to afford automatic discovery of services, the materials they serve, and the internal structures and citation schemes.

In winter 2007, the initial contents of the Homer Multitext were published online using the first generation of web-based applications built atop implementations of these protocols. In this presentation we will describe and demonstrate these protocols and discuss how they contribute to the vision of the Homer Multitext.

The TICI Stack

The technical editors of the Center for Hellenic Studies call the set of technologies that implement these protocols the "TICI Stack", with TICI being an acronym for the four categories of materials offered by the Homer Multitext Library: Texts, Images, Collections, and Indices.

Texts

The Texts components of the HMT are electronic editions and translations of ancient works: transcriptions of specific manuscripts of the Iliad and Odyssey, editions of Homeric fragments on papyrus, editions of ancient works that include Homeric quotations, and texts of ancient commentaries on Homeric poetry. These texts are marked up in TEI-Conformant XML, using very minimal markup. Editorial information is included in the markup where appropriate, following the EpiDoc standard. More complex issues that might be handled by internal markup (complex scholarly apparatus, for example, or cross-referencing) is not included; these matters are handled by means of stand-off markup, in indices or collections.

Access to texts is via implementations of the Canonical Text Services Protocol (CTS), which is based on the Functional Requirements for Bibliographic Records (FRBR). CTS extends FRBR, however, by providing hierarchical access from the broadest bibliographic level ("textgroup", or "author"), down through "work", "edition/translation", and into the contents of the text itself. The CTS protocol can point to any citeable unit of a text, or range of such units, and even more precisely to a single character. For example, the CTS URN:

```
urn:cts:tlg0012.tlg001:chsA:10.1
```

points to Homer (tlg0012), the Iliad (tlg001), the edition catalogued as "chsA" (in this case a transcription of the text that appears on the manuscript Marcianus Graecus Z. 454 [= 822]), Book 10, Line 1. A CTS request, handed this URN, would return:

Ἄλλοι μὲν παρὰ νηυσίν ἀριστῆες Παναχαιῶν

A more specific URN would be:

```
urn:cts:tlg0012.tlg001:chsA:10.1:ρ[1]
```

would point to the second Greek letter "rho" in this edition of Iliad 10.1, which appears in the word ἀριστῆες.

Images

The HMT currently offers over 2000 high resolution images of the folios of three Homeric manuscripts, captured in the spring of 2007 at the Biblioteca Nationale Marciana in Venice. These images are accessible online through a basic directory-listing, but also through a web-based application, described below. To relate these images with bibliographic data, technical metadata, and associated passages of texts, the HMT relies on Collections and Indices.

Collections

A collection is a group of like objects, in no particular order. For the CTS, one example is a Collection of images, with each element in the collection represented by an XML fragment that records an id, pointing to a digital image file, and metadata for that image. Another example is a Collection of manuscript folios. Each member of this Collection includes the name of the manuscript, the folio's enumeration and side ("12-recto"), and one or more CTS URNs that identify what text appears on the folio. Collections are exposed via the Collection Service, which allows discovery of the fields of a particular collection, querying on those fields, and retrieval of XML fragments. A final example would be a collection of lexicon of Homeric Greek.

Indices

Indices are simple relations between two elements, a reference and a value. They do much of the work of binding together TICI applications.

The Manuscript Browser

The first data published by the HMT Library were the images of manuscripts from Venice. These are exposed via the Manuscript Browser application, the first published application built on the TICI Stack.

http://chs75.harvard.edu/manuscripts

This application allows users to browse images based either on the manuscript and folio-number, or (more usefully) by asking for a particular manuscript, and then for a particular book and line of the Iliad.

Users are then taken to a page for that folio. Online interaction with the high-resolution images is through the Google Maps API, and AJAX implementation that allows very responsive panning and zooming, without requiring users to download large files.

From the default view, users can select any other views of that page (details, ultraviolet photographs, etc.), and can choose to download one of four versions of the image they are viewing: an uncompressed TIFF, or one of three sizes of JPEG.

This application draws on a CTS service, two indices, and a collection of images. It queries a CTS service to retrieve bibliographic information about manuscripts whose contents are online--the CTS protocol is not limited to electronic texts, but can deliver information about printed editions as well. When a user requests a particular book and line of the Iliad, it queries one index whose data looks like this:

```
<record>
<ref>urn:cts:greekLit:
tlg0012.tlg001:1.1</ref>
<value>msA-12r</value>
</record>
```

Here the <ref> element contains a URN that, in this case, points to Book 1, Line 1 of the Iliad, and the <value> element identifies folio 12, recto, of "manuscript A". This allows the application to query a collection of manuscript folios and to determine which images are available for this folio.

The display of an image via the Google Maps API requires it to be rendered into many tiles--the number depends on the size and resolution of the image, and the degree of zooming desired. For the images from Venice, we produced over 3,000,000 image tiles. This work was made possible by the Google Maps Image Cutter software from the UCL Centre for Advanced Spatial Analysis, who kindly added batch-processing capability to their application at our request.

Next Steps

Currently under development, as of November 2007, is the TICI Reader, a web-based application to bring texts together with indexed information and collections. This is intended to coincide with the completion of six editions of the Iliad, electronic transcriptions of important medieval manuscripts of the Iliad, and a new electronic edition of the medieval scholarly notes, the scholia, that appear on these manuscripts.

Availability

All of the work of the HMT is intended to be open and accessible. The images and texts are licensed under a Creative Commons License, all tools are based on open standards and implemented in open source technologies. And work is progressing, and more planned, on translations of as many of these materials as possible, to ensure that they reach the widest possible audience.

Bibliography

Canonical Texts Services Protocol: http://chs75.harvard.edu/projects/diginc/techpub/cts

Creative Commons: http://creativecommons.org/

EpiDoc: Epigraphic Documents in TEI XML: http://epidoc.sourceforge.net/

Google Maps API: http://www.google.com/apis/maps/

IFLA Study Group on the Functional Requirements of Bibliographic Records. "Functional Requirements of Bibliographic Records: final report." München: K. G. Saur, 1998: http://www.ifla.org/VII/s13/frbr/frbr.pdf

TEI P4: Guidelines for Electronic Text Encoding and Interchange. Edited by C. M. Sperberg-McQueen and Lou Burnard. The TEI Consortium: 2001, 2002, 2004.

Defining an International Humanities Portal

Neil Fraistat
nfraistat@gmail.com
Maryland Institute for Technology in the Humanities, USA

Domenico Fiormonte
fiormont@uniroma3.it
Università Roma Tre, Italy

Ian Johnson
johnson@acl.arts.usyd.edu.au
University of Sydney, Australia

Jan Christoph Meister
jan-c-meister@uni-hamburg.de
Universität Hamburg, Germany

John Unsworth
unsworth@uiuc.edu
University of Illinois, Urbana-Champaign, USA

What would a portal for humanists look like? Who would it serve and what services would it provide?

centerNet, an international network of digital humanities centers, proposes a panel to discuss an ongoing conversation about the need for an International Humanities Portal (IHP) to serve students and researchers. The panelists will present different perspectives a number of questions we have posed to the larger network. The panel will both present their own positions on the case for a Portal and will report back on the collaborative development of a common Needs Analysis that identifies the need and the potential scope of such a portal. We propose a panel as a way to reflect back to the community the discussions so far and to solicit further input on the need and case for a common infrastructure. The participants bring an international perspective to the discussion that is central to what we imagine.

Some of the questions the panelists will address are:

1. Who would be the users of a humanities portal? Would a portal serve only researchers or would it serve students and anyone else interested in the humanities? Would a portal serve primarily digital humanists or would it serve all humanists?

1. The need for a portal starts with the definition of the audience and how a portal might serve their interests. As the title suggests and the participants in the panel represent, we have also started from the position that research crosses national and linguistic boundaries and that we should therefore imagine an international portal that can be multilingual and support the needs of international research.

2. Who might support such a portal? How might the development of such a portal be funded and how would it be maintained? centerNet includes centers with access to development and support resources, but no center has the capacity to support a successful international portal. We therefore imagine a distributed model that will draw on expertise and support from the centerNet community and beyond. We also imagine that the development of a consensus about the need and scope of a portal by centerNet could help individual centers to secure support from national and other funding bodies to develop and maintain components. Ultimately we hope the support needed at any one center will be easier to secure and easier to maintain if backed up by an articulated Need Analysis from a larger body like centerNet.

3 What a humanities portal might look like? Will it actually be a portal or should we imagine providing services to existing university portals? Many universities are developing student and faculty portals as are scholarly associations and projects. Portals now seem dated in light of Web 2.0 social technologies. For this reason we imagine that an IHP will need to interoperate with other portals through "portlets" or OpenSocial (code.google.com/apis/opensocial/) plug-ins that can be added to project sites. The IHP will itself need to play well with other resources rather than aim to organize them.

4. What services might it provide? What services are already provided by centers and projects? Can the portal provide an entry point into the richness of existing resources? The panelists will survey the variety of resources already available to their communities and discuss what new resources are needed. Above all a portal should be a door into an area of inquiry -- panelists will reflect on what they and their community would expect to have easy access to from something that promised to be an International Humanities Portal.

5. Just as important as defining the services is imagining how services might interoperate. A humanist familiar with the web will know of resources from Voice of the Shuttle to Intute: Arts and Humanities, but can we imagine how these services might be brought together so that one can search across them? Can we imagine a mashup of services providing new and unanticipated possibilities?

6. What are the next steps? How can the case for an IHP be strengthened? How can centerNet help, not hinder, projects that want to develop and support components that meet the needs of the community? One of the purposes of this panel is to solicit feedback from the larger digital humanities community.

Portals are typically characterized by two features. First, their users can customize their account to show or hide particular resources.. Thus users of the TAPoR portal, for example, can define texts and tools that they want to use and ignore the rest.

The ease of customization and the depth of customization are important to users if they use portals regularly. Second, users expect that a portal provides access to the breadth of services and resources they need for a domain of inquiry. Thus students expect a student portal to provide access to all the online resources they need as a student, from the library account to their course information. A portal should be just that -- a door into a domain. Is it possible for an IHP to be on the one hand easy to use (and customizable), and on the other hand truly provide broad access to resources? Is it possible that an IHP is too ambitious and would either be to complicated for humanists to use or not broad enough in scope to be useful? The answer in part lies in imagining an initial audience and conducting the usability studies needed to understand what they would expect. Thus we imagine an iterative process of designing for a first set of users and redesigning as we learn more. This means a long term development commitment which is beyond most project funding. The challenges are enormous and we may be overtaken by commercial portals like Google or Yahoo that provide most of what humanists need. We believe, however, that the process of defining the opportunities and challenges is a way to recognize what is useful and move in a direction of collaboration. We hope that in the discussion modest first steps will emerge.

Bibliography

centerNet: http://www.digitalhumanities.org/centernet/

Detlor, Brian. *Towards knowledge portals: From human issues to intelligent agents.* Dordrecht, The Netherlands: Kluwer Academic Publishers, 2004.

Intute: Arts and Humanities: http://www.intute.ac.uk/artsandhumanities/

TAPoR: Text Analysis Portal for Research: http://portal.tapor.ca

VoS: Voice of the Shuttle: http://vos.ucsb.edu/

Beyond Search: Literary Studies and the Digital Library

Matthew Jockers
mjockers@stanford.edu
Stanford University, USA

Glen Worthey
gworthey@stanford.edu
Stanford University, USA

Joe Shapiro
jpshapir@stanford.edu
Stanford University, USA

Sarah Allison
sdalliso@stanford.edu
Stanford University, USA,

The papers in this session are derived from the collaborative research being conducted in the "Beyond Search and Access: Literary Studies and the Digital Library" workshop at Stanford University. The workshop was set up to investigate the possibilities offered by large literary corpora and how recent advances in information technology, especially machine learning and data-mining, can offer new entry points for the study of literature. The participants are primarily interested in the literary implications of this work and what computational approaches can or do reveal about the creative enterprise. The papers presented here share a common interest in making literary-historical arguments based on statistical evidence harvested from large text collections. While the papers are not, strictly speaking, about testing computational methods in a literary context, methodological evaluations are made constantly in the course of the work and there is much to interest both the literary minded and the technically inclined.

Abstract 1: "The Good, the Bad and the Ugly: Corralling an Okay Text Corpus from a Whole Heap o' Sources."

Glen Worthey

An increasingly important variety of digital humanities activity has recently arisen among Stanford's computing humanists, an approach to textual analysis that Franco Moretti has called "distant reading" and Matthew Jockers has begun to formally develop under the name of macro-analysis. At its most simple, the approach involves making literary-historical arguments based on statistical evidence from "comprehensive" text corpora. One of the peculiarities of this activity is evident from the very beginning of any foray into it: it nearly always requires the creation of "comprehensive" literary corpora. This paper

discusses the roles and challenges of the digital librarian in supporting this type of research with specific emphasis on the curation (creation, licensing, gathering, and manipulation) of a statistically significant corpus.

Because the arguments at stake are largely statistical, both the size of the "population" of texts under examination, and any inherent bias in the composition of the corpus, are potentially significant factors. Although the number of available electronic texts from which to draw a research corpus is increasing rapidly, thanks both to mass digitization efforts and to expanding commercial offerings of digital texts, the variability of these texts is likewise increasing: no longer can we count on working only with well-behaved (or even well-formed) TEI texts of canonical literary works; no longer can we count on the relatively error-free re-keyed, proofread, marked-up editions for which the humanities computing community has spent so much effort creating standards and textual corpora that abide by these standards.

In addition to full texts with XML (or SGML) markup that come to us from the model editions and exemplary projects that we have come to love (such as the Brown Women Writers Project), and the high-quality licensed resources of similar characteristics (such as Chadwyck Healey collections), we are now additionally both blessed and cursed with the fruits of mass digitization (both by our own hands and by commercial and consortial efforts such as the Google Library Project and Open Content Alliance), which come, for the most part, as uncorrected (or very lightly corrected) OCR. Schreibman, et al., 2008, have described University of Maryland efforts to combine disparate sources into a single archive. And a similar effort to incorporate "messy data" (from Usenet texts) into usable corpora has been described in Hoffman, 2007; there is also, of course, a rich literature in general corpus design. This paper will discuss the application of lessons from these other corpus-building projects to our own inclusion of an even more variant set of sources into a corpus suitable for this particular flavor of literary study.

This digital librarian's particular sympathies, like perhaps those of many or most who participate in ADHO meetings, lie clearly with what we might now call "good," "old-fashioned," TEI-based, full-text digital collections. We love the perfectionism of "true" full text, the intensity and thoughtfulness and possibility provided by rich markup, the care and reverence for the text that have always been hallmarks of scholarly editing, and that were embraced by the TEI community even for less formal digital editions.

Now, though, in the days of "more" and "faster" digital text production, we are forced to admit that most of these less-than-perfect full-text collections and products offer significant advantages of scale and scope. While one certainly hesitates to claim that quantity actually trumps quality, it may be true that quantity can be transformed into quality of a sort different from what we're used to striving for. At the same time, the "macro-analysis" and "distant reading" approaches demand corpora on scales that we in the "intensely curated" digital library had not really anticipated before. As in economics, it is perhaps difficult to say whether increased supply has influenced demand or vice versa, but regardless, we are certainly observing something of a confluence between mass approaches to the study of literary history and mass digitization of its textual evidence.

This paper examines, in a few case studies, the gathering and manipulation of existing digital library sources, and the creation of new digital texts to fill in gaps, with the aim of creating large corpora for several "distant reading" projects at Stanford. I will discuss the assessment of project requirements; the identification of potential sources; licensing issues; sharing of resources with other institutions; and the more technical issues around determining and obtaining "good-enough" text accuracy; "rich-enough" markup creation, transformation and normalization; and provision of access to the corpora in ways that don't threaten more everyday access to our digital library resources. Finally, I'll discuss some of the issues involved in presenting, archiving, and preserving the resources we've created for "special" projects so that they both fit appropriately into our "general" digital library, and will be useful for future research.

References

Sebastian Hoffmann: Processing Internet-derived Text—Creating a Corpus of Usenet Messages. Literary and Linguistic Computing Advance Access published on June 1, 2007, DOI 10.1093/llc/fqm002. *Literary and Linguistic Computing* 22: 151-165.

Susan Schreibman, Jennifer O'Brien Roper, and Gretchen Gueguen: Cross-collection Searching: A Pandora's Box or the Holy Grail? Literary and Linguistic Computing Advance Access published on April 1, 2008, DOI 10.1093/llc/fqm039. *Literary and Linguistic Computing* 23: 13-25.

Abstract 2: "Narrate or Describe: Macro-analysis of the 19th-century American Novel"

Joe Shapiro

Does the 19th-century American novel narrate, or does it describe? In 1936, Georg Lukács argued that over the course of the nineteenth century description replaced narration as the dominant mode of the European (and especially French) novel. What happens, we want to know, for the American novel? We begin with a structuralist distinction between the logics of narration and description: narration as the articulation of the events that make up a story; description as the attribution of features to characters, places, and objects in the story. We then set out, with the help of computational linguistics, to identify the computable sentence-level stylistic "signals" of narration and description from a small training sample (roughly

2,000 sentences), and thus to develop a model that can classify narration and description in "the wild"—a corpus of over 800 American novels from 1789 to 1875. This paper will discuss how we have refined our research problem and developed our classifying model—trial-and-error processes both; our initial results in "the wild"; and finally how macro-analysis of this kind leads to new problems for literary history.

Existing scholarship suggests that "realist" description enters the American novel with the historical novel; thus our initial training set of samples was taken from 10 American historical novels from the 1820s, 1830s, and 1840s (by J.F. Cooper and his rivals). Participants in the Beyond Search workshop have tagged random selections from these 10 novels. The unit of selection is, for convenience, the chapter. Selected chapters are broken (tokenized) into individual sentences and human-tagged using a custom XML schema that allows for a "type" attribute for each sentence element. Possible values for the type attribute include "Description," "Narration," "Both," "Speech," and "Other." Any disagreement about tagging the training set has been resolved via consensus. (Since the signals for description may change over time—indeed, no small problem for this study—we plan to add an additional training sample from later in the corpus.)

Using a maximum-entropy classifier we have begun to investigate the qualities of the evolving training set and to identify the stylistic "signals" that are unique to, or most prevalent in, narrative and descriptive sentences. In the case of description, for example, we find a marked presence of spatial prepositions, an above average percentage of nouns and adjectives, a relatively low percentage of first and second person pronouns, above average sentence lengths, and a high percentage of diverse words (greater lexical richness). From this initial work it has become clear, however, that our final model will need to include not simply word usage data, but also grammatical and lexical information, as well as contextualizing information (i.e., the kinds of sentence that precede and follow a given sentence, the sentence's location in a paragraph). We are in the process of developing a model that makes use of part of speech sequences and syntactic tree structures, as well as contextualizing information.

After a suitable training set has been completed and an accurate classifying model has been constructed, our intention is to "auto-tag" the entire corpus at the level of sentence. Once the entire corpus has been tagged, a straightforward quantitative analysis of the relative frequency of sentence types within the corpus will follow. Here, the emphasis will be placed on a time-based evaluation of description as a feature of 19th-century American fiction. But then, if a pattern emerges, we will have to explain it—and strictly quantitative analysis will need to be supplemented by qualitative analysis, as we interrogate not just what mode is prevalent when, but what the modes might mean at any given time and how the modes themselves undergo mutation.

Abstract 3: "Tracking the 'Voice of Doxa' in the Victorian Novel."

Sarah Allison

The nineteenth-century British novel is known for its moralizing: is it possible to define this "voice of doxa," or conventional wisdom, in terms of computable, sentence-level stylistic features? A familiar version of the voice of doxa is the "interrupting narrator," who addresses the reader in the second person in order to clarify the meaning of the story. This project seeks to go beyond simple narrative interruption to the explication of ethical signals emitted in the process of characterization (how the portrayal of a character unfolds over the course of a novel). It also takes a more precise look at the shift in tense noted by narratologists from the past tense of the story to the present-tense of the discourse, in which meaning can be elaborated in terms of proverbs or truisms. (An example from *Middlemarch*: "[Fred] had gone to his father and told him one vexatious affair, and he had left another untold: in such cases the complete revelation **always produces** the impression of a previous duplicity" 23). This project is the first attempt to generate a set of micro- stylistic features that indicate the presence of ethical judgment.

Through an analysis of data derived through ad hoc harvesting of frequently occurring lexical and syntactic patterns (e.g. word frequencies, part of speech saturation, frequent grammatical patterns, etc) and data derived through the application of supervised classification algorithms this research attempts to determine a set of grammatical features that tend to cluster around these moments of direct narrative discourse.

This research is an alternative application of the method developed in Joe Shapiro's Beyond Search workshop project (see abstract above), which seeks to identify computable stylistic differences between narrative and descriptive prose in 19th century American fiction. In this work we seek to create another category within the "descriptive," a subcategory that captures moments of explicitly moralized description: the Voice of Doxa. We identify the formal aspects of this authorial "voice" in order to "hunt" for similar moments, or occurrences, in a series of novels. The work begins with a limited search for patterns of characterization evident among characters in a single George Eliot novel; from this we develop a model, which we will then apply to the entire Eliot corpus. In the end, the target corpus is extended to include 250 19th century British novels wherein we roughly chart the evolutionary course of the "ethical signal."

Designing, Coding, and Playing with Fire: Innovations in Interdisciplinary Research Project Management

Stephen Ramsay
sramsay@unlserve.unl.edu
University of Nebraska at Lincoln, USA

Stéfan Sinclair
sgsinclair@gmail.com
McMaster University, Canada

John Unsworth
unsworth@uiuc.edu
University of Illinois, Urbana-Champaign, USA

Milena Radzikowska
mradzikowska@gmail.com
Mount Royal College, Canada

Stan Ruecker
sruecker@ualberta.ca
University of Alberta, Canada

From its inception, Digital Humanities has relied on interdisciplinary research. Nationally and internationally, our experience of doing research involving people from various disciplines has become increasingly important, not only because the results are often innovative, fascinating, and significant, but also because the management tools and practices we have been developing are potentially useful for a wide range of researchers. In this session, we will discuss three very different aspects of interdisciplinary project management, based on our combined experience of decades of collaboration.

The first paper, "Hackfests, Designfests, and Writingfests," is an attempt to pull together a summary of the best practices we have been developing in the area of planning and carrying out face-to-face work sessions. Over the past three years, we have carried out a total of fourteen Hackfests, involving programmers, designers, and writers. In studying our own results, we have generated practical guidelines on topics such as the choice of locations, the assignment of tasks, and the handling of logistics.

The second paper, "Hackey," provides a description of a new online tool that we have been developing for use by programmer/designer pairs, to encourage small but frequent activity on a project. It often happens that people make commitments, especially to virtual teams, that result in them working in large but infrequent blocks of time, as at the Hackfests, when it may be more productive in the long run if they are able to devote instead an hour a day. We've attempted to address this problem through the design of a game that uses the concept of volleying back and forth on a specific topic, all under the watchful eye of a benevolent third-party pit boss.

The final paper, entitled "Rules of the Order," considers the sociology of large, multi-institutional software development projects. Most people would be able to cite the major obstacles standing in the way of such collaborations, like the lack of face-to-face communication. But even with the best methods of communication available, researchers still need to acknowledge the larger "sociology" of academia and develop ways of working that respect existing institutional boundaries, workflows, and methodologies.

Hackfests, Designfests, and Writingfests: The Role of Intense Periods of Face-to-Face Collaboration in International Research Teams

Stan Ruecker, Milena Radzikowska, and Stéfan Sinclair

Extended Abstract

As the authors of this paper have become increasingly active in international, interdisciplinary, collaborative research projects, we have progressively adopted a growing range of online tools to help manage the activities. We've used Basecamp for task assignment, Meeting Wizard for scheduling conference calls, project listserves for discussion, and research wikis for archiving the results. We have SVN for software versioning, and Jira for tracking and prioritizing programming tasks. Our websites provide a public face, with project rationales, contact information, and repositories for research presentations and articles. We also try whenever possible to provide online prototypes, so that our readers have the opportunity to try our interface experiments for themselves, with appropriate warnings that things might be held together on the back end with duct tape, and participant agreements for those occasions when we hope to capture log records of what they are doing.

All of the online tools we use contribute to our productivity. However, we have also become increasingly convinced of the benefits of periodic face-to-face working sessions, where we gather from the twelve-winded sky to focus our attention on the project. In the words of Scott Ambler, "have team, will travel." The value of such face-to-face meetings is well understood by distance educators (e.g. Davis and Fill 2007), who see such focused time as a significant factor in student learning. We have been speculatively trying variations of our working sessions, such as choice of location, numbers of participants, types of work, and other logistical details like meals and entertainment. Over the past three years, we have held fourteen such events. They have involved a number of participants ranging from as few as two to as many as fifteen. Nine of the events were dedicated to writing and designing; three were used for programming and designing; one was devoted just to programming, and another to just designing. Our goal in this paper is to provide the insights we've gained from these variations for the benefit of other researchers who are interested in trying their own project-oriented work fests.

Types of Work

It is important that these not be meetings. No one expects to get work done at a meeting, not in the sense of writing a paper, designing a system, or programming, so the expectations are at odds with the goal of the exercise. Meetings typically carry a lot of connotative baggage, not least of which is inefficiency and frustration, as expressed by the educational videos with John Cleese entitled "Meetings, Bloody Meetings." Humanists are rarely trained in conducting meetings efficiently and productively, and it shows. By emphasizing to all the participants that the goal is to work productively together in the same room, we can hit the ground running and get a lot accomplished in a short period of time. We have in almost every case combined people who are working on different kinds of tasks. In the two instances where we had only a single task (programming and designing, respectively) we had a sense of accomplishing much less. It is difficult in some ways to verify this perception, but the energy levels of participants in both cases seemed lower. Our interpretation would be, not that there is competition per se between different disciplines, but that it is harder to judge how much another disciplinary group is accomplishing, and that uncertainty helps to keep people motivated.

Location, Location, Location

There are many temptations to locate the events in places that will be convenient. We have found that in this instance it is better to avoid temptation if possible. For example, if there are five participants and three are from the same city, why not hold the event in that city? It reduces costs, in particular for travel and accommodations, and the participants can also serve as hosts for their out-of-town colleagues. In practice, however, we have found that holding a hackfest in some of the participant's hometown, lowers the hackfest's productivity level. The biggest obstacle is the local demands on the attention of the resident participants. We have found that local participants are constantly on call for work and domestic commitments, which means they tend to absent themselves from the event. One local host for a comparatively large number of participants is less problematic, since the periodic absence of one participant is not as critical a factor in maintaining the momentum of the entire group, but for small groups it can be significant. An alternative is therefore to transport the participants to a pleasant working location, provided that the attractions of the location are not so powerful as to be distracting. We are all familiar with conferences in exotic locations where participants have difficulty focusing on the conference because the beach is calling. Given this mixture of considerations, our current guideline is therefore to use a location no closer than an hour's travel from some cluster of the participants, so that both the overall travel costs and local demands for time can be reduced.

Participants

In general, the participants on our research projects know what they are doing. In fact, many of them are eager to get going on tasks that have fallen behind due to the other demands on their time. So in general, it has not been necessary to provide a large administrative overhead in terms of controlling and monitoring their activities during the event. Given the nature of interdisciplinary collaboration, too much attempt to control is counter-productive in any case. However, the nature of task assignment during a hackfest is a matter that may require some finesse. We have made errors where we had someone working on an assignment that was not the kind of assignment they could tackle with enthusiasm. In these cases, a little preliminary consultation with the individual participant would have made a big improvement. This consultation could have occurred either before the hackfest or even upon arrival, but it was a misstep not to have it early. In terms of numbers of participants, it appears that you need enough people to generate momentum. Two or three is not enough, although four seems to be. Our largest group so far has been fifteen, which is somewhat large to co-ordinate, and they end up working anyway in smaller groups of 2-5, but a larger group also makes the event more dynamic and can provide greater opportunities for cross-disciplinary consultation.

Our hackfests are set up according to a flat organizational structure. All participants are treated as equals and decisions are, as much as possible, based on consensus. Even when planning a hackfest, we consult with all invited participants on potential locations, duration, and objectives. In rare cases where leadership is required during the event, it is clear to everyone that the overriding structure is based on equality. For mid-sized groups, we find that it is useful to have a "rover" or two, who constantly moves between the different groups and tries to identify potential problems and opportunities for bridging group efforts.

A further point related to participants is to, whenever possible, consider the past, present, and future of the project. It is too easy to focus on what needs to be done immediately and to only include those who are currently active in the project. However, we have found that including previous participants – even if they are not currently active on the project – helps to remind the team of some of the institutional history, providing some context for past decisions and the current situation. Similarly, a hackfest is an excellent opportunity to recruit prospective new colleagues in an intense and enjoyable training experience.

Meals and Accommodations

We want these events to be highlights for people on the projects, so it is important that the meals and accommodations be the best that we can reasonably afford. Eating, snacking, and coffee breaks provide time for discussions and developing social capital, both of which are significant for the success of a research project. We have a senior colleague in computing

science, not related to our group, who holds hackfests at his cottage, where the graduate students take turns doing the cooking. We would hesitate to recommend this approach, since there would be lost time for the cook that could be more productively used on the project. At our most recent event, we hired a caterer. This strategy also worked well because it removed the added stress of finding, three times a day, a good place to feed everyone. In terms of sleeping accommodations, we have tended to provide one room per participant, since the additional stress of living with a roommate seems an unnecessary potential burden to introduce; private rooms can also be a space to individually unwind away from what may be an unusually concentrated period of time with the same group of colleagues.

Work Rooms

The situated environment of work is very important. We have worked in the common areas of cluster housing, in the lounges of hotels, in boardrooms, in university labs, and on one occasion in a rustic lodge. We hesitate to say it, because there were problems with the internet connectivity and task assignment, but our four days in the rustic lodge were probably the most successful so far. We were in rural Alberta, but there was a nearby pond, a place for ultimate Frisbee games, and there were fire pits, hot tubs, and wood chopping that we could amuse ourselves with during breaks. There is a value value in having varied forms of entertainment and leisure, which can provide mental breaks. We prefer if possible to have a mix of physical activities and pure entertainment (music, movies), although it is important to be sensitive to different tastes.

Pair Programming

Even when there are a dozen participants, it is normally the case that we end up working in smaller groups of 2, 3, or 4. The dynamics of the group and the task can indicate the appropriate group size. We have also recently tried pair programming, where two programmers both work on the same computer. It seems to be a useful approach, and we intend to do more of it. However, the type of activity that can be paired-up seems to depend on the nature of the participants' discipline and typical work approach. For example, we have found that two designers work well together on concept planning and paper prototypes, but they need divided tasks, and two computers, when it comes to creating a digital sketch.

Wrapping-up

We think that if participants feel, at the end of the hackfest, that the project has moved forward and their time was well spent, they tend to continue progressing on the project. We have tried a few different methods to create that sense of conclusion and, therefore, satisfaction. At Kramer pond, for example, we concluded the hackfest by a show-andtell of tasks accomplished during the event, and an agreement about the next steps participants would take upon returning home. We have also tried to end a hackfest with a conference paper submission. This strategy has had mixed results – about half of the intended papers managed to reach the submission stage.

Documentation

Increasingly, we have focused a portion of our hackfest energy on documenting the event. We have used image posting sites, such as flickr.com, wiki posts, and video recording. Documentation serves as a reminder of agreed-upon and accomplished tasks, fosters a sense of satisfaction with the activity, and maintains a record of successes (and failures) emergent from the activity.

Future Directions

Although there are costs associated with hackfesting, we feel that the benefits so far have been tremendous. Projects typically take a quantum leap forward around the momentum that can be gained at a hackfest. We need to learn more about ramping up, and we are also interested in better documentation of the events themselves, perhaps in the form of videography.

References

Ambler, Scott. 2002. "Bridging the Distance." http://www.ddj.com/article/printableArticle.jhtml?articleID=184414899&dept_url=/architect/

Davis, Hugh, and Karen Fill. 2007. Embedding Blended Learning in a University's Teaching Culture: Experiences and Reflections. *British Journal of Educational Technology*, 38/5, pp. 817-828.

Hackey: A Rapid Online Prototyping Game for Designers and Programmers

Stéfan Sinclair and Stan Ruecker

Extended Abstract

Researchers in Humanities Computing are frequently involved in the production of new online software systems. The specific roles vary, with some researchers work primarily as humanists interested in building a tool in order to carry out their own next project, while others may be archivists creating a digital collection as a public resource, programmers developing prototype systems for testing a concept, or designers experimenting with new approaches to interfaces or visualizations. The semi-serious slogan of the Text Analysis Developer's Alliance (TADA) summarizes the activity in this way: "Real humanists make tools." In the case of design of the tools, it is useful to consider three levels, which in some contexts we have referred to as basic, advanced, and experimental. A basic tool is one that will run without difficulty in every contemporary browser. It provides functionality that everyone expects. A search function is a good example. In a basic tool, it is usual to follow the current best practices for online design, with due attention to the standard heuristics that can help

ensure that the result is usable. Nielsen (2000) provides a list of ten such criteria: (1) Use simple and natural dialogue. (2) Speak the users' language. (3) Minimize user memory load. (4) Consistency. (5) Feedback. (6) Clearly marked exists. (7) Shortcuts. (8) Good error messages. (9) Prevent errors. (10) Help and documentation.

An advanced tool is a basic tool with more features. For example, an advanced search function might provide support for Boolean operators, or allow nested queries, or include proximity searches, where a search is successful if the target words occur with in a specified number of words from each other. Advanced tools may have an experimental dimension, but they are still largely recognizable as something a typical user can be expected to have seen before. They are still within the realm of contemporary best practices.

In the case of experimental tools, however, the goal is not primarily to implement the current best practices, but instead to experiment with new visual forms for information display and online interaction. Often the prototype will serve as the basis for user testing involving a new opportunity for action. Examples might include Shneiderman et al's (1992) concept of direct manipulation, where the data and the interface are so tightly coupled that the visualization is the interface, Bederson's (2001) approach to zoomable user interfaces (ZUIs) where levels of magnification become central to interaction, or Harris and Kamvar's (2007) constellation approach to emotional expression in the blogosphere, where the results of blog scrapes are plotted as patterns that dynamically emerge from a constantly changing visual field. As Lev Manovich put it during a question and answer period in a paper session at Digital Humanities 2007, "in this line of research, a prototype is itself a theory."

The experimental design of interfaces and visualization tools is a special case for researchers in humanities computing, because it involves the extension of concepts into areas that are not well defined. In our research projects, we typically have a content expert, programmer, and visual communication designer working closely together in a cycle of static sketching, kinetic sketching, prototyping, and refinement, where each iteration may require as much as several months or as little as a few hours. One of our ongoing interests is in finding ways to facilitate the process.

In this paper, we describe our initial experiments in creating an online system that encourages the research team to focus on details taken one at a time, in order to support rapid turnaround in design and development. Hackey provides the designer and programmer with a means of focusing on individual issues in the process, and rapidly exchanging idea, sketches, and prototypes. The game also includes a third role, in the form of the "pit boss" who lurks on the conversation and can step in to provide guidance or suggestions. Either player can also choose during a turn to address a question to the pit boss, usually in the form of clarification or appeal for arbitration of some kind.

In addition to encouraging rapid turnaround, Hackey is also intended to encourage fun. By treating the exchanges as a kind of collaborative game, there is room for the players to contextualize their activity as more than a task list or a series of Jira tickets. Hackey provides a framework for discussions that can be at the same time serious and lighthearted. The principle is that the bounds of discourse are opened slightly by the game metaphor, to allow for mutual interrogation of the underlying assumptions held by both the designer and the programmer. This framework addresses an ongoing difficulty with interdisciplinary collaboration, where experts from two fields need to be able to find common ground that can accommodate the expertise of both, without undo compromise of the disciplinary expertise of either party. What we would like to avoid is the situation where the programmers feel that they are "working for" the designers, or the designers feel that they are not considered important partners in the process by the programmers, and that their proposals can be rejected out of hand without discussion. In this respect, the third party "pit boss" role is useful, since in extreme situations it can serve as a court of second appeal for either of the players, and under normal conditions it constitutes a presence that helps maintain the spirit of free and open exchange of ideas.

Future Directions

The Hackey prototype is still in its early stages, and we are still in the phase of experimentation, but we hope to make it more widely available for interdisciplinary teams who are seeking new methods for collaborative work. By logging their dialogues in the system and treating these records as an objective of interpretive study, we hope to be able to elaborate the original concept with additional features appropriate to the general goals.

References

Bederson, B. 2001. PhotoMesa: A zoomable image browser using quantum treemaps and bubblemaps. *Proceedings of the 14th annual ACM symposium on user interface software and technology.* 71–80. http://doi.acm.org/10.1145/502348.502359.

Harris, J. and S. Kamvar. 2007. We feel fine. http://wefeelfine.org.

MONK. 2007. Metadata Offer New Knowledge. www.monkproject.org.

Nielsen, J. 2000. *Designing web usability: The practice of simplicity*. Indianapolis, IN: New Riders.

Shneiderman, B., Williamson, C. and Ahlberg, C. (1992). Dynamic Queries: Database Searching by Direct Manipulation. *Proceedings of the SIGCHI conference on human factors in computing systems*. pp. 669-670. http://doi.acm.org/10.1145/142750.143082. Accessed March 10, 2006.

Rules of the Order: The Sociology of Large, Multi-Institutional Software Development Projects

Stephen Ramsay

Digital Humanities is increasingly turning toward tool development, both as a research modality and as a practical means of exploiting the vast collections of digital resources that have been amassing over the last twenty years. Like the creation of digital collections, software development is an inherently collaborative activity, but unlike digital collections, complex software systems do not easily allow for loosely aggregated collaborations. With digital text collections, it might be possible to have several contributors working at their own pace—even at several institutions—with a high level of turnover and varying local workflows, but this is rarely possible with software development. However discretized the components of a system might be, the final product is invariably a tightly integrated whole with numerous dependencies and a rigidly unified architectural logic.

For most digital humanities practitioners, amassing a team of developers almost requires that work be distributed across institutions and among a varied group of people. Any non-trivial application requires experts from a number of different development subspecialties, including such areas as interface design, relational database management, programming, software engineering, and server administration (to name only a few). Few research centers in DH have the staff necessary for undertaking large application development projects, and even the ones that do quickly find that crossdepartmental collaborations are needed to assemble the necessary expertise.

There are a great many works on and studies of project management for software development, but very few of these resources address the particular situation of a widely distributed group of developers. Most of the popular methodologies (Rational Unified Process, Theory of Contraints, Event Chain) assume a group of developers who are all in one place with a clear management structure. More recent "agile methods" (such as Extreme Project Management) invariably stipulate a small groups of developers. Reading the ever useful (if slightly dispiriting) works on project failure, such as Brooks's famous 1975 book The Mythical Man Month, one wonders whether project success is even possible for a distributed group.

Development of open source software might seem an instructive counter-example of project success arising from massive, distributed teams of developers, but as several studies have shown, the core development team for even the largest software development projects (Apache, Mozilla, and the Linux kernel, for example) are considerably smaller than the hundreds of people listed as contributors might suggest, and in most cases, the project was able to build upon a core code base that was originally developed in a single location with only a few developers.

The MONK Project, which endeavors to build a portable system for undertaking text analysis and visualization with large, full text literary archives, might seem a perfect storm of management obstacles. The project has two primary investigators, half a dozen co-investigators, nearly forty project participants at eight different institutions, and includes professors of all ranks, staff programmers, research fellows, and graduate students. The intended system, moreover, involves dozens of complex components requiring expertise in a number of programming languages, architectural principles, and data storage methods, and facility with a broad range of text analytical techniques (including text mining and statistical natural language processing).

This paper offers some reflections on the trials and triumphs of project management at this scale using the MONK Project as an example. It focuses particularly on the sociology of academia as a key element to be acknowledged in management practice, and concludes that honoring existing institutional boundaries and work patterns is essential to maintaining a cohesive (if, finally, hierarchical) management structure. It also considers the ways in which this apparently simple notion runs counter to the cultural practices and decision making methods of ordinary academic units—practices which are so ingrained as to seem almost intuitive to many.

The MONK Project has from the beginning studiously avoided references to cloisters and habits in its description of itself. This paper offers a playful analogy in contrast to that tradition, by suggesting that most of what one needs to know about project management was set forth by Benedict of Nursia in his sixth-century Rule for monks. Particular emphasis is placed on Benedict's belief that social formations become increasingly fractured as they move from the cenobitic ideal (which, I suggest, is roughly analogous to the structure of the typical research center), and his belief that the Abbot (like the modern project manager) must assert a kind of benevolent despotism over those within his charge, even if the dominant values of the group privilege other, more democratic forms of leadership.

References

Benedict of Nursia. *Regula*. Collegeville, Minn.: Liturgical Press, 1981.

Brooks, Frederick. *The Mythical Man Month: Essays on Software Development.* Reading, MA: Addison-Wesley, 1975.

DeCarlo, Douglas. *Extreme Project Management.* New York: Wiley-Jossey, 2004.

Kruchten, Philippe. *The Rational Unified Process: An Introduction*. 3rd ed. Reading, Mass: Addison-Wesley, 2003.

Leach, Lawrence P. *Critical Chain Project Management.* 2nd ed. Norwood, MA.: Artech, 2004.

Mockus, Audris. Roy T. Fielding, James D. Herbsleb. "Two Case Studies of Open Source Software Development: Apache and Mozilla." *ACM Transactions on Software Engineering and Methodology* 11.3 (2002): 309-46.

Skinner, David C. *Introduction to Decision Analysis*. 2nd ed. Sugar Land, TX: Probabilistic, 1999.

Aspects of Sustainability in Digital Humanities

Georg Rehm
georg.rehm@uni-tuebingen.de
Tübingen University, Germany

Andreas Witt
andreas.witt@uni-tuebingen.de
Tübingen University, Germany

The three papers of the proposed session, "Aspects of Sustainability in Digital Humanities", examine the increasingly important topic of sustainability from the point of view of three different fields of research: library and information science, cultural heritage management, and linguistics.

Practically all disciplines in science and the humanities are nowadays confronted with the task of providing data collections that have a very high degree of sustainability. This task is not only concerned with the long-term archiving of digital resources and data collections, but also with aspects such as, for example, interoperability of resources and applications, data access, legal issues, field-specific theoretical approaches, and even political interests.

The proposed session has two primary goals. Each of the three papers will present the most crucial problems that are relevant for the task of providing sustainability within the given field or discipline. In addition, each paper will talk about the types of digital resources and data collections that are in use within the respective field (for example, annotated corpora and *syntactic treebanks* in the field of linguistics). The main focus, however, lies in working on the distinction between field-specific and universal aspects of sustainability so that the three fields that will be examined – library and information science, cultural heritage management, linguistics – can be considered case studies in order to come up with a more universal and all-encompassing angle on sustainability. Especially for introductory texts and field – *independent* best-practice guidelines on sustainability it is extremely important to have a solid distinction between universal and field-specific aspects. The same holds true for the integration of sustainability-related informational units into field-independent markup languages that have a very broad scope of potential applications, such as the TEI guidelines published by the Text Encoding Initiative.

Following are short descriptions of the three papers:

The paper "Sustainability in Cultural Heritage Management" by Øyvind Eide, Christian-Emil Ore, and Jon Holmen discusses technical and organisational aspects of sustainability with regard to cultural heritage information curated by institutions such as, for example, museums. Achieving organisational sustainability is a task that not only applies to the staff of a museum but also to education and research institutions, as well as to national and international bodies responsible for our common heritage.

Vital to the sustainability of collections is information about the collections themselves, as well as individual items in those collections. "Sustaining Collection Value: Managing Collection/ Item Metadata Relationships", by Allen H. Renear, Richard Urban, Karen Wickett, Carole L. Palmer, and David Dubin, examines the difficult problem of managing collection level metadata in order to ensure that the context of the items in a collection is accessible for research and scholarship. They report on ongoing research and also have preliminary suggestions for practitioners.

The final paper "Sustainability of Annotated Resources in Linguistics", by Georg Rehm, Andreas Witt, Erhard Hinrichs, and Marga Reis, provides an overview of important aspects of sustainability with regard to linguistic resources. The authors demonstrate which of these several aspects can be considered specific for the field of linguistics and which are more general.

Paper 1: Sustainability in Cultural Heritage Management

Øyvind Eide, Christian-Emil Ore, Jon Holmen
University of Oslo, Norway

Introduction

During the last decades, a large amount of information in cultural heritage institutions have been digitised, creating the basis for many different usage scenarios. We have been working in this area for the last 15 years, through projects such as the Museum Project in Norway (Holmen et al., 2004). We have developed routines, standardised methods and software for digitisation, collection management, research and education. In this paper, we will discuss long term sustainability of digital cultural heritage information. We will discuss the creation of sustainable digital collections, as well as some problems we have experienced in this process.

We have divided the description of sustainability in three parts. First we will describe briefly the technical part of sustainability work (section 2). After all, this is a well known research area on its own, and solutions to many of the problems at hand are known, although they may be hard to implement. We will then use the main part of the paper to discuss what we call organisational sustainability (section 3), which may be even more important than the technical part in the future — in our opinion, it may also be more difficult to solve. Finally, we briefly address the scholarly part of sustainability (section 4).

Technical Sustainability

Technical sustainability is divided in two parts: preservation of the digital bit patterns and the ability to interpret the bit pattern according to the original intention. This is an area where important work is being done by international bodies such as UNESCO (2003), as well as national organisations such as the Library of Congress in the USA (2007), and the Digital Preservation Coalition in the UK (DPC,2001).

It is evident that the use of open, transparent formats make sit easier to use preserved digital content. In this respect XML encoding is better compared to proprietary word processor formats, and uncompressed TIFF is more transparent than company-developed compressed image formats. In a museum context, though, there is often a need to store advanced reproductions of objects and sites, and there have been problems to find open formats able to represent, in full, content exported from proprietary software packages. An example of this is CAD systems, where the open format SVG does not have the same expressive power as the proprietary DXF format (Westcott, 2005, p.6). It is generally a problem for applications using new formats, especially when they are heavily dependent upon presentation.

Organisational Sustainability

Although there is no sustainability without the technical part, described above, taken care of, the technical part alone is not enough. The organisation of the institution responsible for the information also has to be taken into consideration.

In an information system for memory institutions it is important to store the history of the information. Digital versions of analogue sources should be stored as accurate replica, with new information linked to this set of historical data so that one always has access to up-to-date versions of the information, as well as to historical stages in the development of the information (Holmenetal.,2004, p.223).

To actually store the files, the most important necessity is large, stable organisations taking responsibility. If the responsible institution is closed without a correct transfer of custody for the digital material, it can be lost easily. An example of this is the Newham Archive (Dunning, 2001) incident. When the Newham Museum Archaeological Service was closed down, only a quick and responsible reaction of the sacked staff saved the result of ten years of work in the form of a data dump on floppies.

Even when the data files are kept, lack of necessary metadata may render them hard to interpret. In the Newham case the data were physically saved but a lot of work was needed to read the old data formats, and some of the information was not recoverable. Similar situations may even occur in large, stable organisations. The Bryggen Museum in Bergen, Norway, is a part of the University Museum in Bergen and documents the large excavation of the medieval town at the harbour in Bergen which took place from the 1950s to the 1970s. The museum stored the information in a large database.

Eventually the system became obsolete and the database files were stored at the University. But there were no explicit routines for the packing and future unpacking of such digital information. Later, when the files were imported into a new system, parts of the original information were not recovered. Fortunately all the excavation documentation was originally done on paper so in principle no information was lost.

Such incidents are not uncommon in the museum world. A general problem, present in both examples above, is the lack of metadata. The scope of each database table and column is well known when a system is developed, but if it is not documented, such meta-information is lost.

In all sectors there is a movement away from paper to born digital information. When born digital data based on archaeological excavations is messed up or lost – and we are afraid this will happen – then parts of our cultural heritage are lost forever. An archaeological excavation destroys its own sources and an excavation cannot be repeated. For many current excavation projects a loss of data like the Bryggen Museum incident would have been a real catastrophe.

The Newham example demonstrates weak planning for negative external effects on information sustainability, whereas the Bergen example shows how a lack of proper organisational responsibility for digital information may result in severe information loss. It is our impression that in many memory institutions there is too little competence on how to introduce information technologies in an organisation to secure both interchange of information between different parts of the organisation and long-term sustainability of the digital information. A general lack of strategies for long term preservation is documented in a recent Norwegian report (Gausdal, 2006, p.23).

When plans are made in order to introduce new technology and information systems into an organisation one has to adapt the system to the organisation or the organisation to the system. This is often neglected and the information systems are not integrated in the everyday work of the staff. Thus, the best way to success is to do this in collaboration and understanding with the employees. This was pointed out by Professor Kristen Nygaard. In a paper published in 1992 describing the uptake of Simula I from 1965 onwards, Nygaard states: "It was evident that the Simula-based analyses were going to have a strong influence on the working conditions of the employees: job content, work intensity and rhythm, social cooperation patterns were typical examples" (Nygaard, 1992, p. 53). Nygaard focused on the situation in the ship building industry, which may be somewhat distant from the memory institutions. Mutate mutandis, the human mechanisms are the same. There is always a risk of persons in the organisation sabotaging or neglecting new systems.

Scholarly Sustainability

When a research project is finished, many researchers see the report or articles produced as the only output, and are confident that the library will take care of their preservation. But research in the humanities and beyond are often based on material collected by the researcher, such as ethnographic objects, sound recordings, images, and notes. The scholarly conclusions are then based on such sources. To sustain links from sources to testable conclusions, they have to be stored so that they are accessible to future researchers. But even in museums, this is often done only in a partial manner. Objects may find their way into the collections. But images, recordings and notes are often seen as the researcher's private property and responsibility, and may be lost when her career is ended. Examples of this are hard to document, though, because such decisions are not generally made public.

Conclusion

Sustainability of data in the cultural heritage sector is, as we have seen, not just a technical challenge. The sustainability is eased by the use of open and transparent standards. It is necessary to ensure the existence of well funded permanent organisation like national archives and libraries. Datasets from museums are often not considered to lie within the preservation scope of the existing organisations. Either this has to be changed or the large museums have to act as digital archives of museum data in general. However, the most important measure to ensure sustainability is to increase the awareness of the challenge among curators and scholars. If not, large amounts of irreplaceable research documentation will continue to be lost.

References

DPC(2001): "Digital preservation coalition". http://www.dpconline.org/graphics.

Dunning, Alastair(2001): "Excavating data: Retrieving the Newham archive". http://ahds.ac.uk/creating/case-studies/newham/.

Gausdal, Ranveig Låg (editor) (2006): *Cultural heritage for all — on digitisation, digital preservation and digital dissemination in the archive, library and museum sector.* A report by the Working Group on Digitisation, the Norwegian Digital Library. ABM-Utvikling.

Holmen, Jon; Ore, Christian-Emil and Eide, Øyvind(2004): "Documenting two histories at once: Digging into archaeology". In: Enter the Past. The E-way into the Four Dimensions of Cultural Heritage. BAR, BAR International Series 1227, pp. 221-224

LC(2007): "The library of congress. Digital preservation. http://www.digitalpreservation.gov.

Nygaard, Kristen(1992): "How many choices do we make? How many are difficult? In: *Software Development and Reality Construction*, edited by Floyd C., Züllighoven H., Budde R., and R., Keil-Slawik. Springer-Verlag, Berlin, pp. 52-59.

UNESCO (2003): "Charter on the preservation of digital heritage. Adopted at the 32nd session of the 9/10 general conference of UNESCO" Technical Report, UNESCO. http://portal.unesco.org/ci/en/files/13367/10700115911Charter_en.pdf/Charter_en.pdf.

Westcott, Keith(2005): Preservation Handbook. *Computer Aided Design (CAD)*. Arts and Humanities Data Service. http://ahds.ac.uk/preservation/cad-preservation-handbook.pdf.

Paper 2: Sustaining Collection Value: Managing Collection/Item Metadata Relationships

Allen H. Renear, Richard Urban, Karen Wickett, Carole L. Palmer, David Dubin
University of Illinois at Urbana-Champaign, USA

Introduction

Collections of texts, images, artefacts, and other cultural objects are usually designed to support particular research and scholarly activities. Toward that end collections themselves, as well as the items in the collections, are carefully developed and described. These descriptions indicate such things as the purpose of the collection, its subject, the method of selection, size, nature of contents, coverage, completeness, representativeness, and a wide range of summary characteristics, such as statistical features. This information enables collections to function not just as aggregates of individual data items but as independent entities that are in some sense more than the sum of their parts, as intended by their creators and curators [Currall et al., 2004,Heaney, 2000,Lagoze et al., 2006, Palmer, 2004]. Collection-level metadata, which represents this information in computer processable form, is thus critical to the distinctive intellectual and cultural role of collections as something more than a set of individual objects.

Unfortunately, collection-level metadata is often unavailable or ignored by retrieval and browsing systems, with a corresponding loss in the ability of users to find, understand, and use items in collections [Lee, 2000, Lee, 2003, Lee, 2005, Wendler, 2004]. Preventing this loss of information is particularly difficult, and particularly important, for "metasearch", where item-level descriptions are retrieved from a number of different collections simultaneously, as is the case in the increasingly distributed search environment [Chistenson and Tennant, 2005, Dempsey, 2005, Digital Library Federation, 2005, Foulonneau et al., 2005, Lagoze et al., 2006, Warner et al., 2007].

The now familiar example of this challenge is the "'on a horse' problem", where a collection with the collection-level subject "Theodore Roosevelt" has a photograph with the item-level annotation "on a horse" [Wendler, 2004] Item-level access across multiple collections (as is provided not only by popular Internet search engines, but also specialized federating systems, such as OAI portals) will not allow the user to effectively use a query with keywords "Roosevelt" and "horse" to find this item, or, if the item is retrieved using item-level metadata alone, to use collection-level information to identify the person on the horse as Roosevelt.

The problem is more complicated and consequential than the example suggests and the lack of a systematic understanding of the nature of the logical relationships between collection-level metadata and item-level metadata is an obstacle to the development of remedies. This understanding is what is required not only to guide the development of context-aware search and exploitation, but to support management and curation policies as well. The problem is also timely: even as recent research continues to confirm the key role that collection context plays in the scholarly use of information resources [Brockman et al., 2001, Palmer, 2004], the Internet has made the context-free searching of multiple collections routine.

In what follows we describe our plans to develop a framework for classifying and formalizing collection-level/item-level metadata relationships. This undertaking is part of a larger project, recently funded by US Institute for Museum and Library Services (IMLS), to develop tools for improved retrieval and exploitation across multiple collections.[1]

Varieties of Collection/Item Metadata Relationships

In some cases the relationship between collection-level metadata and item-level metadata attributes appears similar to non-defeasible inheritance. For instance, consider the Dublin Core Collections Application Profile element *marcrel:OWN*, adapted from the MARC cataloging record standard. It is plausible that within many legal and institutional contexts whoever owns a collection owns each of the items in the collection, and so if a collection has a value for the *marcrel:OWN* attribute then each member of the collection will have the same value for *marcrel:OWN*. (For the purpose of our example it doesn't matter whether or not this is actually true of *marcrel:OWN*, only that some attributes are sometimes used by metadata librarians with an understanding of this sort, while others, such as *dc:identifier*, are not).

In other cases the collection-level/item-level metadata relationship is almost but not quite this simple. Consider the collection-level attribute *myCollection:itemType*, intended to characterize the type of objects in a collection, with values such as "image," "text," "software," etc. (we assume heterogeneous collections).[2] Unlike the preceding case we cannot conclude that if a collection has the value "image" for *myCollection:itemType* then the items in that collection also have the value "image" for that same attribute. This is because an item which is an image is not itself a collection of images and therefore cannot have a non-null value for *myCollection:itemType*.

However, while the rule for propagating the information represented by *myCollection:itemType* from collections to items is not simple propagation of attribute and value, it is nevertheless simple enough: if a collection has a value, say "image," for *myCollection:itemType*, then the items in the collection have the same value, "image" for a corresponding attribute, say, *myItem:type*, which indicates the type of item (cf. the Dublin Core metadata element *dc:type*). The attribute myItem:type has the same domain of values as *myCollection:itemType*, but a different semantics.

When two metadata attributes are related as *myCollection: itemType* and *myItem:type* we might say the first can be v-converted to the other. Roughly: a collection-level attribute **A** v-converts to an item-level attribute **B** if and only if whenever a collection has the value *z* for **A**, every item in the collection has the value *z* for **B**. This is the simplest sort of convertibility — the attribute changes, but the value remains the same. Other sorts of conversion will be more complex. We note that the sort of propagation exemplified by *marcrel:0WN* is a special case of v-convertibility: *marcrel:0WN* v-converts to itself.

This analysis suggests a number of broader issues for collection curators. Obviously the conversion of collection-level metadata to item-level metadata, when possible, can improve discovery and exploitation, especially in item-focused searching across multiple collections. But can we even in the simplest case be confident of conversion without loss of information? For example, it may be that in some cases an "image" value for *myCollection:itemType* conveys more information than the simple fact that each item in the collection has "image" value for *myItem:type*.

Moreover there are important collection-level attributes that both (i) resist any conversion and (ii) clearly result in loss of important information if discarded. Intriguingly these attributes turn out to be carrying information that is very tightly tied to the distinctive role the collection is intended to play in the support of research and scholarship. Obvious examples are metadata indicating that a collection was developed according to some particular method, designed for some particular purpose, used in some way by some person or persons in the past, representative (in some respect) of a domain, had certain summary statistical features, and so on. This is precisely the kind of information that makes a collection valuable to researchers, and if it is lost or inaccessible, the collection cannot be useful, as a collection, in the way originally intended by its creators.

The DCC/CIMR Project

These issues were initially raised during an IMLS Digital Collections and Content (DCC) project, begun at the University of Illinois at Urbana-Champaign in 2003. That project developed a collection-level metadata schema[3] based on the RSLP[4] and Dublin Core Metadata Initiative (DCMI) and created a collection registry for all the digital collections funded through the Institute of Museum and Library Services National Leadership Grant (NLG) since 1998, with some Library Services and Technology Act (LSTA) funded collections included since 2006[5]. The registry currently contains records for 202 collections. An item-level metadata repository was also developed, which so far has harvested 76 collections using the OAI-PMH protocol[6].

Our research initially focused on overcoming the technical challenges of aggregating large heterogeneous collections of item-level records and gathering collections descriptions from contributors. We conducted studies on how content contributors conceived of the roles of collection descriptions in digital environments [Palmer and Knutson, 2004, Palmer et al., 2006], and conducted preliminary usability work. These studies and related work on the CIC Metadata Portal[7], suggest that while the boundaries around digital collections are often blurry, many features of collections are important for helping users navigate and exploit large federated repositories, and that collection and item-level descriptions should work in concert to benefit certain kinds of user queries [Foulonneau et al., 2005].

In 2007 we received a new three year IMLS grant to continue the development of the registry and to explore how a formal description of collection-level/item-level metadata relationships could help registry users locate and use digital items. This latter activity, CIMR, (Collection/Item Metadata Relationships), consists of three overlapping phases. The first phase is developing a logic-based framework of collection-level/ itemlevel metadata relationships that classifies metadata into varieties of convertibility with associated rules for propagating information between collection and item levels and supporting further inferencing. Next we will conduct empirical studies to see if our conjectured taxonomy matches the understanding and behavior of metadata librarians, metadata specification designers, and registry users. Finally we will design and implement pilot applications using the relationship rules to support searching, browsing, and navigation of the DCC Registry. These applications will include non-convertible and convertible collection-level/item-level metadata relationships.

One outcome of this project will be a proposed specification for a metadata classification code that will allow metadata specification designers to indicate the collectionlevel/item-level metadata relationships intended by their specification. Such a specification will in turn guide metadata librarians in assigning metadata and metadata systems designers in designing systems that can mobilize collection level metadata to provide improved searching, browsing, understanding, and use by end users. We will also draft and make electronically available RDF/OWL bindings for the relationship categories and inference rules.

Preliminary Guidance for Practitioners

A large part of the problem of sustainability is ensuring that information will be valuable, and as valuable as possible, to multiple audiences, for multiple purposes, via multiple tools, and over time. Although we have only just begun this project, already some preliminary general recommendations can be made to the different stakeholders in collection management. Note that tasks such as propagation must be repeated not only as new objects are added or removed but, as new information about objects and collections becomes available.

For metadata standards developers:

1. Metadata standards should explicitly document the relationships between collectionlevel metadata and item-level metadata. Currently we have neither the understanding nor the formal mechanisms for such documentation but they should be available soon.

For systems designers:

2. Information in convertible collection-level metadata should be propagated to items in order to make contextual information fully available to users, especially users working across multiple collections. (This is not a recommendation for how to manage information internally, but for how to represent it to the user; relational tables may remain in normal forms.)

3. Information in item-level metadata should, where appropriate, be propagated to collection level metadata.

4. Information in non-convertible collection-level metadata must, to the fullest extent possible, be made evident and available to users.

For collection managers:

5. Information in non-convertible metadata must be a focus of data curation activities if collections are to retain and improve their usefulness over time.

When formal specifications and tools based on them are in place, relationships between metadata at the collection and item levels will integrated more directly into management and use. In the mean time, attention and sensitivity to the issues we raise here can still improve matters through documentation and policies, and by informing system design.

Acknowledgments

This research is supported by a 2007 IMLS NLG Research & Demonstration grant hosted by the GSliS Center for Informatics Research in Science and Scholarship (CIRSS). Project documentation is available at http://imlsdcc.grainger.uiuc. edu/about.asp#documentation. We have benefited considerably from discussions with other DCC/CIMR project members and with participants in the IMLS DCC Metadata Roundtable, including: Timothy W. Cole, Thomas Dousa, Dave Dubin, Myung-Ja Han, Amy Jackson, Mark Newton , Carole L. Palmer, Sarah L. Shreeves, Michael Twidale, Oksana Zavalina

References

[Brockman et al., 2001] Brockman, W., Neumann, L., Palmer, C. L., and Tidline, T. J. (2001). *Scholarly Work in the Humanities and the Evolving Information Environment*. Digital Library Federation/Council on Library and Information Resources, Washington, D.C.

[Chistenson and Tennant, 2005] Chistenson, H. and Tennant, R. (2005). *Integrating information resources: Principles, technologies, and approaches*. Technical report, California Digital Library.

[Currall et al., 2004] Currall, J., Moss, M., and Stuart, S. (2004). What is a collection? *Archivaria*, 58:131–146.

[Dempsey, 2005] Dempsey, L. (2005). From metasearch to distributed information environments. Lorcan Dempsey's weblog. Published on the World Wide Web at http:// orweblog.oclc.org/archives/000827.html.

[Digital Library Federation, 2005] Digital Library Federation (2005). *The distributed library: OAI for digital library aggregation: OAI scholars advisory panel meeting, June 20–21, Washington, D.C.* Published on the World Wide Web at http:// www.diglib.org/architectures/oai/imls2004/OAISAP05.htm.

[Foulonneau et al., 2005] Foulonneau, M., Cole, T., Habing, T. G., and Shreeves, S. L. (2005). Using collection descriptions to enhance an aggregation of harvested itemlevel metadata. In ACM, editor, *Proceedings of the 5th ACM/IEEE-CS Joint Conference on Digital Libraries*, pages 32–41, New York. ACM/ IEEE-CS, ACM Press.

[Heaney, 2000] Heaney, M. (2000). *An analytical model of collections and their catalogues*. Technical report, UK Office for Library and Information Science.

[Lagoze et al., 2006] Lagoze, C., Krafft, D., Cornwell, T., Dushay, N., Eckstrom, D., and Saylor, J. (2006). Metadata aggregation and automated digital libraries: A retrospective on the NSDL experience. In *Proceedings of the 6th ACM/IEEE-CS Joint Conference on Digital Libraries*, pages 230–239, New York. ACM/IEEE-CS, ACM Press.

[Lee, 2000] Lee, H. (2000). What is a collection? *Journal of the American Society for Information Science*, 51(12):1106–1113.

[Lee, 2003] Lee, H. (2003). Information spaces and collections: Implications for organization. *Library & Information Science Research,* 25(4):419–436.

[Lee, 2005] Lee, H. (2005). The concept of collection from the user's perspective. *Library Quarterly*, 75(1):67–85.

[Palmer, 2004] Palmer, C. (2004). *Thematic research collections*, pages 348–365. Blackwell, Oxford.

[Palmer and Knutson, 2004] Palmer, C. L. and Knutson, E. (2004). Metadata practices and implications for federated collections. In *Proceedings of the 67th ASIS&T Annual Meeting (Providence, RI, Nov. 12–17, 2004)*, volume 41, pages 456–462.

[Palmer et al., 2006] Palmer, C. L., Knutson, E., Twidale, M., and Zavalina, O. (2006). Collection definition in federated digital resource development. In *Proceedings of the 69th ASIS&T Annual Meeting (Austin, TX, Nov. 3–8, 2006)*, volume 43.

[Warner et al., 2007] Warner, S., Bakaert, J., Lagoze, C., Lin, X., Payette, S., and Van de Sompel, H. (2007). Pathways: Augmenting interoperability across scholarly repositories. *International Journal on Digital Libraries*, 7(1-2):35–52.

[Wendler, 2004] Wendler, R. (2004). *The eye of the beholder: Challenges of image description and access at Harvard*., pages 51–56. American Library Association, Chicago.

Notes

1 Information about the IMLS Digital Collections and Content project can be found at: http://imlsdcc.grainger.uiuc.edu/about.asp.

2 In our examples we will use imaginary metadata attributes. The namespace prefix "*myCollection:*" indicates collection-level attributes and the prefix "*myItem:*" indicates item-level attributes. No assumptions should be made about the semantics of these attributes other than what is stipulated for illustration. The current example, *myCollection: itemType*, does intentionally allude to *cld:itemType* in the Dublin Core Collections Application Profile, and "image," "text," "software," are from the DCMI Type Vocabulary; but our use of *myCollection:itemType* differs from *cld:itemType* in entailing that all of the items the collection are of the indicated type.

3 General overview and detailed description of the IMLS DCC collection description scheme are available at: http://imlsdcc.grainger.uiuc.edu/CDschema_overview.asp

4 http://www.ukoln.ac.uk/metadata/rslp/

5 http://www.imls.gov/

6 http://www.openarchives.org/OAI/openarchivesprotocol.html

7 http://cicharvest.grainger.uiuc.edu/

Paper 3: Sustainability of Annotated Resources in Linguistics

Georg Rehm, Andreas Witt, Erhard Hinrichs, Marga Reis

Introduction

In practically all scientific fields the task of ensuring the sustainability of resources, data collections, personal research journals, and databases is an increasingly important topic – linguistics is no exception (Dipper et al., 2006, Trilsbeek and Wittenburg, 2006). We report on ongoing work in a project that is concerned with providing methods, tools, best-practice guidelines, and solutions for *sustainable* linguistic resources. Our overall goal is to make sure that a large and very heterogeneous set of ca. 65 linguistic resources will be accessible, readable, and processible by interested parties such as, for example, other researchers than the ones who originally created said resources, in five, ten, or even 20 years time. In other words, the agency that funded both our project as well as the projects who created the linguistic resources – the German Research Foundation – would like to avoid a situation in which they have to fund yet another project to (re)create a corpus for whose creation they already provided funding in the past, but the "existing" version is no longer available or readable due to a proprietary file format, because it has been locked away in an academic's hidden vault, or the person who developed the annotation format can no longer be asked questions concerning specific details of the custom-built annotation format (Schmidt et al., 2006).

Linguistic Resources: Aspects of Sustainability

There are several text types that linguists work and interact with on a frequent basis, but the most common, by far, are linguistic corpora (Zinsmeister et al., 2007). In addition to rather simple word and sentence collections, empirical sets of grammaticality judgements, and lexical databases, the linguistic resources our sustainability project is primarily confronted with are linguistic corpora that contain either texts or transcribed speech in several languages; they are annotated using several incompatible annotation schemes. We developed XML-based tools to normalise the existing resources into a common approach of representing linguistic data (Wörner et al., 2006, Witt et al., 2007b) and use interconnected OWL ontologies to represent knowledge about the individual annotation schemes used in the original resources (Rehm et al., 2007a).

Currently, the most central aspects of sustainability for linguistic resources are:

- markup languages
- metadata encoding
- legal aspects (Zimmermann and Lehmberg, 2007, Lehmberg et al., 2007a,b, Rehm et al., 2007b,c, Lehmberg et al., 2008),
- querying and search (Rehm et al., 2007a, 2008a, Söhn et al., 2008), and
- best-practice guidelines (see, for example, the general guidelines mentioned by Bird and Simons, 2003).

None of these points are specific to the field of linguistics, the solutions, however, are. This is exemplified by means of two of these aspects.

The use of markup languages for the annotation of linguistic data has been discussed frequently. This topic is also subject to standardisation efforts. A separate ISO Group, ISO TC37 SC4, deals with the standardisation of linguistic annotations.

Our project developed an annotation architecture for linguistic corpora. Today, a linguistic corpus is normally represented by a single XML file. The underlying data structures most often found are either trees or unrestricted graphs. In our approach we transform an original XML file to several XML files, so that each file contains the same textual content. The markup of these files is different. Each file contains annotations which belong to a single annotation layer. A data structure usable to model the result documents is a multi-rooted tree. (Wörner et al., 2006, Witt et al., 2007a,b, Lehmberg and Wörner, 2007).

The specificities of linguistic data also led to activities in the field of metadata encoding and its standardisation. Within our project we developed an approach to handle the complex nature of linguistic metadata (Rehm et al., 2008b) which is based on the metadata encoding scheme described by the TEI. (Burnard and Bauman, 2007). This method of metadata representation splits the metadata into the 5 different levels the primary information belongs to. These levels are: (1) setting, i.e. the situation in which the speech or dialogue took place; (2) raw data, e.g., a book, a piece of paper, an audio or video recording of a conversation etc.; (3) primary data, e.g., transcribed speech, digital texts etc.; (4) annotations, i.e., (linguistic) markup that add information to primary data; and (5) corpus, i.e. a collection of primary data and its annotations.

All of these aspects demonstrate that it is necessary to use field specific as well as generalised methodologies to approach the issue "Sustainability of Linguistic Resources".

References

Bird, S. and Simons, G. (2003). Seven dimensions of portability for language documentation and description. *Language*, 79:557–582.

Burnard, L. and Bauman, S., editors (2007). *TEI P5: Guidelines for Electronic Text Encoding and Interchange*. Text Encoding Initiative Consortium.

Dipper, S., Hinrichs, E., Schmidt, T., Wagner, A., and Witt, A. (2006). Sustainability of Linguistic Resources. In Hinrichs, E., Ide, N., Palmer, M., and Pustejovsky, J., editors, *Proceedings of the LREC 2006 Satellite Workshop Merging and Layering Linguistic Information*, pages 48–54, Genoa, Italy.

Lehmberg, T., Chiarcos, C., Hinrichs, E., Rehm, G., and Witt, A. (2007a). Collecting Legally Relevant Metadata by Means of a Decision-Tree-Based Questionnaire System. In Schmidt, S., Siemens, R., Kumar, A., and Unsworth, J., editors, *Digital Humanities 2007*, pages 164–166, Urbana-Champaign, IL, USA. ACH, ALLC, Graduate School of Library and Information Science, University of Illinois, Urbana-Champaign.

Lehmberg, T., Chiarcos, C., Rehm, G., and Witt, A. (2007b). Rechtsfragen bei der Nutzung und Weitergabe linguistischer Daten. In Rehm, G., Witt, A., and Lemnitzer, L., editors, *Datenstrukturen für linguistische Ressourcen und ihre Anwendungen – Data Structures for Linguistic Resources and Applications: Proceedings of the Biennial GLDV Conference 2007*, pages 93–102. Gunter Narr, Tübingen.

Lehmberg, T., Rehm, G., Witt, A., and Zimmermann, F. (2008). Preserving Linguistic Resources: Licensing – Privacy Issues – Mashups. *Library Trends*. In print.

Lehmberg, T. and Wörner, K. (2007). Annotation Standards. In Lüdeling, A. and Kytö, M., editors, *Corpus Linguistics, Handbücher zur Sprach-und Kommunikationswissenschaft (HSK)*. de Gruyter, Berlin, New York. In press.

Rehm, G., Eckart, R., and Chiarcos, C. (2007a). An OWL-and XQuery-Based Mechanism for the Retrieval of Linguistic Patterns from XML-Corpora. In Angelova, G., Bontcheva, K., Mitkov, R., Nicolov, N., and Nikolov, N., editors, *International Conference Recent Advances in Natural Language Processing (RANLP 2007)*, pages 510–514, Borovets, Bulgaria.

Rehm, G., Eckart, R., Chiarcos., C., Dellert, J. (2008a). Ontology-Based XQuery'ing of XML-Encoded Language Resources on Multiple Annotation Layer. In *Proceedings of LREC 2008*, Marrakech, Morocco.

Rehm, G., Schonefeld, O., Witt, A., Lehmberg, T., Chiarcos, C., Bechara, H., Eishold, F., Evang, K., Leshtanska, M., Savkov, A., and Stark, M. (2008b). The Metadata-Database of a Next Generation Sustainability Web-Platform for Language Resources. In *Proceedings of LREC 2008*, Marrakech, Morocco.

Rehm, G., Witt, A., Zinsmeister, H., and Dellert, J. (2007b). Corpus Masking: Legally Bypassing Licensing Restrictions for the Free Distribution of Text Collections. In Schmidt, S., Siemens, R., Kumar, A., and Unsworth, J., editors, *Digital Humanities 2007*, pages 166–170, Urbana-Champaign, IL, USA. ACH, ALLC, Graduate School of Library and Information Science, University of Illinois, Urbana-Champaign.

Rehm, G., Witt, A., Zinsmeister, H., and Dellert, J. (2007c). Masking Treebanks for the Free Distribution of Linguistic Resources and Other Applications. In *Proceedings of the Sixth International Workshop on Treebanks and Linguistic Theories (TLT 2007)*, number 1 in Northern European Association for Language Technology Proceedings Series, pages 127–138, Bergen, Norway.

Schmidt, T., Chiarcos, C., Lehmberg, T., Rehm, G., Witt, A., and Hinrichs, E. (2006). Avoiding Data Graveyards: From Heterogeneous Data Collected in Multiple Research Projects to Sustainable Linguistic Resources. In *Proceedings of the E-MELD 2006 Workshop on Digital Language Documentation: Tools and Standards – The State of the Art*, East Lansing, Michigan.

Söhn, J.-P., Zinsmeister, H., and Rehm, G. (2008). Requirements of a User-Friendly, General-Purpose Corpus Query Interface. In *Proceedings of LREC 2008 workshop on Sustainability of Language Resources and Tools for Natural Language Processing*, Marrakech, Morocco.

Trilsbeek, P. and Wittenburg, P. (2006). Archiving Challenges. In Gippert, J., Himmelmann, N. P., and Mosel, U., editors, *Essentials of Language Documentation,* pages 311–335. Mouton de Gruyter, Berlin, New York.

Witt, A., Rehm, G., Lehmberg, T., and Hinrichs, E. (2007a). Mapping Multi-Rooted Trees from a Sustainable Exchange Format to TEI Feature Structures. In *TEI@20: 20 Years of Supporting the Digital Humanities*. The 20th Anniversary Text Encoding Initiative Consortium Members' Meeting, University of Maryland, College Park.

Witt, A., Schonefeld, O., Rehm, G., Khoo, J., and Evang, K. (2007b). On the Lossless Transformation of Single-File, Multi-Layer Annotations into Multi-Rooted Trees. In Usdin, B. T., editor, *Proceedings of Extreme Markup Languages 2007*, Montréal, Canada.

Wörner, K., Witt, A., Rehm, G., and Dipper, S. (2006). Modelling Linguistic Data Structures. In Usdin, B. T., editor, *Proceedings of Extreme Markup Languages 2006*, Montréal, Canada.

Zimmermann, F. and Lehmberg, T. (2007). Language Corpora – Copyright – Data Protection: The Legal Point of View. In Schmidt, S., Siemens, R., Kumar, A., and Unsworth, J., editors, *Digital Humanities 2007*, pages 162–164, Urbana-Champaign, IL, USA. ACH, ALLC, Graduate School of Library and Information Science, University of Illinois, Urbana-Champaign.

Zinsmeister, H., Kübler, S., Hinrichs, E., and Witt, A. (2008). Linguistically Annotated Corpora: Quality Assurance, Reusability and Sustainability. In Lüdeling, A. and Kytö, M., editors, *Corpus Linguistics, HSK*. de Gruyter, Berlin etc. In print.

ALLC SESSION: e-Science: New collaborations between information technology and the humanities

Speakers

David Robey
d.j.b.robey@reading.ac.uk
Arts and Humanities Research Council, UK

Stuart Dunn
stuart.dunn@kcl.ac.uk
King's College London, UK

Laszlo Hunyadi
hunyadi@ling.arts.unideb.hu
University of Debrecen, Hungary

Dino Buzzetti
buzzetti@philo.unibo.it
University of Bologna, Italy

e-Science in the UK and elsewhere stands for a broad agenda as important for the humanities as it is for the natural sciences: to extend the application of advanced information technologies to develop new kinds of research across the whole range of academic disciplines, particularly through the use of internet-based resources. The aim of this session is to introduce some recent UK developments in this area, side by side with related developments in other parts of Europe.

• David Robey: Introduction

• Stuart Dunn: e-Science developments in the UK: temporal mapping and location-aware web technologies for the humanities

• Laszlo Hunyadi: Virtual Research Organizations for the Humanities in Europe: technological challenges, needs and opportunities

• Dino Buzzetti: Interfacing Biological and Textual studies in an e-Science Environment

Understanding TEI(s): A Presentation and Discussion Session

Chairs

Susan Schreibman
sschreib@umd.edu
University of Maryland, USA

Ray Siemens
siemens@uvic.ca
University of Victoria, Canada

Speakers will include

Peter Boot
peter.boot@huygensinstituut.knaw.nl
Huygens Institute The Netherlands

Arianna Ciula
arianna.ciula@kcl.ac.uk
King's College London, UK

James Cummings
James.Cummings@oucs.ox.ac.uk
University of Oxford, UK

Kurt Gärtner
gaertnek@staff.uni-marburg.de
Universität Trier, Germany

Martin Holmes
mholmes@uvic.ca
University of Victoria, Canada

John Walsh
jawalsh@indiana.edu
Indiana University, USA

The Text Encoding Initiative (TEI) is a consortium which collectively develops and maintains a standard for the representation of texts in digital form. Its chief deliverable is a set of Guidelines which specify encoding methods for machine-readable texts, primarily in the humanities, social sciences and linguistics. Since 1994, the TEI Guidelines have been widely used by libraries, museums, publishers, and individual scholars to present texts for online research, teaching, and preservation. In addition to the Guidelines themselves, the Consortium provides a variety of supporting resources, including resources for learning TEI <http://www.tei-c.org/Support/Learn/>, information on projects using the TEI <http://www.tei-c.org/Activities/Projects/>, TEI-related publications <http://www.tei-c.org/Support/Learn/>, and software <http://www.tei-c.org/Tools/> developed for or adapted to the TEI. The TEI Consortium is a non-profit membership organization composed of academic institutions, research projects, and individual scholars from around the world. For more information, see http://www.tei-c.org/.

This panel provides an opportunity beyond the project- and researcher-centred papers of the DH2008 conference for members of the TEI community to present on and discuss aspects of the TEI community itself, among them the nature and influence of Special Interest Groups (SIGs), recent trends in the integration of TEI with other schema and systems, issues of the annual conference and community support, and education and outreach now, and in the future.

DH2008: ADHO Session 'Digital resources in humanities research: Evidence of value (2)'

Chair

Harold Short
harold.short@kcl.ac.uk
King's College London, UK

Panelists

David Hoover
david.hoover@nyu.edu
New York University, USA

Lorna Hughes
lorna.hughes@kcl.ac.uk
King's College London, UK

David Robey
d.j.b.robey@reading.ac.uk
Arts and Humanities Research Council, UK

John Unsworth
unsworth@uiuc.edu
University of Illinois, Urbana-Champaign, USA

This takes further the issues discussed at the ALLC session at DH2007.

While most of us who do humanities computing need no convincing of its value, academic colleagues - including those on appointment and promotion panels - still need to be convinced, and even more so funders. If we want backing for the use and further development of digital resources, both data and processes, we need to collect more extensive concrete evidence of the ways in which they enable us to do research better, or to do research we would not otherwise be able to do, or to generate new knowledge in entirely new ways. Since the value of humanities research as a whole is qualitative, not quantitative, it is qualitative evidence in particular we should be looking for: digital resources providing the means not simply of doing research, but of doing excellent research.

The DH2007 panel session discussed a wide range of general issues arising in the discussion of the value of humanities computing, both in terms of its impact and results, and in terms of the intrinsic structures and qualities of digital objects created for research purposes.

The present panel session takes this further one the one hand by presenting recent work in the UK that has systematically tried to capture the value of humanities computing support activities, of digital tools development projects, and in general of the impact of of digital methods on research in the arts and humanities. Lorna Hughes and David Robey will discuss the results of the evaluation process at the end of the AHRC ICT Methods Network at King's College London, and of related work on a set of resource-development projects funded by the AHRC ICT in Arts and Humanities Research Programme.

John Unsworth and David Hoover will take a somewhat more anecdotal approach, and one that emphasizes North America rather than the UK. They will focus on a variety of ways of assessing and enhancing the value of digital humanities research in the areas of access, analysis, and advancing one's career. What kinds and levels of access to digital material have the most impact? What kinds of analysis and presentation, and what venues of publication or dissemination are most persuasive and effective? How does or can the exploitation of digital materials enhance the career of the (digital) humanist?

We hope that participants from other countries will contribute their points of view in the discussion.

The Building Blocks of the New Electronic Book

Ray Siemens
siemens@uvic.ca
University of Victoria, Canada

Claire Warwick
c.warwick@ucl.ac.uk
University College London, UK

Kirsten C. Uszkalo
kirsten@uszkalo.com
St. Francis Xavier University, Canada

Stan Ruecker
sruecker@ualberta.ca
University of Alberta, Canada

Paper 1: A New Context for the Electronic Book

Ray Siemens

The power of the book has preoccupied readers since before the Renaissance, a crucial time during which the human record entered print and print, in turn, facilitated the radical, crucial social and societal changes that have shaped our world as we know it. It is with as much potential and as much complexity, that the human record now enters the digital world of electronic media. At *this* crucial time, however, we must admit to some awkward truths. The very best of our electronic books are still pale reflections of their print models, and the majority offer much less than a reflection might. Even the electronic document – the webpage mainstay populating the World Wide Web – does not yet afford the same basic functionality, versatility, and utility as does the printed page. Nevertheless, more than half of those living in developed countries make use of the computer and the internet to read newspaper pieces, magazine and journal articles, electronic copies of books, and other similar materials on a daily basis; the next generation of adults already recognises the electronic medium as their chief source of textual information; our knowledge repositories increasingly favour digital products over the print resources that have been their mainstay for centuries; and professionals who produce and convey textual information have as a chief priority activities associated with making such information available, electronically – even if they must do so in ways that do not yet meet the standards of functionality, interactivity, and fostering of intellectual culture that have evolved over 500 years of print publication.

Why, then, are our electronic documents and books not living up to the promise realized in their print counterparts? What is it about the book that has made it so successful? How can we understand that success? How can we duplicate that success, in order to capitalize on economic and other advantages in the electronic realm?

This introductory paper frames several responses to the above, centring on issues related to textual studies, reader studies, interface design, and information management, and framing the more detailed responses in the areas below.

Paper 2: 'Humanities scholars, research and reading, in physical and digital environments'

Claire Warwick

It may initially seem strange to suggest that research on what seems to be such a traditional activity as reading might be a vital underpinning to the study of digital resources for the humanities. Yet, to design digital resources fit for research purposes, we must understand what scholars do in digital environments, what kind of resources they need, and what makes some scholars, particularly in the humanities, decline to use existing digital resources. One of the most important things that humanities scholars undertake in any environment is to read. Reading is the fundamental activity of humanities researchers, and yet we have little knowledge of how reading processes are modified or extended in new media environments. In the early to mid 1990s, many humanities scholars expressed excitement about the possibilities of electronic text, predicting that the experience of reading would change fundamentally (e.g., Bolter, 1991; Landow, 1992; Nunberg, 1996; Sutherland, 1997). Such prophetic writings, however, were not based on and rarely followed up by studies with readers, particularly readers in humanities settings. Through the last fifteen years critical interest within humanities circles with respect to reading has waned and little progress has been made in understanding how electronic textuality may affect reading practices, both of academic and non-academic readers.

We are, however, beginning to understand the relationship between reading in print and online, and how connections may be made between the two (Blandford, Rimmer, & Warwick 2006), and how humanities scholars relate to information environments, by reading and information seeking. In this paper we will discuss our research to date, undertaken as part of the UCIS project on the users of digital libraries, and on the LAIRAH project, which studied levels of use of a variety of digital humanities resources. (http://www.ucl.ac.uk/slais/research/circah/)

Research has been based on a use-in- context approach, in which participants have been interviewed about their preferences for the use of digital resources in a setting which is as naturalistic as possible. We also report on the results of user workshops, and of one to one observations of the use of digital resources.

As a result of these studies, we have found that humanities researchers have advanced information skills and mental models of their physical information environment, although they may differ in that they find these skills and models difficult

to apply to the digital domain (Makri et al. 2007). Humanities researchers are aware of the available functions as well as the problems of digital environments, and are concerned with accuracy, selection methods, and ease of use (Warwick et al. 2007). They require information about the original item when materials are digitized and expect high-quality content: anything that makes a resource difficult to understand – a confusing name, a challenging interface, or data that must be downloaded – will deter them from using it (Warwick et al. 2006).

It is therefore vital that such insights into the information behaviour of humanities scholars should be used to inform future studies of reading. It is also vital that this research is informed by the insights from other members of the wider research group, who work on textual studies and interface and resource design. The paper will therefore end with a short discussion of future plans for such collaboration.

References

Bolter, J. David. (1991). *Writing space: the computer, hypertext, and the history of writing*. Hillsdale, NJ: L. Erlbaum Associates.

Blandford, A., Rimmer, J., & Warwick, C. (2006). Experiences of the library in the digital age. Paper presented at 3rd International Conference on Cultural Covergence and Digital Technology, Athens, Greece.

Landow, G. P. (1992). *Hypertext: the convergence of contemporary critical theory and technology*. Baltimore, MD: Johns Hopkins University Press, 1992.

Makri, S., Blandford, A., Buchanan, G., Gow, J., Rimmer, J., & Warwick, C. (2007). A library or just another information resource? A case study of users' mental models of traditional and digital libraries. *Journal of the American Society of Information Science and Technology,* 58(3), 433-445.

Nunberg, G (ed.) (1996) *The Future of the Book*. Berkeley, University of California Press.

Sutherland, K. (Ed) (1997) *Electronic Text: Investigations in Method and Theory.* Oxford, Clarendon Press

Warwick, C., Terras, M., Huntington, P., & Pappa, N. (2007) "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*, forthcoming

Paper 3: A Book is not a Display: A Theoretical Evolution of the E-Book Reader

Kirsten C. Uszkalo and Stan Ruecker

In our part of this panel session, we will discuss the design of the physical electronic book—the reading device—from a user perspective. There has been a tendency on the part of producers of information to disassociate the information from its media. This makes sense in a digital age: there are inefficiencies in production that can be made more efficient by remediation. Additionally, there is little intrinsic justification for preferring one medium over another, unless one wishes to take into account the situated context of use. Designers, however, insofar as they are distinct from producers of information, have had a longstanding tendency to become absorbed in the situated context of use.

When the conventions of HTML were determined in such a way that they discounted formatting, and for very good reasons of cross-platform compatibility, designers began the uphill battle of changing the conventions, because they felt that formatting, and not just content, was important. For example, Travis Albers says of his recent project Bookglutton.com: "BookGlutton is working to facilitate adoption of on-line reading. Book design is an important aspect of the reader, and it incorporates design elements, like dynamic dropcaps."

The issue at stake with existing electronic book readers is their function as an interface, which can, according to Gui Bonsiepe, create the "possibilities for effective action or obstruct them" (62). Interface issues like eye strain, contrast, and resolution are being addressed through the use of E-Ink's electronic paper in Sony's 2007 / 2007 Sony E Book, Bookeen, iRex Technologies, and Amazon's Kindle (2008) electronic book readers. Samsung and LG Philips are addressing object size of ebook readers with 14 inch lightweight epaper formats; and Polymer Vision's Readius, "a mobile device" is a "flexible 5-inch e-paper display that unfurls like a scroll." However, these formats lack the intuitive, familiar feeling of the book. Acknowledging that ebook proprietary software issues, copyright, and cost might keep people from investing in a reading device in order to be able to read a book, and noting the continued superiority of the paper book in both form and function over the electronic book reader, we will look to the range of design issues which continue to plague the potential viable sales for the electronic book reader.

References

Balas, Janet L. "Of iPhones and Ebooks: Will They Get Together?" *Computers in Libraries*. Westport: Nov/Dec 2007. Vol. 27, Iss. 10; pg. 35, 4 pgs

Bina. "DON'T Upgrade the REB 1100!!," Published November 6, 2004. Accessed November 17, 2007. http://www.amazon.com/RCA-REB1100-eBook-Reader/dp/B00005T3UH

Geist, Michael. "Music Publisher's Takedown Strikes The Wrong Chord" Published Tuesday October 30, 2007. Accessed November 17, 2007 http://www.michaelgeist.ca/content/view/2336/135/

Teaching and learning: Dual educational electronic textbooks: the starlight platform

Grammenos, Dimitris. et al "Teaching and learning: Dual educational electronic textbooks: the starlight platform" Proceedings of the 9th international ACM SIGACCESS conference on Computers and accessibility Assets '07 October 2007

Hodge, Steve, et al. "ThinSight: versatile multi-touch sensing for thin form-factor displays" Symposium on User Interface Software and Technology archive. *Proceedings of the 20th annual ACM*, Newport, Rhode Island, USA, 2007, pp. 259 - 268

Farber, Dan. "Sony's misguided e-book" Published February 22nd, 2006. Accessed November 17, 2007. http://blogs.zdnet.com/BTL/?p=2619

Forest, Brady. "Illiad, a new ebook reader" Published September 21, 2006. Accessed November 17, 2007 http://radar.oreilly.com/archives/2006/09/illiad_a_new_ebook_reader.html

Mangalindan, Mylene. "Amazon to Debut E-Book Reader" *Wall Street Journal*. Published November 16, 2007; Page B2.

Morrison, Chris. "10 Taking a Page Out of the E-Book." *Business 2.0*. San Francisco: Oct 2007. Vol. 8, Iss. 9; pg. 34

Oswald, Ed "Sony Updates Electronic Book Reader" Published October 2, 2007, 2:50. Accessed November 17, 2007. <http://www.betanews.com/article/Sony_Updates_Electronic_Book_Reader/1191350925>

Stone, Brad. "Envisioning the Next Chapter for Electronic Books" Published: September 6, 2007. Accessed November 17, 2007. http://www.nytimes.com/2007/09/06/technology/06amazon.html

SDH/SEMI panel: Text Analysis Developers' Alliance (TADA) and T-REX

Organiser:

Stéfan Sinclair
sgsinclair@gmail.com
McMaster University, Canada

Chair:

Ray Siemens
siemens@uvic.ca
University of Victoria, Canada

Participants:

Stéfan Sinclair
sgsinclair@gmail.com
McMaster University, Canada

Matt Jockers
mjockers@stanford.edu
Stanford University, USA

Susan Schreibman
sschreib@umd.edu
University of Maryland, USA

Patrick Juola
juola@mathcs.duq.edu
Duquesne University, USA

David Hoover
david.hoover@nyu.edu
New York University, USA

Jean-Guy Meunier
meunier.jean-guy@uqam.ca
Université du Québec à Montréal, Canada

Dominic Forest
dominic.forest@umontreal.ca
University of Montreal, Canada

The Text Analysis Developers' Alliance (TADA) is an informal affiliation of designers, developers, and users of text analysis tools in the digital humanities. The TADA wiki (http://tada.mcmaster.ca/) has steadily grown since the group's first meeting three years ago, and now contains a wide range of useful documents, including general information about text analysis, pedagogically-oriented text analysis recipes, and content from several projects engaged in open research (publicly documenting themselves). Most recently TADA has

announced its inaugural T-REX event (http://tada.mcmaster.ca/trex/), a community initiative with the following objectives:

- to identify topics of shared interest for further development in text analysis
- to evaluate as a community ideas and tools that emerge from the topics
- to present results from this process as research in various venues

This panel will provide a more detailed account of the history and activities of TADA, as well as a preliminary account of T-REX from the participants themselves, where possible.

Agora.Techno.Phobia.Philia2: Feminist Critical Inquiry, Knowledge Building, Digital Humanities

Martha Nell Smith
mnsmith@umd.edu
University of Maryland, USA

Susan Brown
sbrown@uoguelph.ca
University of Guelph, Canada

Laura Mandell
mandellc@muohio.edu
Miami University, USA

Katie King
katking@umd.edu
University of Maryland, USA

Marilee Lindemann
mlindema@umd.edu
University of Maryland, USA

Rebecca Krefting
beckortee@starpower.net
University of Maryland, USA

Amelia S. Wong
awong22@umd.edu
University of Maryland, USA

In the later twentieth century, the humanities were transformed by the rise of gender studies and related critical movements that investigate the impact of socialized power distribution on knowledge formations and institutions. Throughout the humanities, the very research questions posed have been fundamentally altered by the rise of feminist and, more recently, critical race, queer, and related social justice-oriented critiques. These critical perspectives, as speakers in last year's DH2007 panel "Agora.Techno.Phobia.Philia: Gender, Knowledge Building, and Digital Media" amply demonstrated, have much to contribute not only to thinking about women in relation to information and communication technologies but also to advancing the work and efficacies of information and communications technologies. The session we propose extends and deepens the critical inquiries posed in last year's session. Though Carolyn Guertin will not be able to join us in Finland this year, we each will take into account the telling survey (available upon request) she presented of virtual space and concomitant debates devoted to women and information technology, as well as her compelling analysis of the stakes involved for female participants in such debates. Important also for our considerations are the special issue of Frontiers devoted to gender, race, and information technology (http://

muse.jhu.edu/journals/frontiers/toc/fro26.1.html); recent books by feminist thinkers such as Isabel Zorn, J. McGrath Cohoon and William Aspray, Lucy Suchman; and the most recent issue of Vectors devoted to Difference (Fall 2007; http://www.vectorsjournal.org/index.php?page=6%7C1). Though these publications demonstrate keen and widespread interest in messy and ambiguous questions of diversity and technology and new media, such work and debates seem oddly absent from or at least underappreciated by the digital humanities (DH) community itself.

It's not that the Digital Humanities community feels hostile to women or to feminist work per se. Women have historically been and currently are visible, audible, and active in ADHO organizations, though not in equal numbers as men. The language used in papers is generally gender-aware. Conference programs regularly include work about feminist projects, that is to say ones that focus on feminist content (such as the interdisciplinary The Orlando Project, The Poetess Archive, Women Writers Project, Dickinson Electronic Archives). Though other DH projects and initiatives are inter- or trans-disciplinary, their interdisciplinarities have tended not to include feminist elements, presumably because those feminist lines of inquiry are assumed to be discipline- or content-based (applicable only to projects such as those mentioned above), hence are largely irrelevant to what the digital humanities in general are about. Thus DH conference exchanges tend to have little to do with gender. In other words, DH debates over methodological and theoretical approaches, or over the substance of digital humanities as a field, have not been informed by the research questions fundamentally altered by feminist, critical race, queer, and related social justice-oriented critiques, the very critiques that have transformed the humanities as a whole.

Why? The timing of the rise of the "field" perhaps?-much of the institutionalization and expansion, if not the foundation, of digital humanities took place from the late 20th century onwards. As recent presentations at major DH centers make clear (Smith @ MITH in October 2007, for example), there is a sense among some in the DH community that technologies are somehow neutral, notwithstanding arguments by Andrew Feenberg, Lucy Suchman, and others in our bibliography that technology and its innovations inevitably incorporate patterns of domination and inequity or more generalized social power distributions. Yet if we have learned one thing from critical theory, it is that interrogating the naturalized assumptions that inform how we make sense of the world is crucial in order to produce knowledge. Important for our panel discussion (and for the companion poster session we hope will be accepted as part of this proposal) is feminist materialist inquiry, which urges us to look for symptomatic gaps and silences in intellectual inquiry and exchange and then analyze their meanings in order to reinflect and thereby enrich the consequent knowledge-building discourses. So what happens if we try to probe this silence on gender in digital humanities discourse, not only in regard to institutions and process, but in regard to defining and framing DH debates?

Building on the work of Lucy Suchman and the questions of her collaborators on the panel, Martha Nell Smith will show how questions basic to feminist critical inquiry can advance our digital work (not just that of projects featuring feminist content): How do items of knowledge, organizations, working groups come into being? Who made them? For what purposes? Whose work is visible, what is happening when only certain actors and associated achievements come into public view? What happens when social orders, including intellectual and social framings, are assumed to be objective features of social life (i.e., what happens when assumptions of objectivity are uninformed by ethnomethodology, which reminds us that social order is illusory and that social structures appear to be orderly but are in reality potentially chaotic)? Doing so, Smith will analyze the consequences of the politics of ordering within disambiguating binary divisions (Subject and object; Human and nonhuman; Nature and culture; Old and new) and the consequent constructions and politics of technology: Where is agency located? Whose agencies matter? What counts as innovation: why are tools valorized? How might the mundane forms of inventive, yet taken for granted labor, necessary (indeed VITAL) to the success of complex sociotechnical arrangements, be more effectively recognized and valued for the crucial work they do? Smith will argue that digital humanities innovation needs to be sociological as well as technological if digital humanities is to have a transformative impact on traditional scholarly practices.

Susan Brown will open up the question-what happens if we try to probe this silence on gender in digital humanties discourse-in a preliminary way by focusing on a couple of terms that circulate within digital humanities discourse in relation to what the digital humanities are (not) about: delivery and service. Delivery and interface work are intriguingly marginal rather than central to DH concerns, despite what we know about the impact of usability on system success or failure. Willard McCarty regards the term "delivery" as metaphorically freighted with connotations of knowledge commodification and mug-and-jug pedagogy (6). However, informed by feminist theory and the history of human birthing, we might alternatively mobilize the less stable connotations of delivery as "being delivered of, or act of bringing forth, offspring," which offers a model open to a range of agents and participants, in which the process and mode of delivery have profound impacts on what is delivered. The history of the forceps suggests a provocative lens on the function of tools in processes of professionalization and disciplinary formation. An emphasis on delivery from conception through to usability studies of fully developed interfaces will-if it takes seriously the culturally and historically grounded analyses of technology design and mobilization modeled by Lucy Suchman-go a considerable distance towards breaking down the oppositions between designer and user, producer and consumer, technologist and humanist that prevent digital humanities from having transformative impacts on traditional scholarly practices. This sense of delivery foregrounds boundaries as unstable and permeable, and boundary issues, as Donna Haraway was among the first to argue, have everything to do with the highly politicized-and gendered-category of the

human, the subject/object of humanist knowledge production. The relationships among service (often married to the term "support"), disciplinarity, and professionalization within debates over the nature and identity of humanities computing are similarly important for reimagining how DH work might be advanced and might have a greater impact on the humanities as a whole. Service, a feminized concept, recurs as a danger to the establishment of legitimacy (Rommel; Unsworth). Sliding, as Geoffrey Rockwell points out, easily into the unequivocally denigrating "servile," service too challenges us to be aware of how institutional power still circulates according to categories and hierarchies that embed social power relations. As Susan Leigh Star has argued, marginality, liminality, or hybridity, multiple membership that crosses identities or communities, provide valuable vantage points for engagement with shifting technologies (Star 50-53). These two liminal terms within digital humanities debates may open up digital humanities debates to new questions, new prospects, and Brown will offer examples of some ways in which this might be done in order to enrich DH work in general.

By focusing on the terms gynesis and ontology, Laura Mandell examines the relationship between concepts in traditional epistemological discourses and those concepts as they appear in discussions about and implementations of the semantic web: hierarchy, ontology, individual, and class. First, she will look at traditional meanings of these terms, performing philological analysis of their origin and province as well as locating them in their disciplinary silos - in the manner of Wikipedia's disambiguation pages. Next, she will look at how the terms are defined in discussions of the emerging semantic web, and also at how they function practically in XML Schema, database design, and OWL, the Web Ontology Language. Much work has been done by feminist theorists and philosophers from Luce Irigaray to more recently Liz Stanley and Sue Wise in critiquing the sexism implicit in concepts of hierarchy, being, and individuality; Mary Poovey has tracked the emergence of modern notions of class in her book about the history of the fact. Mandell will extend these foundational critiques, asking whether any of the insights achieved by that work apply to the semantic web. Of particular concern for Mandell is the category of error. Simone de Beauvoir's Second Sex first brought to light the notion of woman as Other, as "error" to the man, and Julia Kristeva (following anthropologist Mary Douglas) finds female agency lurking in the dirt or ambiguity of any logical system. If computational and generational languages such as those producing the semantic web cannot tolerate ambiguity and error, what will be repressed? Does the syntax of natural language allow for women's subjectivity in a way that ontological relationships might not? What are the opportunities for adding relationships to these systems that capitalize upon the necessary logical errors that philosopher Luce Irigaray has designated as necessary if "The Human are two"?

Rounding out this panel will be two respondents to these three 12-15 minute presentations by experts in feminist and queer theory and practice, both of whom have expressed keen interest in DH through the evolving practices of their work in knowledge-building and pedagogy. Professors Katie King, who has worked for the past two decades on writing technologies and their implications for knowledge production, and Marilee Lindemann, who is a leading queer studies theorist and a blogger who has been exploring new forms of scholarly writing in her creative nonfiction, Roxie's World (http://roxies-world.blogspot.com/) will respond to the questions posed by Smith, Brown, and Mandell, formulating questions that aim to engage the audience in discourse designed to advance both our practices and our theories about what we are doing, where we might be going, and how augmenting our modes and designs for inquiry might improve both our day-to-day DH work and its influence on humanities knowledge production. Their questions will be posed ahead of time (by the end of May) on the website we will devote to this session (http://www.mith.umd.edu/mnsmith/DH2008), as will "think piece" interactive essays contributed by all of the participants (we are mostly likely to use a wiki).

As a result of ongoing work with feminist graduate students interested in new media, the possibilities realized and unrealized, we propose extending this panel asynchronously so that current dissertation research be featured as case studies more than suitable for testing some of the hypotheses, exploring the questions, and reflecting upon the issues raised by the panel presentations. Thus we propose two poster sessions (or one with two components) by University of Maryland graduate students: Rebecca Krefting, "Because HTML Text is Still a Text: The (False) Promise of Digital Objectivity," and Amelia Wong, "Everyone a Curator?: The Potential and Limitations of Web 2.0 for New Museology."

Papers

Unicode 5.0 and 5.1 and Digital Humanities Projects

Deborah Winthrop Anderson
dwanders@sonic.net
UC Berkeley, USA

In theory, digital humanities projects should rely on standards for text and character encoding. For character encoding, the standard recommended by TEI P5 (TEI Consortium, eds. Guidelines for Electronic Text Encoding and Interchange [last modified: 03 Feb 2008], http://www.tei-c.org/P5/) is the Unicode Standard (http://www.unicode.org/standard/standard.html). The choices made by digital projects in character encoding can be critical, as they impact text analysis and language processing, as well as the creation, storage, and retrieval of such textual digital resources. This talk will discuss new characters and important features of Unicode 5.0 and 5.1 that could impact digital humanities projects, discuss the process of proposing characters into Unicode, and provide the theoretical underpinnings for acceptance of new characters by the standards committees. It will also give specific case studies from recent Unicode proposals in which certain characters were not accepted, relaying the discussion in the standards committees on why they were not approved. This latter topic is important, because decisions made by the standards committees ultimately will affect text encoding.

For those characters not in Unicode, the P5 version of the TEI Guidelines deftly describes what digital projects should do in Chapter 5 (TEI Consortium, eds. "Representation of Non-standard Characters and Glyphs," Guidelines for Electronic Text Encoding and Interchange [last modified: 03 Feb 2008], http://www.tei-c.org/release/doc/tei-p5-doc/en/html/WD.html [accessed: 24 March 2008]), but one needs to be aware of the new characters that are in the standards approval process. The presentation will briefly discuss where to go to look for the new characters on public websites, which are "in the pipeline."

The release of Unicode 5.0 in July 2007 has meant that an additional 1,369 new characters have been added to the standard, and Unicode 5.1, due to be released in April 2008, will add 1,624 more (http://www.unicode.org/versions/Unicode5.1.0/) In order to create projects that take advantage of what Unicode and Unicode-compliant software offers, one must be kept abreast of developments in this standard and make appropriate changes to fonts and documents as needed. For projects involving medieval and historic texts, for example, the release of 5.1 will include a significant number of European medieval letters, as well as new Greek and Latin epigraphic letters, editorial brackets and half-brackets, Coptic combining marks, Roman weights and measures and coin symbols, Old Cyrillic letters and Old Slavonic combining letters. The Menota project (http://www.menota.org/guidelines-2/convertors/convert_2-0-b.page), EMELD's "School of Best Practice" (http://linguistlist.org/emeld/school/classroom/conversion/index.html), and SIL's tools (http://scripts.sil.org/Conversion) all provide samples of conversion methods for upgrading digital projects to include new Unicode characters.

Since Unicode is the accepted standard for character encoding, any critical assessment of Unicode made to the body in charge of Unicode, the Unicode Technical Committee, is generally limited to comments on whether a given character is missing in Unicode or--if proposed or currently included in Unicode--critiques of a character's glyph and name, as well as its line-breaking properties and sorting position. In Chapter 5 of the TEI P5 Guidelines, mention is made of character properties, but it does not discuss line-breaking or sorting, which are now two components of Unicode proposals and are discussed in annexes and standards on the Unicode Consortium website (Unicode Standard Annex #14 "Line Breaking Properties," Unicode Technical Standard #10, "Unicode Collation Algorithm," both accessible from www.unicode.org). Users should pay close attention to these two features, for an incorrect assignment can account for peculiar layout and sorting features in software. Comments on missing characters, incorrect glyphs or names, and properties should all be directed to the Unicode online contact page (http://www.unicode.org/reporting.html). It is recommended that an addition to Chapter 5 of P5 be made regarding word-breaking and collation when defining new characters.

The Unicode Standard will, with Unicode 5.1, have over 100,000 characters encoded, and proposals are underway for several unencoded historic and modern minority scripts, many through the Script Encoding Initiative at UC Berkeley (http://www.linguistics.berkeley.edu/sei/alpha-script-list.html). Reviewing the glyphs, names, and character properties for this large number of characters is difficult. Assistance from the academic world is sought for (a) authoring and review of current proposals of unencoded character and scripts, and (b) proofing the beta versions of Unicode. With the participation of digital humanists, this character encoding standard can be made a reliable and useful standard for such projects.

Variation of Style: Diachronic Aspect

Vadim Andreev
smol.an@mail.ru
Smolensk State University, Russian Federation

Introduction

Among works, devoted to the quantitative study of style, an approach prevails which can be conventionally called as *synchronic*. Synchronic approach is aimed at solving various classification problems (including those of attribution), making use of average (mean) values of characteristics, which reflect the style of the whole creative activity of an author. This approach is based on the assumption that the features of an individual style are not changing during lifetime or vary in time very little, due to which the changes can be disregarded as linguistically irrelevant.

This assumption can be tested in experiments, organised within a *diachronic* approach, whose purpose is to compare linguistic properties of texts, written by the same author at different periods of his life.

This paper presents the results of such a diachronic study of the individual style of famous American romantic poet E.A.Poe. The study was aimed at finding out whether there were linguistically relevant differences in the style of the poet at various periods of his creative activity and if so, at revealing linguistic markers for the transition from one period to the other.

Material

The material includes iambic lyrics published by Poe in his 4 collections of poems. Lyrics were chosen because this genre expresses in the most vivid way the essential characteristics of a poet. In order to achieve a common basis for the comparison only iambic texts were taken, they usually did not exceed 60 lines. It should be noted that iamb was used by Poe in most of his verses. Sonnets were not taken for analysis because they possess specific structural organization. Poe's life is divided into three periods: (1) from Poe's first attempts to write verses approximately in 1824 till 1829, (2) from 1830 till 1835 and (3) from 1836 till 1849.

Characteristics

For the analysis 27 characteristics were taken. They include morphological and syntactic parameters.

Morphological characteristics are formulated in terms of traditional morphological classes (noun, verb, adjective, adverb and pronoun). We counted how many times each of them occurs in the first and the final strong (predominantly stressed) syllabic positions – ictuses.

Most of syntactic characteristics are based on the use of traditional notions of the members of the sentence (subject, predicate, object, adverbial modifier) in the first and the final strong positions in poems. Other syntactic parameters are the number of clauses in (a) complex and (b) compound sentences.

There are also several characteristics which represent what can be called as poetical syntax. They are the number of enjambements, the number of lines divided by syntactic pauses and the number of lines, ending in exclamation or question marks. Enjambement takes place when a clause is continued on the next line (And what is not a dream by day / To him whose eyes are cast / On things around him <…>). Pause is a break in a line, caused by a subordinate clause or another sentence (I feel ye now – I feel ye in your strength – <…>).

The values of the characteristics, which were obtained as a result of the analysis of lyrics, were normalised over the size of these texts in lines.

Method

One of multivariate methods of statistical analyses – discriminant analysis – was used. This method has been successfully used in the study of literary texts for authorship detection (Stamatatos, Fakatakis and Kokkinakis 2001; Baayen, Van Halteren, and Tweedie 1996, etc.), genre differentiation (Karlgen, Cutting 1994; Minori Murata 2000, etc.), gender categorization (Koppel et al. 2002; Olsen 2005), etc.

Discriminant analysis is a procedure whose purpose is to find characteristics, discriminating between naturally occurring (or a priori formed) classes, and to classify into these classes separate (unique) cases which are often doubtful and "borderline". For this purpose linear functions are calculated in such a way as to provide the best differentiation between the classes. The variables of these functions are characteristics of objects, relevant for discrimination. Judging by the coefficients of these variables we can single out the parameters which possess maximum discriminating force. Besides, the procedure enables us to test the statistical significance of the obtained results (Klecka, 1989).

In this paper discriminant analysis is used to find out if there is any difference between groups of texts written during Periods 1–3, reveal characteristics differentiating these text groups and establish their discriminating force.

Results

It would be natural to expect that due to Poe's relatively short period of creative activity (his first collection of poems was published in 1827, his last collection – in 1845) his individual style does not vary much, if at all. Nevertheless the results show that there are clearly marked linguistic differences between his texts

written during these three periods. Out of 27 characteristics, used in the analysis, 14 proved to possess discriminating force, distinguishing between the verse texts of different periods of the author's life. The strongest discriminating force was observed in morphological characteristics of words both in the first and final strong positions and syntactic characteristics of the initial part of verse lines. These parameters may be used for automatic classification of Poe's lyrics into three groups corresponding to three periods of his creative activity with 100% correctness.

The transition from the first to the second period is mainly characterised by changes in the number of verbs, nouns and pronouns in the first and the last strong positions, as well as in the number of subordinate clauses in complex sentences, words in the function of adverbial modifier in the initial position in the line. The development in Poe's style from the second to the third period is also marked by changes in the number of morphological classes of words in the initial and final strong positions of the line (nouns, adverbs and pronouns).

It should be stressed that these changes reflect general tendencies of variation of frequencies of certain elements and are not present in all the texts. In the following examples the shift of verbs from the final part of the line, which is characteristic of the first period, to the initial strong position of the line (i.e. second syllable) in the second period is observed.

Period 1

But when within thy waves she looks –

Which glistens then, and trembles –

Why, then, the prettiest of brooks

Her worshipper resembles –

For in my heart – as in thy streem –

Her image deeply lies <...>

(To the River)

Period 2

You know the most enormous flower –

That rose – <...>

I tore it from its pride of place

And shook it into pieces <...>

(Fairy Land)

On the whole the results show that there are certain linguistic features which reflect the changes in the style of E.A.Poe. Among important period markers are part of speech characteristics and several syntactic parameters.

Bibliography

Baayen, R.H., Van Halteren, H., and Tweedie, F. (1996) Outside the cave of shadows: Using syntactic annotation to enhance authorship attribution. *Literary and Linguistic Computing*, 11: 121–131.

Karlgen, J., Cutting, D. (1994) Recognizing text genres with simple metrics using discriminant analysis. *Proceedings of COLING 94*, Kyoto: 1071–1075.

Klecka, W.R. (1989). *Faktornyj, diskriminantnyj i klasternyj analiz*. [Factor, discriminant and cluster analysis]. Moscow: Finansi i statistica.

Koppel, M, Argamon, S., and Shimoni, A.R. (2002) Automatically Categorizing Written Texts by Author Gender. *Literary & Linguistic Computing*, 17: 401–412.

Murata, M. (2000) Identify a Text's Genre by Multivariate Analysis – Using Selected Conjunctive Words and Particle-phrases. *Proceedings of the Institute of Statistical Mathematics*, 48: 311–326.

Olsen, M. (2005) Écriture Féminine: Searching for an Indefinable Practice? *Literary & Linguistic Computing*, 20: 147–164.

Stamatatos, E., Fakatakis, N., & Kokkinakis, G. (2001). Computer-Based Authorship Attribution Without Lexical Measures. *Computers and the Humanities*, 35: 193–214.

Exploring Historical Image Collections with Collaborative Faceted Classification

Georges Arnaout
garna001@odu.edu
Old Dominion University, USA

Kurt Maly
maly@cs.odu.edu
Old Dominion University, USA

Milena Mektesheva
mmekt001@odu.edu
Old Dominion University, USA

Harris Wu
hwu@odu.edu
Old Dominion University, USA

Mohammad Zubair
zubair@cs.odu.edu
Old Dominion University, USA,

The US Government Photos and Graphics Collection include some of the nation's most precious historical documents. However the current federation is not effective for exploration. We propose an architecture that enables users to collaboratively construct a faceted classification for this historical image collection, or any other large online multimedia collections. We have implemented a prototype for the American Political History multimedia collection from usa.gov, with a collaborative faceted classification interface. In addition, the proposed architecture includes automated document classification and facet schema enrichment techniques.

Introduction

It is difficult to explore a large historical multimedia humanities collection without a classification scheme. Legacy items often lack textual description or other forms of metadata, which makes search very difficult. One common approach is to have librarians classify the documents in the collection. This approach is often time or cost prohibitive, especially for large, growing collections. Furthermore, the librarian approach cannot reflect diverse and ever-changing needs and perspectives of users. As Sir Tim Berners-Lee commented: "the exciting thing [about Web] is serendipitous reuse of data: one person puts data up there for one thing, and another person uses it another way." Recent social tagging systems such as del.icio.us permit individuals to assign free-form keywords (tags) to any documents in a collection. In other words, users can contribute metadata. These tagging systems, however, suffer from low quality of tags and lack of navigable structures.

The system we are developing improves access to a large multimedia collection by supporting users collaboratively build a faceted classification. Such a collaborative approach supports diverse and evolving user needs and perspectives. Faceted classification has been shown to be effective for exploration and discovery in large collections [1]. Compared to search, it allows for recognition of category names instead of recalling of query keywords. Faceted classification consists of two components: the facet schema containing facets and categories, and the association between each document and the categories in the facet schema. Our system allows users to collaboratively 1) evolve a schema with facets and categories, and 2) to classify documents into this schema. Through users' manual efforts and aided by the system's automated efforts, a faceted classification evolves with the growing collection, the expanding user base, and the shifting user interests.

Our fundamental belief is that a large, diverse group of people (students, teachers, etc.) can do better than a small team of librarians in classifying and enriching a large multimedia collection.

Related Research

Our research builds upon popular wiki and social tagging systems. Below we discuss several research projects closest to ours in spirit.

The Flamenco project [1] has developed a good browsing interface based on faceted classification, and has gone through extensive evaluation with digital humanities collections such as the fine art images at the museums in San Francisco. Flamenco, however, is a "read-only" system. The facet schema is pre-defined, and the classification is pre-loaded. Users will not be able to change the way the documents are classified.

The Facetag project [2] guides users' tagging by presenting a predetermined facet schema to users. While users participate in classifying the documents, the predetermined facet schema forces users to classify the documents from the system's perspective. The rigid schema is insufficient in supporting diverse user perspectives.

A few recent projects [4, 7] attempt to create classification schemas from tags collected from social tagging systems. So far these projects have generated only single hierarchies, instead of multiple hierarchies as in faceted schemas. Also just as any other data mining systems, these automatic classification approaches suffers from quality problems.

So far, no one has combined user efforts and automated techniques to build a faceted classification, both to build the schema and to classify documents into it, in a collaborative and interactive manner.

Architecture and Prototype Implementation

The architecture of our system is shown in Figure 1. Users can not only tag (assign free-form keywords to) documents but also collaboratively build a faceted classification in a wiki fashion. Utilizing the metadata created by users' tagging efforts and harvested from other sources, the system help improve the classification. We focus on three novel features: 1) to allow users collaboratively build and maintain a faceted classification, 2) to systematically enrich the user-created facet schema, 3) to automatically classify documents into the evolving facet schema.

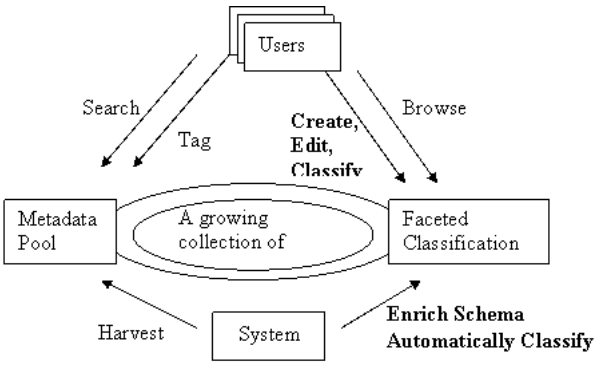


Figure 1. System Architecture

We have developed a Web-based interface that allows users create and edit facets/categories similar to managing directories in the Microsoft File Explorer. Simply by clicking and dragging documents into faceted categories, users can classify (or re-classify) historic documents. All the files and documents are stored in a MySQL database. For automatic classification, we use a support vector machine method [5] utilizing users' manual classification as training input. For systematic facet enrichment, we are exploring methods that create new faceted categories from free-form tags based on a statistical co-occurrence model [6] and also WordNet [8].

Note that the architecture has an open design so that it can be integrated with existing websites or content management systems. As such the system can be readily deployed to enrich existing digital humanity collections.

We have deployed a prototype on the American Political History (APH) sub-collection (http://teachpol.tcnj.edu/amer_pol_hist) of the US Government Photos and Graphics Collection, a federated collection with millions of images (http://www.usa.gov/Topics/Graphics.shtml). The APH collection currently contains over 500 images, many of which are among the nation's most valuable historical documents. On the usa.gov site, users can explore this collection only by two ways: either by era, such as 18th century and 19th century, or by special topics, such as "presidents" (Figure 2). There are only four special topics manually maintained by the collection administrator, which do not cover most items in the collection. This collection is poor with metadata and tools, which is common to many digital humanity collections that contain legacy items that have little pre-existing metadata, or lack resources for maintenance.

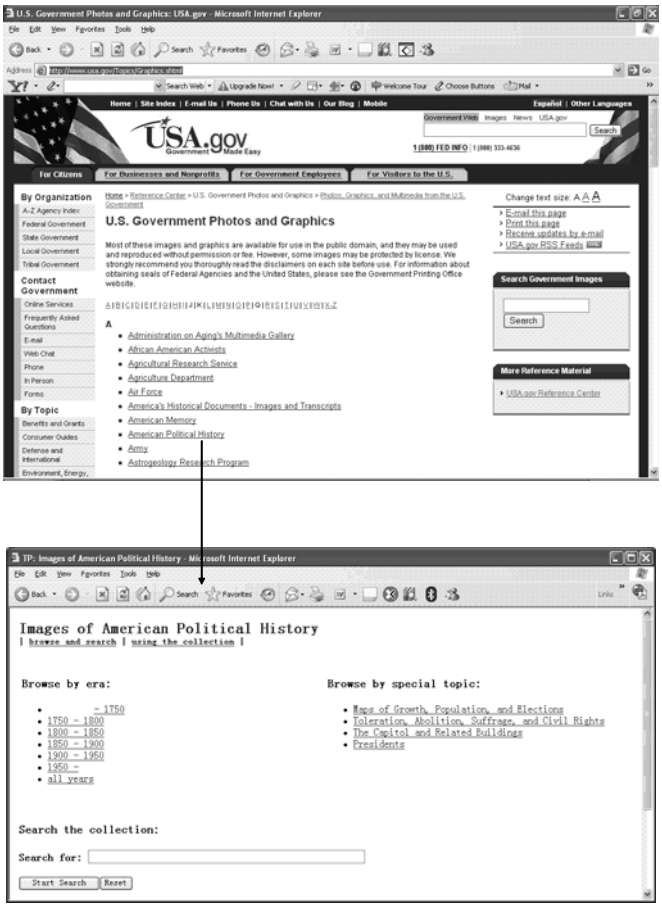
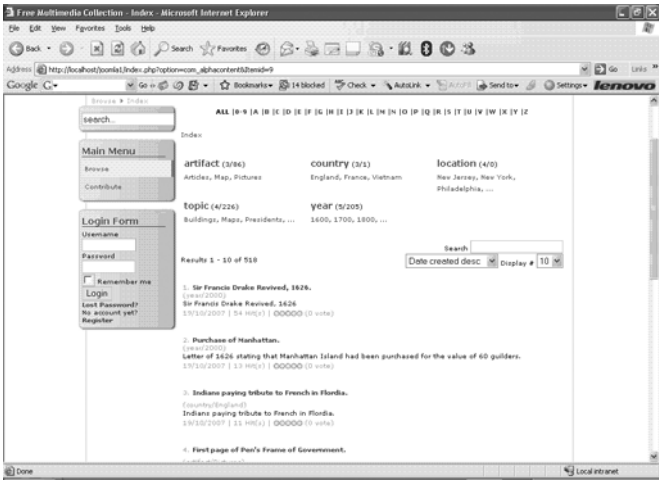


Figure 2. American Political History Collection at usa.gov

The prototype focused on the collaborative classification interface. After deploying our prototype, the collection has been collaboratively classified into categories along several facets. To prove the openness of system architecture, the prototype has been integrated with different existing systems. (Figure 3)

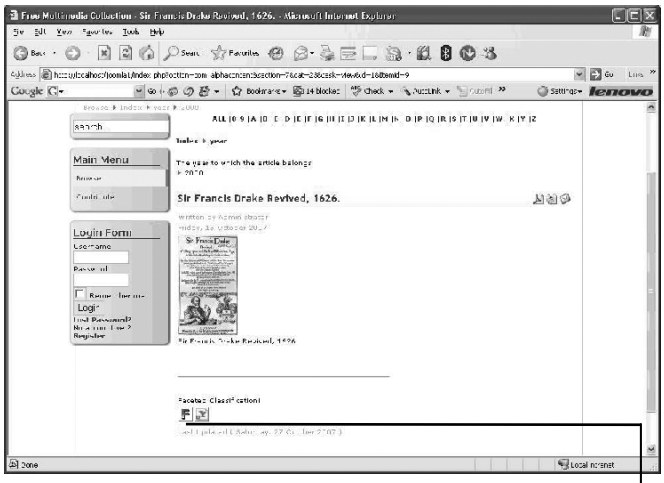
The system integrated with a Flamenco-like Interface



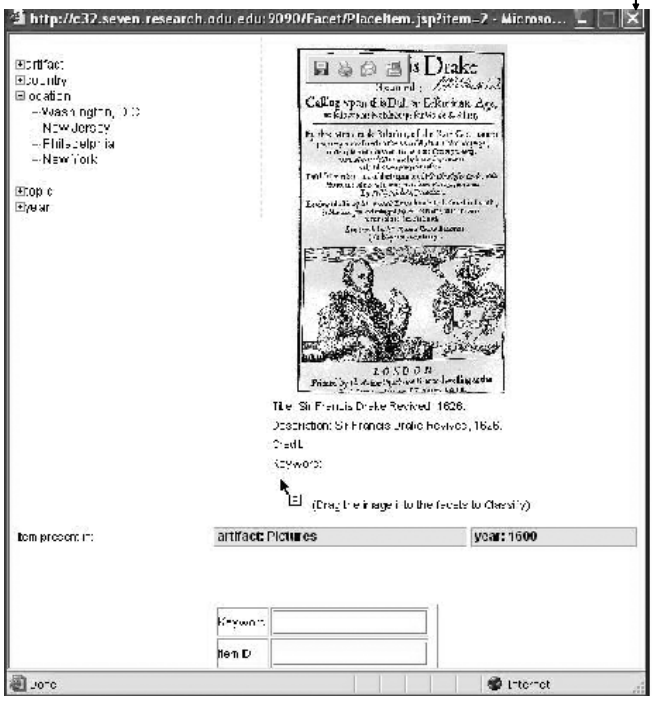
The system integrated with Joomla!, a popular content management system

Figure 3. Multi-facet Browsing

As users explore the system (such as by exploring faceted categories or through a keyword search), besides each item there is a "classify" button which leads to the classification interface. The classification interface shows the currently assigned categories in various facets for the selected item. It allows user to drag and drop an item into a new category. At this level user can also add or remove categories from a facet, or add or remove a facet.



Faceted Classification button on the bottom of the screen (the button to the right links to a social tagging system, del.icio.us)



The classification interface. Users can create/edit facets and categories, and drag items into categories

Figure 4. Classification Interface

Evaluation and Future Steps

Initial evaluation results in a controlled environment show great promise. The prototype was tested by university students interested in American political history. The collection was collaboratively categorized into facets such as Artifact (map, photo, etc.), Location, Year, and Topics (Buildings, Presidents, etc.) The prototype is found to be more effective than the original website in supporting user's retrieval tasks, in terms of both recall and precision. At this time, our prototype does not have all the necessary support to be deployed on public Internet for a large number of users. For this we need to work on the concept of hardening a newly added category or facet. The key idea behind hardening is to accept a new category or

facet only after reinforcement from multiple users. In absence of hardening support our system will be overwhelmed by the number of new facets and categories. We are also exploring automated document classification and facet schema enrichment techniques. We believe that collaborative faceted classification can improve access to many digital humanities collections.

Acknowledgements

This project is supported in part by the United States National Science Foundation, Award No. 0713290.

References

[1] Hearst, M.A., Clustering versus Faceted Categories for Information Exploration. *Communications of the ACM*, 2006, 49(4).

[2] Quintarelli, E., L. Rosati, and Resmini, A. *Facetag: Integrating Bottom-up and Top-down Classification in a Social Tagging System*. EuroIA 2006, Berlin.

[3] Wu, H. and M.D. Gordon, Collaborative filing in a document repository. *SIGIR* 2004: p. 518-519

[4] Heymann, P. and Garcia-Molina, H., *Collaborative Creation of Communal Hierarchical Taxonomies in Social Tagging Systems*. Stanford Technical Report InfoLab 2006-10, 2006.

[5] Joachims, T. Text categorization with support vector machines. In *Proceedings of 10th European Conference on Machine Learning*, pages 137–142, April 1998.

[6] Sanderson, M. and B. Croft, Deriving concept hierarchies from text. *SIGIR* 1999: p. 206-213.

[7] Schmitz and Patrick, Inducing Ontology from Flickr Tags. *Workshop in Collaborative Web Tagging*, 2006.

[8] *WordNet: An Electronic Lexical Database*. Christiane Fellbaum (editor). 1998. The MIT Press, Cambridge, MA.

Annotated Facsimile Editions: Defining Macro-level Structure for Image-Based Electronic Editions

Neal Audenaert
neal.audenaert@gmail.com
Texas A&M University, USA

Richard Furuta
furuta@cs.tamu.edu
Texas A&M University, USA

Introduction

Facsimile images form a major component in many digital editing projects. Well-known projects such as the *Blake Archive* [Eaves 2007] and the *Rossetti Archive* [McGann 2007] use facsimile images as the primary entry point to accessing the visually rich texts in their collections. Even for projects focused on transcribed electronic editions, it is now standard practice to include high-resolution facsimile.

Encoding standards and text processing toolkits have been the focus of significant research. Tools, standards, and formal models for encoding information in image-based editions have only recently begun to receive attention. Most work in this area has centered on the digitization and presentation of visual materials [Viscomi 2002] or detailed markup and encoding of information within a single image [Lecolinet 2002, Kiernan 2004, Dekhtyar 2006]. Comparatively little has been work has been done on modeling the large-scale structure of facsimile editions. Typically, the reading interface that presents a facsimile determines its structure.

Separating the software used to model data from that used to build user interfaces has well-known advantages for both engineering and digital humanities practices. To achieve this separation, it is necessary to develop a model of a facsimile edition that is independent of the interface used to present that edition.

In this paper, we present a unified approach for representing linguistic, structural, and graphical content of a text as an Annotated Facsimile Edition (AFED). This model grows out of our experience with several digital facsimile edition projects over more than a decade, including the *Cervantes Project* [Furuta 2001], the *Digital Donne* [Monroy 2007a], and the *Nautical Archaeology Digital Library* [Monroy 2007b]. Our work on these projects has emphasized the need for an intuitive conceptual model of a digital facsimile. This model can then serve as the basis for a core software module that can be used across projects without requiring extensive modification by software developers. Drawing on our prior work we have distilled five primary goals for such a model:

• Openness: Scholars' focused research needs are highly specific, vary widely between disciplines, and change over time. The model must accommodate new information needs as they arise.

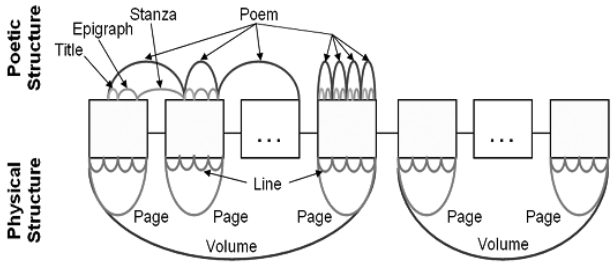


Figure 1: A simplified diagram showing an edition of collected poems (in two volumes) represented as an annotated facsimile edition.

• Non-hierarchical: Facsimile editions contain some information that should be presented hierarchically, but they cannot be adequately represented as a single, properly nested hierarchy.

• Restructuring: A facsimile is a representation of the physical form of a document, but the model should enable applications to restructure the original form to meet specific needs.

• Alignment: Comparison between varying representations of the same work is a fundamental task of humanities research. The model must support alignment between facsimiles of different copies of a work.

Annotated Facsimile Editions

The Annotated Facsimile Edition (AFED) models the macro level structure of facsimile editions, representing them as a stream of images with annotations over that stream. Figure 1 shows a simplified diagram illustrating a two-volume edition of collected poems. Annotations encode the structure of the document and properties of the structural elements they represent. Separate annotation streams encode multiple analytical perspectives. For example, in figure 1, the annotations shown below the image stream describe the physical structure of the edition (volumes, pages, and lines) while the annotations shown above the image stream describe the poetic structure (poems, titles, epigraphs, stanzas). Annotations within a single analytical perspective—but not those from different perspectives—follow a hierarchical structure.

The Image Stream

The image stream intuitively corresponds to the sequential ordering of page images in a traditional printed book. These images, however, need not represent actual "pages." An image might show a variety of artifacts including an opening of a book, a fragment of a scroll, or an unbound leaf of manuscript notes. While it is natural to treat facsimile images sequentially, any particular linear sequence represents an implementation decision—a decision that may not be implied by the physical document. For example, an editor may choose to arrange an edition of letters according to the date written, recipient, or thematic content. The image stream, therefore, is an implementation detail of the model. The structure of the edition is specified explicitly by the annotations.

Annotation Management	
Perspective	Analytical perspective e.g., physical structure, narrative elements, poetic.
Type	The name of this type of annotation, e.g., page, volume, chapter, poem, stanza
Start Index	The index into the image stream where this annotation starts.
Stop Index	The index into the image stream where this annotation ends.
Sequence	A number for resolving the sequence of multiple annotations on the same page.
Content	
Canonical Name	A canonical name that uniquely identifies this content relative to a domain specific classification scheme.
Display Name	The name to be displayed when referring to an instance this annotation
Properties	A set of key/value pairs providing domain specific information about the annotation.
Transcriptions	A set of transcriptions of the content that this annotation specifies.
Structural Information	
Parent	A reference to the parent of this annotation.
Children	A list of references to the children of this annotation

Table 1: Information represented by an annotation.

Many historical texts exist only as fragments. Many more have suffered damage that results in the lost of a portion of the original text. Despite this damage, the general content and characteristics of the text may be known or hypothesized based on other sources. In other cases, while the original artifact may exist, a digital representation of all or part of the artifact may unavailable initially. To enable scholars to work with missing or unavailable portions of a facsimile, we introduce the notion of an abstract image. An abstract image is simply a placeholder for a known or hypothesized artifact of the text for which no image is available. Annotations attach to abstract images in the same way they attach to existing images.

Annotations

Annotations are the primary means for representing structural and linguistic content in the AFED. An annotation identifies a range of images and specifies properties about those images. Table 1 lists the information specified by each annotation. Properties in italics are optional. As shown in this table, annotations support three main categories of information: annotation management, content, and structural information.

The annotation management and structural information categories contain record keeping information. Structural information describes the hierarchical structure of annotation within an analytical perspective. The annotation management category specifies the annotation type and identifies the image content referenced by the annotation. The sequence number is an identifier used by AFED to determine the relative ordering of multiple annotations that have the same starting index. AFED is agnostic to the precise semantics of this value. The annotation type determines these semantics. For example, a paragraph annotation may refer to the paragraph number relative to a page, chapter, or other structural unit.

The content category describes the item referenced by the annotation. Annotations support two naming conventions. To facilitate comparison between documents, an annotation may specify a canonical name according to a domain specific naming convention. Canonical names usually do not match the name given to the referenced item by the artifact itself and are rarely appropriate for display to a general audience. Accordingly, the annotation requires the specification of a name suitable for display.

Descriptive metadata can be specified as a set of key/value properties. In addition to descriptive metadata, annotations support multiple transcriptions. Multiple transcriptions allow alternate perspectives of the text; for example, a paleographic transcription to support detailed linguistic analysis and a normalized transcription to facilitate reading. Transcriptions may also include translations.

AFED's annotation mechanism defines a high-level syntactical structure that is sufficient to support the basic navigational needs of most facsimile projects. By remaining agnostic to semantic details, it allows for flexible, project specific customization. Where projects need to support user interactions that go beyond typical navigation scenarios, these interactions can be integrated into the user interface without requiring changes to the lower-level tools used to access the facsimile.

Discussion

AFED has proven to be a useful model in our work. We have deployed a proof of concept prototype based on the AFED model. Several of the facsimile editions constructed by the *Cervantes Project* use this prototype behind the scenes. Given its success in these reader's interfaces, we are working to develop a Web-based editing toolkit. This application will allow editors to quickly define annotations and use those annotations to describe a facsimile edition. We anticipate completing this tool by the summer of 2008.

By using multiple, hierarchical annotation streams, AFED's expressive power falls under the well-studied class of document models, known as OHCO (ordered hierarchy of content objects). Specifically, it is an instance of a revised form of this generic model known as OHCO-3, [Renear 1996]. Whereas most prior research and development associated with the OHCO model has focused on XML-based, transcribed content, we have applied this model to the task of representing macro-level structures in facsimile editions.

Focusing on macro-level document structure partially isolates the AFED model from the non-hierarchical nature of documents both in terms of the complexity of the required data structures, and in terms of providing simplified model to facilitate system implementation. If warranted by future applications, we can relax AFED's hierarchical constraint. Relaxing this constraint poses no problems with the current prototype; however, further investigation is needed to determine potential benefits and drawbacks.

In addition to macro-level structures, a document model that strives to represent the visual content of a document for scholarly purposes must also account for fine-grained structures present in individual images and provide support for encoded content at a higher level of detail. We envision using the AFED model in conjunction with models tailored for these low-level structures. We are working to develop a model for representing fine-grained structure in visually complex documents grounded in spatial hypermedia theory.

Acknowledgements

This material is based upon work support by National Science Foundation under Grant No. IIS-0534314.

References

[Dekhtyar 2006] Dekhtyar, A., et al. Support for XML markup of image-based electronic editions. *International Journal on Digital Libraries* 6(1) 2006, pp. 55-69.

[Eaves 2007] Eaves, M., Essick, R.N., Viscomi, J., eds. *The William Blake Archive*. http://www.blakearchive.org/ [24 November 2007]

[Furuta 2001] Furuta, R., et al. The Cervantes Project: Steps to a Customizable and Interlinked On-Line Electronic Variorum Edition Supporting Scholarship. In *Proceedings of ECDL 2001, LNCS, 2163*. Springer-Verlag: Heidelberg, pp. 71-82.

[Kiernan 2004] Kiernan K., et al. The ARCHway Project: Architecture for Research in Computing for Humanities through Research, Teaching, and Learning. *Literary and Linguistic Computing* 2005 20(Suppl 1):69-88.

[Lecolinet 2002] Lecolinet, E., Robert, L. and Role. F. Text-image Coupling for Editing Literary Sources. *Computers and the Humanities* 36(1): 2002 pp 49-73.

[McGann 2007] McGann, J., *The Complete Writings and Pictures of Dante Gabriel Rossetti*. Institute for Advanced Technology in the Humanities, University of Virginia. http://www.rossettiarchive.org/ [24 November 2007]

[Monroy 2007a] Monroy, C., Furuta, R., Stringer, G. Digital Donne: Workflow, Editing Tools and the Reader's Interface of a Collection of 17th-century English Poetry. In *Proceedings of Joint Conference on Digital Libraries JCDL 2007* (Vancouver, BC, June 2007), ACM Press: New York, NY, pp. 411-412.

[Monroy 2007b] Monroy, C., Furuta, R., Castro, F. Texts, Illustrations, and Physical Objects: The Case of Ancient Shipbuilding Treatises. In *Proceedings of ECDL 2007, LNCS, 4675*. Springer-Verlag: Heidelberg, pp. 198-209.

[Renear 1996] Renear, A., Mylonas, E., Durand, D. Refining our Notion of What Text Really Is: The Problem of Overlapping Hierarchies. In Ide, N., Hockey, S. *Research in Humanities Computing.* Oxford: Oxford University Press, 1996.

[Viscomi 2002] Viscomi, J. (2002). 'Digital Facsimiles: Reading the William Blake Archive'. Kirschenbaum, M. (ed.) Image-based Humanities Computing. spec. issue of *Computers and the Humanities*, 36(1): 27-48.

CritSpace: Using Spatial Hypertext to Model Visually Complex Documents

Neal Audenaert
neal.audenaert@gmail.com
Texas A&M University, USA,

George Lucchese
george_lucchese@tamu.edu
Texas A&M University, USA,

Grant Sherrick
sherrick@csdl.tamu.edu
Texas A&M University, USA

Richard Furuta
furuta@cs.tamu.edu
Texas A&M University, USA

In this paper, we present a Web-based interface for editing visually complex documents, such as modern authorial manuscripts. Applying spatial hypertext theory as the basis for designing this interface enables us to facilitate both interaction with the visually complex structure of these documents and integration of heterogeneous sources of external information. This represents a new paradigm for designing systems to support digital textual studies. Our approach emphasizes the visual nature of texts and provides powerful tools to support interpretation and creativity. In contrast to purely image-based systems, we are able to do this while retaining the benefits of traditional textual analysis tools.

Introduction

Documents express information as a combination of written words, graphical elements, and the arrangement of these content objects in a particular media. Digital representations of documents—and the applications built around them—typically divide information between primarily textual representations on the one hand (e.g., XML encoded documents) and primarily graphical based representations on the other (e.g., facsimiles).

Image-based representations allow readers access to high-quality facsimiles of the original document, but provide little support for explicitly encoded knowledge about the document. XML-based representations, by contrast, are able to specify detailed semantic knowledge embodied by encoding guidelines such as the TEI [Sperberg-McQueen 2003]. This knowledge forms the basis for building sophisticated analysis tools and developing rich hypertext interfaces to large document collections and supporting materials. This added power comes at a price. These approaches are limited by the need to specify all relevant content explicitly. This is, at best, a time consuming and expensive task and, at worst, an impossible one [Robinson 2000]. Furthermore, in typical systems, access to these texts

mediated almost exclusively by the transcribed linguistic content, even when images alongside their transcriptions.

By adopting spatial hypertext as a metaphor for representing document structure, we are able to design a system that emphasizes the visually construed contents of a document while retaining access to structured semantic information embodied in XML-based representations. Dominant hypertext systems, such as the Web, express document relationships via explicit links. In contrast, spatial hypertext expresses relationships by placing related content nodes near each other on a two-dimensional canvas [Marshall 1993]. In addition to spatial proximity, spatial hypertext systems express relationships through visual properties such as background color, border color and style, shape, and font style. Common features of spatial hypermedia systems include parsers capable of recognizing relationships between objects such as lists, list headings, and stacks, structured metadata attributes for objects, search capability, navigational linking, and the ability to follow the evolution of the information space via a history mechanism.

The spatial hypertext model has an intuitive appeal for representing visually complex documents. According to this model, documents specify relationships between content objects (visual representations of words and graphical elements) based on their spatial proximity and visual similarity. This allows expression of informal, implicit, and ambiguous relationships—a key requirement for humanities scholarship. Unlike purely image-based representations, spatial hypertext enables users to add formal descriptions of content objects and document structure incrementally in the form of structured metadata (including transcriptions and markup). Hypermedia theorists refer to this process as "incremental formalism" [Shipman 1999]. Once added, these formal descriptions facilitate system support for text analysis and navigational hypertext.

Another key advantage of spatial hypertext is its ability to support "information triage" [Marshall 1997]. Information triage is the process of locating, organizing, and prioritizing large amounts of heterogeneous information. This is particularly helpful in supporting information analysis and decision making in situations where the volume of information available makes detailed evaluation of it each resource impossible. By allowing users to rearrange objects freely in a two-dimensional workspace, spatial hypertext systems provide a lightweight interface for organizing large amounts of information. In addition to content taken directly from document images, this model encourages the inclusion of visual surrogates for information drawn from numerous sources. These include related photographs and artwork, editorial annotations, links to related documents, and bibliographical references. Editors/readers can then arrange this related material as they interact with the document to refine their interpretive perspective. Editors/readers are also able to supply their own notes and graphical annotations to enrich the workspace further.

System Design

We are developing CritSpace as a proof of concept system using the spatial hypertext metaphor as a basis for supporting digital textual studies. Building this system from scratch, rather than using an existing application, allows us to tailor the design to meet needs specific to the textual studies domain (for example, by including document image analysis tools). We are also able to develop this application with a Web-based interface tightly integrated with a digital archive containing a large volume of supporting information (such as artwork, biographical information, and bibliographic references) as well as the primary documents.

Initially, our focus is on a collection of manuscript documents written by Picasso [Audenaert 2007]. Picasso's prominent use of visual elements, their tight integration with the linguistic content, and his reliance on ambiguity and spatial arrangement to convey meaning make this collection particularly attractive [Marin 1993, Michaël 2002]. These features also make his work particularly difficult to represent using XML-based approaches. More importantly, Picasso's writings contain numerous features that exemplify the broader category of modern manuscripts including documents in multiple states, extensive authorial revisions, editorial drafts, interlinear and marginal scholia, and sketches and doodles.

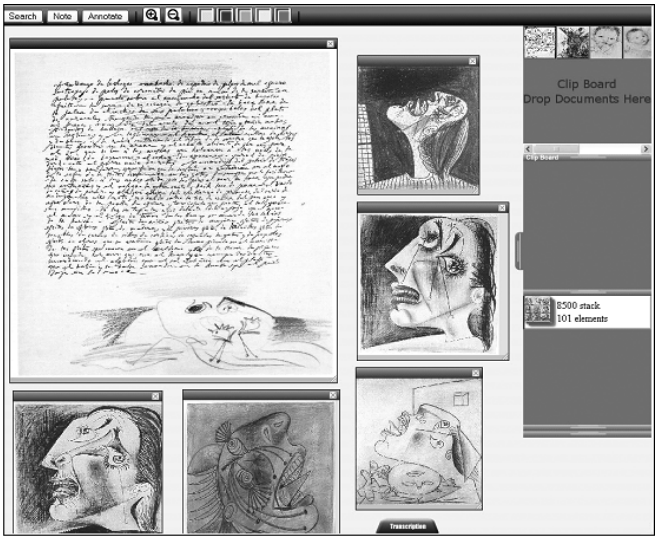


Figure 1: Screenshot of CritSpace showing a document along with several related preparatory sketches Picasso made for Guernica about the same time.

CritSpace provides an HTML based interface for accessing the collection maintained by the *Picasso Project* [Mallen 2007] that contains nearly 14,000 artworks (including documents) and 9,500 biographical entries. CritSpace allows users to arrange facsimile document images in a two dimensional workspace and resize and crop these images. Users may also search and browse the entire holdings of the digital library directly from CritSpace, adding related artworks and biographical information to the workspace as desired. In addition to content taken from the digital library, users may add links to other

material available on the Web, or add their own comments to the workspace in the form of annotations. All of these items are displayed as content nodes that can be freely positioned and whose visual properties can be modified. Figure 1 shows a screenshot of this application that displays a document and several related artworks.

CritSpace also introduces several features tailored to support digital textual studies. A tab at the bottom of the display opens a window containing a transcription of the currently selected item. An accordion-style menu on the right hand side provides a clipboard for temporarily storing content while rearranging the workspace, an area for working with groups of images, and a panel for displaying metadata and browsing the collection based on this metadata. We also introduce a full document mode that allows users to view a high-resolution facsimile. This interface allows users to add annotations (both shapes and text) to the image and provides a zooming interface to facilitate close examination of details.

Future Work

CritSpace provides a solid foundation for understanding how to apply spatial hypertext as a metaphor for interacting with visually complex documents. This perspective opens numerous directions for further research.

A key challenge is developing tools to help identify content objects within a document and then to extracted these object in a way that will allow users to manipulate them in the visual workspace. Along these lines, we are working to adapt existing techniques for foreground/background segmentation [Gatos 2004], word and line identification [Manmatha 2005], and page segmentation [Shafait 2006]. We are investigating the use of Markov chains to align transcriptions to images semi-automatically [Rothfeder 2006] and expectation maximization to automatically recognize dominant colors for the purpose of separating information layers (for example, corrections made in red ink).

Current implementations of these tools require extensive parameter tuning by individuals with a detailed understanding of the image processing algorithms. We plan to investigate interfaces that will allow non-experts to perform this parameter tuning interactively.

Modern spatial hypertext applications include a representation of the history of the workspace [Shipman 2001]. We are interested in incorporating this notion, to represent documents, not as fixed and final objects, but rather objects that have changed over time. This history mechanism will enable editors to reconstruct hypothetical changes to the document as authors and annotators have modified it. It can also be used to allowing readers to see the changes made by an editor while constructing a particular form of the document.

While spatial hypertext provides a powerful model for representing a single workspace, textual scholars will need tools to support the higher-level structure found in documents, such as chapters, sections, books, volumes. Further work is needed to identify ways in which existing spatial hypertext models can be extended express relationships between these structures and support navigation, visualization, and editing.

Discussion

Spatial hypertext offers an alternative to the dominate view of text as an "ordered hierarchy of content objects" (OCHO) [DeRose 1990]. The OCHO model emphasizes the linear, linguistic content of a document and requires explicit formalization of structural and semantic relationships early in the encoding process. For documents characterized by visually constructed information or complex and ambiguous structures, OCHO may be overly restrictive.

In these cases, the ability to represent content objects graphically in a two dimensional space provides scholars the flexibility to represent both the visual aspects of the text they are studying and the ambiguous, multi-facetted relationships found in those texts. Furthermore, by including an incremental path toward the explicit encoding of document content, this model enables the incorporation of advanced textual analysis tools that can leverage both the formally specified structure and the spatial arrangement of the content objects.

Acknowledgements

This material is based upon work support by National Science Foundation under Grant No. IIS-0534314.

References

[Audenaert 2007] Audenaert, N. et al. Viewing Texts: An Art-Centered Representation of Picasso's Writings. In *Proceedings of Digital Humanities 2007* (Urbana-Champaign, IL, June, 2007), pp. 14-17.

[DeRose 1990] DeRose, S., Durand, D., Mylonas, E., Renear, A. What is Text Really? *Journal of Computing in Higher Education*. 1(2), pp. 3-26.

[Gatos 2004] Gatos, B., Ioannis, P., Perantonis, S. J., An Adaptive Binarization Technique for Low Quality Historical Documents. In *Proceedings of Document Analysis Systems 2004*. LNCS 3163 Springer-Verlag: Berlin, pp. 102-113.

[Mallen 2006] Mallen, E., ed. (2007) *The Picasso Project*. Texas A&M University http://picasso.tamu.edu/ [25 November 2007]

[Manmatha 2005] Manmatha, R., Rothfeder, J. L., A Scale Space Approach for Automatically Segmenting Words from Historical Handwritten Documents. In *IEEE Transactions on*

Pattern Analysis and Machine Intelligence, 28(8), pp. 1212-1225.

[Marin 1993] Marin, L. Picasso: Image Writing in Process. trans. by Sims, G. In *October 65* (Summer 1993), MIT Press: Cambridge, MA, pp. 89-105.

[Marshall 1993] Marshall, C. and Shipman, F. Searching for the Missing Link: Discovering Implicit Structure in Spatial Hypertext. In *Proceedings of Hypertext '93* (Seattle WA, Nov. 1993), ACM Press: New York, NY, pp. 217-230.

[Marshall 1997] Marshall, C. and Shipman, F. Spatial hypertext and the practice of information triage. In *Proceedings of Hypertext '97* (Southampton, UK, Nov. 1997), ACM Press: New York, NY, pp. 124-133.

[Michaël 2002] Michaël, A. Inside Picasso's Writing Laboratory. Presented at Picasso: The Object of the Myth. November, 2002. http://www.picasso.fr/anglais/cdjournal.htm [December 2006]

[Renear 1996] Renear, A., Mylonas, E., Durand, D. Refining our Notion of What Text Really Is: The Problem of Overlapping Hierarchies. In Ide, N., Hockey, S. *Research in Humanities Computing*. Oxford: Oxford University Press, 1996.

[Robinson 2000] Robinson, P. Ma(r)king the Electronic Text: How, Why, and for Whom? In Joe Bray et. al. *Ma(r)king the Text: The Presentation of Meaning on the Literary Page*. Ashgate: Aldershot, England, pp. 309-28.

[Rothfeder 2006] Rothfeder, J., Manmatha, R., Rath, T.M., Aligning Transcripts to Automatically Segmented Handwritten Manuscripts. In *Proceedings of Document Analysis Systems 2006*. LNCS 3872 Springer-Verlag: Berlin, pp. 84-95.

[Shafait 2006] Shafait, F., Keysers, D., Breuel, T. Performance Comparison of Six Algorithms for Page Segmentation. In *Proceedings of Document Analysis Systems 2006.* LNCS 3872 Springer-Verlag: Berlin, pp. 368-379.

[Shipman 1999] Shipman, F. and McCall, R. Supporting Incremental Formalization with the Hyper-Object Substrate. In *ACM Transactions on Information Systems* 17(2), ACM Press: New York, NY, pp. 199-227.

[Shipman 2001] Shipman, F., Hsieh, H., Maloor, P. and Moore, M. The Visual Knowledge Builder: A Second Generation Spatial Hypertext. In *Proceedings of Twelfth ACM Conference on Hypertext and Hypermedia* (Aarhus Denmark, August 2001), ACM Press, pp. 113-122.

[Sperberg-McQueen 2003] Sperberg-Mcqueen, C. and Burnard, L. *Guidelines for Electronic Text Encoding and Interchange: Volumes 1 and 2: P4*, University Press of Virginia, 2003.

Glimpses though the clouds: collocates in a new light

David Beavan

d.beavan@englang.arts.gla.ac.uk

University of Glasgow, UK

This paper demonstrates a web-based, interactive data visualisation, allowing users to quickly inspect and browse the collocational relationships present in a corpus. The software is inspired by tag clouds, first popularised by on-line photograph sharing website Flickr (www.flickr.com). A paper based on a prototype of this Collocate Cloud visualisation was given at Digital Resources for the Humanities and Arts 2007. The software has since matured, offering new ways of navigating and inspecting the source data. It has also been expanded to analyse additional corpora, such as the British National Corpus (http://www.natcorp.ox.ac.uk/), which will be the focus of this talk.

Tag clouds allow the user to browse, rather than search for specific pieces of information. Flickr encourages its users to add tags (keywords) to each photograph uploaded. The tags associated with each individual photograph are aggregated; the most frequent go on to make the cloud. The cloud consists of these tags presented in alphabetical order, with their frequency displayed as variation in colour, or more commonly font size. Figure 1 is an example of the most popular tags at Flickr:

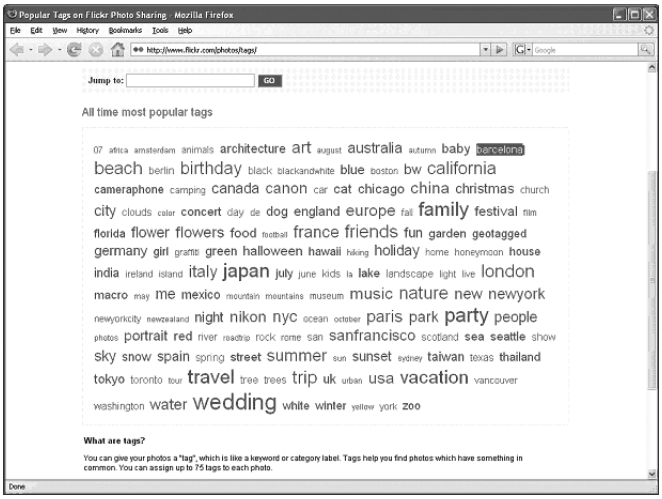


Figure 1. Flickr tag cloud showing 125 of the most popular photograph keywords

http://www.flickr.com/photos/tags/ (accessed 23 November 2007)

The cloud offers two ways to access the information. If the user is looking for a specific term, the alphabetical ordering of the information allows it to be quickly located if present. More importantly, as a tool for browsing, frequent tags stand out visually, giving the user an immediate overview of the data. Clicking on a tag name will display all photographs which contain that tag.

The cloud-based visualisation has been successfully applied to language. McMaster University's TAPoR Tools (http://taporware.mcmaster.ca/) features a 'Word Cloud' module, currently in beta testing. WMatrix (http://ucrel.lancs.ac.uk/wmatrix/) can compare two corpora by showing log-likelihood results in cloud form. In addition to other linguistic metrics, internet book seller Amazon provides a word cloud, see figure 2.

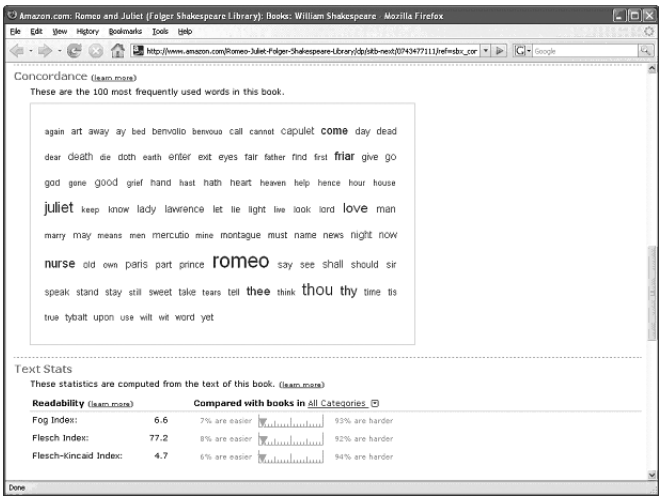


Figure 2. Amazon.com's 'Concordance' displaying the 100 most frequent words in Romeo and Juliet

http://www.amazon.com/Romeo-Juliet-Folger-Shakespeare-Library/dp/sitb-next/0743477111/ref=sbx_con/104-4970220-2133519?ie=UTF8&qid=1179135939&sr=1-1#concordance (accessed 23 November 2007)

In this instance a word frequency list is the data source, showing the most frequent 100 words. As with the tag cloud, this list is alphabetically ordered, the font size being proportionate to its frequency of usage. It has all the benefits of a tag cloud; in this instance clicking on a word will produce a concordance of that term.

This method of visualisation and interaction offers another tool for corpus linguists. As developer for an online corpus project, I have found that the usability and sophistication of our tools have been important to our success. Cloud-like displays of information would complement our other advanced features, such as geographic mapping and transcription synchronisation.

The word clouds produced by TAPoR Tools, WMatrix and Amazon are, for browsing, an improvement over tabular statistical information. There is an opportunity for other corpus data to be enhanced by using a cloud. Linguists often use collocational information as a tool to examine language use. Figure 3 demonstrates a typical corpus tool output:

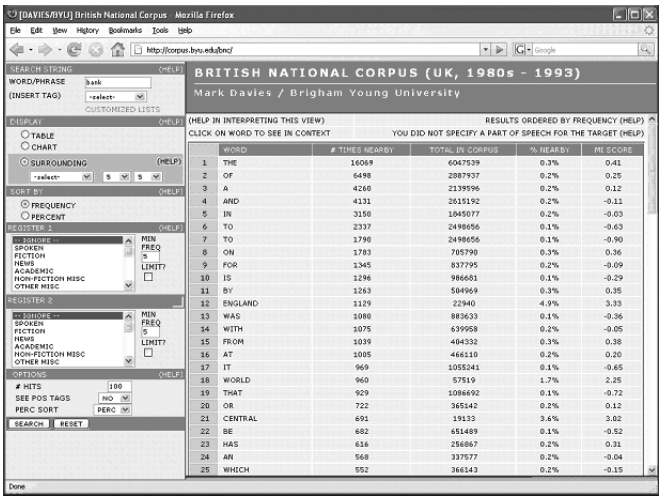


Figure 3. British National Corpus through interface developed by Mark Davies, searching for 'bank', showing collocates

http://corpus.byu.edu/bnc/ (accessed 23 November 2007)

The data contained in the table lends itself to visualisation as a cloud. As with the word cloud, the list of collocates can be displayed alphabetically. Co-occurrence frequency, like word frequency, can be mapped to font size. This would produce an output visually similar to the word cloud. Instead of showing all corpus words, they would be limited to those surrounding the chosen node word.

Another valuable statistic obtainable via collocates is that of collocational strength, the likelihood of two words co-occurring, measured here by MI (Mutual Information). Accounting for this extra dimension requires an additional visual cue to be introduced, one which can convey the continuous data of an MI score. This can be solved by varying the colour, or brightness of the collocates forming the cloud. The end result is shown in figure 4:

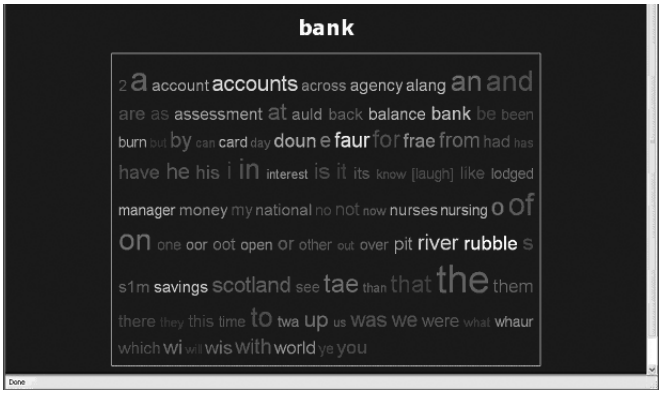


Figure 4. Demonstration of collocate cloud, showing node word 'bank'

The collocate cloud inherits all the advantages of previous cloud visualisations: a collocate, if known, can be quickly located due to the alphabetical nature of the display. Frequently occurring collocates stand out, as they are shown in a larger typeface, with collocationally strong pairings highlighted using brighter

formatting. Therefore bright, large collocates are likely to be of interest, whereas dark, small collocates perhaps less so. Hovering the mouse over a collocate will display statistical information, co-occurrence frequency and MI score, as one would find from the tabular view.

The use of collocational data also presents additional possibilities for interaction. A collocate can be clicked upon to produce a new cloud, with the previous collocate as the new node word. This gives endless possibilities for corpus exploration and the investigation of different domains. Occurrences of polysemy can be identified and expanded upon by following the different collocates. Particular instances of usage are traditionally hidden from the user when viewing aggregated data, such as the collocate cloud. The solution is to allow the user to examine the underlying data by producing an optional concordance for each node/collocate pairing present. Additionally a KWIC concordance can be generated by examining the node word, visualising the collocational strength of the surrounding words. These concordance lines can even by reordered on the basis of collocational strength, in addition to the more traditional options of preceding or succeeding words.

This visualisation may be appealing to members of the public, or those seeking a more practical introduction to corpus linguistics. In teaching use they not only provide analysis, but from user feedback, also act as stimulation in creative writing. Collocate searches across different corpora or document sets may be visualised side by side, facilitating quick identification of differences.

While the collocate cloud is not a substitute for raw data, it does provide a fast and convenient way to navigate language. The ability to generate new clouds from existing collocates extends this further. Both this iterative nature and the addition of collocational strength information gives these collocate clouds greater value for linguistic research than previous cloud visualisations.

The function and accuracy of old Dutch urban designs and maps. A computer assisted analysis of the extension of Leiden (1611)

Jakeline Benavides
j.benavides@rug.nl
University of Groningen, The Netherlands

Charles van den Heuvel
charles.vandenheuvel@vks.knaw.nl
Royal Netherlands Academy of Arts and Sciences, The Netherlands

Historical manuscripts and printed maps of the pre-cadastral[1] period show enormous differences in scale, precision and color. Less apparent are differences in reliability between maps, or between different parts of the same map. True, modern techniques in computer assisted cartography make it possible to measure very accurately such differences between maps and geographic space with very distinct levels of precision and accuracy. However differences in reliability between maps, or between different parts of the same map, are not only due to the accuracy measurement techniques, but also to their original function and context of (re-)use. Historical information about the original context of function and context of (re-)use can give us insight how to measure accuracy, how to choose the right points for geo-referencing and how to rectify digital maps. On the other hand computer assisted cartography enables us to trace and to visualize important information about mapmaking, especially when further historical evidence is missing, is hidden or is distorted, consciously or unconsciously. The proposed paper is embedded in the project: *Paper and Virtual Cities*, (subsidized by the Netherlands Organization for Scientific Research) that aims at developing methodologies that a) permit researchers to use historical maps and related sources more accurately in creating in digital maps and virtual reconstructions of cities and b) allow users to recognize better technical manipulations and distortions of truth used in the process of mapmaking.[2]

In this paper we present as one of the outcomes of this project a method that visualizes different levels of accuracy in and between designs and maps in relation to their original function to assess their quality for re-use today. This method is presented by analyzing different 17th century designs, manuscript and engraved maps of the city of Leiden, in particular of the land surveyor, mapmaker Jan Pietersz. Dou. The choice for Leiden and Dou is no coincidence.

One of the reasons behind differences the accuracy of maps is the enormous variety in methods and measures used in land surveying and mapmaking in the Low Countries.[3] This variety was the result of differences in the private training and the

backgrounds of the surveyors, in the use of local measures and in the exam procedures that differed from province to province.[4] These differences would last until the 19th Century. However, already by the end of the 16th Century we see in and around Leiden the first signs of standardization in surveying techniques and the use of measures.[5] First of all, the Rhineland rod (3.767 meter) the standard of the water-administration body around Leiden is used more and more, along local measures in the Low Countries and in Dutch expansion overseas. A second reason to look at Leiden in more detail is that in 1600 a practical training school in land surveying and fortification would be founded in the buildings of its university, the so-called Duytsche Mathematique, that turned out to be very successful not only in the Low Countries, but also in other European countries. This not only contributed to the spread and reception of the Rhineland rod, but also to the dissemination of more standardized ways of land surveying and fortification.[6] The instructional material of the professors of this training school and the notes of their pupils are still preserved, which allows us to study the process of surveying and mapmaking in more detail.

The reason to look into the work of Jan Pietersz. Dou is his enormous production of maps. Westra (1994) calculated that at least 1150 maps still exist.[7] Of the object of our case study alone, the city of Leiden, Dou produced at least 120 maps between 1600 and 1635, ranging from property maps, designs for extensions of the city and studies for civil engineering works etc. We will focus on the maps that Dou made for the urban extension of the city of Leiden of 1611. Sometimes these (partial) maps were made for a specific purpose; in other cases Dou tried in comprehensive maps, combining property estimates and future designs, to tackle problems of illegal economic activities, pollution, housing and fortification. Since these measurements were taken in his official role of, sworn-in land surveyor we can assume that they were supposed to be accurate. This variety in designs and maps for the same area allows us to discuss accuracy in relation to function and (re-)use of maps. We will also explain that the differences between designs and maps require different methods of geo-referencing and analysis.

In particular, we will give attention to one design map of Dou for the northern part of the city of Leiden RAL PV 1002-06 (Regionaal Archief Leiden) to show how misinterpretation of features lead to unreliable or biased decisions when the historical context is not taken into account, even when we can consider the results, in terms of accuracy, satisfactory. Since Dou later also made maps for commercial purposes of the same northern extension of Leiden it is interesting to compare these maps.

Conclusions are drawn addressing the question of whether Dou used the same measurements to produce a commercial map or that he settled for less accuracy given the different purpose of the later map compared to his designs and property maps. To answer this question, we use modern digital techniques of geo-processing[8] to compare the old maps to modern cartographical resources and to the cadastral map of the 1800s in order to determine how accurate the various maps in question are. We do this, by using the American National Standard for Spatial Data Accuracy (NSSDA) to define accuracy at a 95% confidence level. By point-based analysis we link distributional errors to classified features in order to find a relationship between accuracy and map function.[9]

Notes

(1) Cadastral mapping refers to the "mapping of property boundaries, particularly to record the limitation of title or for the assessment of taxation". The term "cadastral" is also used for referring to surveys and resurveys of public lands (Neumann, J. Encyclopedic Dictionary of Cartography in 25 Languages 1.16, 1997).

(2) The project *Paper and Virtual Cities. New methodologies for the use of historical sources in computer-assisted urban cartography* (2003-2008) is a collaboration between the department of Alfa-Informatics of the University Groningen and the Virtual Knowledge Studio of the Royal Netherlands Academy of Arts and Sciences subsidized by the Netherlands Organization for Scientific Research (NWO). http://www.virtualknowledgestudio.nl/projects/paper-virtualcities.php

(3) Meskens, A, *Wiskunde tussen Renaissance en Barok. Aspecten van wiskunde-beoefening te Antwerpen 1550-1620, [Publikaties SBA/MVC 41-43]*, [PhD, University of Antwerp], Antwerp, 1995.

H.C. Pouls, "Landmeetkundige methoden en instrumenten voor 1800", in Stad in kaart, Alphen aan den Rijn, pp. 13-28

Winter, P.J. van, *Hoger beroepsonderwijs avant-la-lettre: Bemoeiingen met de vorming van landmeters en ingenieurs bij de Nederlandse universiteiten van de 17e en 18e eeuw,* Amsterdam/Oxford/New York, 1988.

(4) Muller, E., Zanvliet, K., eds., *Admissies als landmeter in Nederland voor 1811: Bronnen voor de geschiedenis van de landmeetkunde en haar toepassing in administratie, architectuur, kartografie en vesting-en waterbouwkunde*, Alphen aan den Rijn 1987

(5) Zandvliet, K., *Mapping for Money. Maps, plans and topographic paintings and their role in Dutch overseas expansion during the 16th and 17th centuries.* (PhD Rijksuniversiteit Leiden], Amsterdam, 1998 describes this development as part of a process of institutionalization of mapmaking, esp, pp. 75-81.

(6) Taverne, E.R.M., *In 't land van belofte: in de nieue stadt. Ideaal en werkelijkheid van de stadsuitleg in de Republiek 1580-1680,* [PhD, University of Groningen]Maarssen, 1978.

Heuvel, C. van den, 'Le traité incomplet de l'Art Militaire et l'ínstruction pour une école des ingénieurs de Simon Stevin', *Simon Stevin (1548-1620) L'emergence de la nouvelle science*, (tentoonstellingscatalogus/catalogue Koninklijk Bibliotheek Albert I, Brussel/Bibliothèque Royale Albert I, Bruxelles, 17-09-2004-30-10-2004) Brussels 2004, pp. 101-111. idem, "The training of noblemen in the arts and sciences in the Low Countries around 1600. Treatises and instructional materials" in *Alessandro Farnese e le Fiandre/Alexander and the Low Countries* (in print)

(7) Frans Westra, Jan Pietersz. Dou (1573-1635). "Invloedrijk landmeter van Rijnland", *Caert-thresoor*, 13e jaargang 1994, nr. 2, pp. 37-48

(8) During geo-referencing a mathematical algorithm is used for scaling, rotating and translating the old map to give modern coordinates to it and allow further comparisons to modern sources. This algorithm is defined by a kind of transformation we decide to use based on the selection of a number of control points (GCPS). These items are described in detail later in this paper. All this processing was done by using different kind of software for digital and geographical processing and statistics (PCI Geomatics, ARCGIS, Autocad, MS Excel, among others).

(9) Jakeline Benavides and John Nerbonne. Approaching Quantitative Accuracy in Early Dutch City Maps. XXIII International cartographic Conference. ISBN 978-5-9901203-1-0 (CD-ROM). Moscow, 2007

AAC-FACKEL and BRENNER ONLINE. New Digital Editions of Two Literary Journals

Hanno Biber
hanno.biber@oeaw.ac.at
Austrian Academy of Sciences, Austria

Evelyn Breiteneder
evelyn.breiteneder@oeaw.ac.at
Austrian Academy of Sciences, Austria

Karlheinz Mörth
karlheinz.moerth@oeaw.ac.at
Austrian Academy of Sciences, Austria

In this paper two highly innovative digital editions will be presented. The digital editions of the historical literary journals "Die Fackel" (published by Karl Kraus in Vienna from 1899 to 1936) and "Der Brenner" (published by Ludwig Ficker in Innsbruck from 1910 to 1954) have been developed within the corpus research framework of the "AAC - Austrian Academy Corpus" at the Austrian Academy of Sciences in collaboration with other researchers and programmers in the AAC from Vienna together with the graphic designer Anne Burdick from Los Angeles. For the creation of these scholarly digital editions the AAC edition philosophy and principles have been made use of whereby new corpus research methods have been applied for questions of computational philology and textual studies in a digital environment. The examples of these online editions will give insights into the potentials and the benefits of making corpus research methods and techniques available for scholarly research into language and literature.

Introduction

The "AAC - Austrian Academy Corpus" is a corpus research unit at the Austrian Academy of Sciences concerned with establishing and exploring large electronic text corpora and with conducting scholarly research in the field of corpora and digital text collections and editions. The texts integrated into the AAC are predominantly German language texts of historical and cultural significance from the last 150 years. The AAC has collected thousands of texts from various authors, representing many different text types from all over the German speaking world. Among the sources, which systematically cover various domains, genres and types, are newspapers, literary journals, novels, dramas, poems, advertisements, essays on various subjects, travel literature, cookbooks, pamphlets, political speeches as well as a variety of scientific, legal, and religious texts, to name just a few forms. The AAC provides resources for investigations into the linguistic and textual properties of these texts and into their historical and cultural qualities. More than 350 million running words of text have been scanned,

digitized, integrated and annotated. The selection of texts is made according to the AAC's principles of text selection that are determined by specific research interests as well as by systematic historical, empirical and typological parameters. The annotation schemes of the AAC, based upon XML related standards, have in the phase of corpus build-up been concerned with the application of basic structural mark-up and selective thematic annotations. In the phase of application development specific thematic annotations are being made exploring questions of linguistic and textual scholarly research as well as experimental and exploratory mark-up. Journals are regarded as interesting sources for corpus research because they comprise a great variety of text types over a long period of time. Therefore, model digital editions of literary journals have been developed: The AAC-FACKEL was published on 1 January 2007 and BRENNER ONLINE followed in October 2007. The basic elements and features of our approach of corpus research in the field of textual studies will be demonstrated in this paper.

AAC-FACKEL

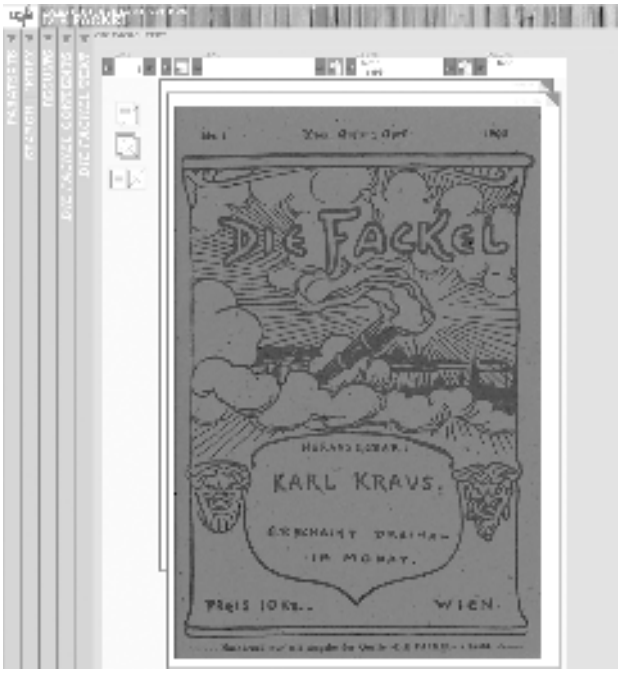


Figure 1. AAC-FACKEL Interface

The digital edition of the journal "Die Fackel" ("The Torch"), published and almost entirely written by the satirist and language critic Karl Kraus in Vienna from 1899 until 1936, offers free online access to 37 volumes, 415 issues, 922 numbers, comprising more than 22.500 pages and 6 million tokens. It contains a fully searchable database of the journal with various indexes, search tools and navigation aids in an innovative and functional graphic design interface, where all pages of the original are available as digital texts and as facsimile images. The work of Karl Kraus in its many forms, of which the journal is the core, can be regarded as one of the most important, contributions to world literature. It is a source for the history of the time, for its language and its moral transgressions. Karl Kraus covers in a typical and idiosyncratic style in thousands of texts the themes of journalism and war, of politics and corruption, of literature and lying. His influential journal comprises a great variety of essays, notes, commentaries, aphorisms and poems. The electronic text, also used for the compilation of a text-dictionary of idioms ("Wörterbuch der Redensarten zu der von Karl Kraus 1899 bis 1936 herausgegebenen Zeitschrift 'Die Fackel'"), has been corrected and enriched by the AAC with information. The digital edition allows new ways of philological research and analysis and offers new perspectives for literary studies.

BRENNER ONLINE

Figure 2. BRENNER ONLINE Interface

The literary journal "Der Brenner" was published between 1910 and 1954 in Innsbruck by Ludwig Ficker. The text of 18 volumes, 104 issues, which is a small segment of the AAC's overall holdings, is 2 million tokens of corrected and annotated text, provided with additional information. "Die Fackel" had set an example for Ludwig Ficker and his own publication. Contrary to the more widely read satirical journal of Karl Kraus, the more quiet "Der Brenner" deals primarily with themes of religion, nature, literature and culture. The philosopher Ludwig Wittgenstein was affiliated with the group and participated in the debates launched by the journal. Among its contributors is the expressionist poet Georg Trakl, the writer Carl Dallago, the translator and cultural critic Theodor Haecker, translator of Søren Kierkegaard and Cardinal Newman into German, the moralist philosopher Ferdinand Ebner and many others. The journal covers a variety of subjects and is an important voice of Austrian culture in the pre and post second world war periods. The digital edition has been made by the AAC in collaboration with the Brenner-Archive of the University

of Innsbruck. Both institutions have committed themselves to establish a valuable digital resource for the study of this literary journal.

Conclusion

The philological and technological principles of digital editions within the AAC are determined by the conviction that the methods of corpus research will enable us to produce valuable resources for scholars. The AAC has developed such model editions to meet these aims. These editions provide well structured and well designed access to the sources. All pages are accessible as electronic text and as facsimile images of the original. Various indexes and search facilities are provided so that word forms, personal names, foreign text passages, illustrations and other resources can be searched for in various ways. The search mechanisms for personal names have been implemented in the interface for BRENNER ONLINE and will be done for "Die Fackel". The interface is designed to be easily accessible also to less experienced users of corpora. Multi-word queries are possible. The search engine supports left and right truncation. The interface of the AAC-FACKEL provides also search mechanisms for linguistic searches, allows to perform lemma queries and offers experimental features. Instead of searching for particular word forms, queries for all the word forms of a particular lemma are possible. The same goes for the POS annotation. The web-sites of both editions are entirely based on XML and cognate technologies. On the character level use of Unicode has been made throughout. All of the content displayed in the digital edition is dynamically created from XML data. Output is produced through means of XSLT style sheets. This holds for the text section, the contents overview and the result lists. We have adopted this approach to ensure the viability of our data for as long a period as possible. Both digital editions have been optimized for use with recent browser versions. One of the basic requirements is that the browser should be able to handle XML-Dom and the local system should be furnished with a Unicode font capable of displaying the necessary characters. The interface has synchronized five individual frames within one single window, which can be alternatively expanded and closed as required. The "Paratext"-section situated within the first frame provides background information and essays. The "Index"-section gives access to a variety of indexes, databases and full-text search mechanisms. The results are displayed in the adjacent section. The "Contents"-section has been developed, to show the reader the whole range of the journal ready to be explored and provides access to the whole run of issues in chronological order. The "Text"-section has a complex and powerful navigational bar so that the reader can easily navigate and read within the journals either in text-mode or in image-mode from page to page, from text to text, from issue to issue and with the help of hyperlinks. These digital editions will function as models for similar applications. The AAC's scholarly editions of "Der Brenner" and "Die Fackel" will contribute to the development of digital resources for research into language and literature.

References

AAC - Austrian Academy Corpus: http://www.aac.ac.at/fackel

AAC - Austrian Academy Corpus: AAC-FACKEL, Online Version: Die Fackel. Herausgeber: Karl Kraus, Wien 1899-1936, AAC Digital Edition No 1, 2007, (http://www.aac.ac.at/fackel)

AAC - Austrian Academy Corpus and Brenner-Archiv: BRENNER ONLINE, Online Version: Der Brenner. Herausgeber: Ludwig Ficker, Innsbruck 1910-1954, AAC Digital Edition No 2, 2007, (http://www.aac.ac.at/brenner)

e-Science in the Arts and Humanities – A methodological perspective

Tobias Blanke
tobias.blanke@kcl.ac.uk
King's College London, UK

Stuart Dunn
stuart.dunn@kcl.ac.uk
King's College London, UK

Lorna Hughes
lorna.hughes@kcl.ac.uk
King's College London, UK

Mark Hedges
mark.hedges@kcl.ac.uk
King's College London, UK

The aim of this paper is to provide an overview of e-Science and e-Research activities for the arts and humanities in the UK. It will focus on research projects and trends and will not cover the institutional infrastructure to support them. In particular, we shall explore the methodological discussions laid out in the Arts and Humanities e-Science Theme, jointly organised by the Arts and Humanities e-Science Support Centre and the e-Science Institute in Edinburgh (http://www.nesc.ac.uk/esi/themes/theme_06/). The second focus of the paper will be the current and future activities within the Arts and Humanities e-Science Initiative in the UK and their methodological consequences (http://www.ahessc.ac.uk).

The projects presented so far are all good indicators of what the future might deliver, as 'grand challenges' for the arts and humanities e-Science programme such as the emerging data deluge (Hey and Trefethen 2003). The Bush administration will have produced over 100 million emails by the end of its term (Unsworth 2006). These can provide the basis for new types of historical and socio-political research that will take advantage of computational methods to deal with digital data. However, for arts and humanities research an information is not just an information. Complicated semantics underlie the archives of human reports. As a simple example, it cannot be clear from the email alone which Bush administration or even which Iraq war are under consideration. Moreover, new retrieval methods for such data must be intuitive for the user and not based on complicated metadata schemes. They have to be specific in their return and deliver exactly that piece of information the researcher is interested in. This is fairly straightforward for structured information if it is correctly described, but highly complex for unstructured information. Arts and humanities additionally need the means to on-demand reconfigure the retrieval process by using computational power that changes the set of information items available from texts, images, movies, etc. This paper argues that a specific methodological agenda in arts and humanities e-Science has been developing over the past two years and explores some of its main tenets. We offer a chronological discussion of two phases in the methodological debates about the applicability of e-science and e-research to arts and humanities.

The first phase concerns the methodological discussions that took place during the early activities of the Theme. A series of workshops and presentations about the role of e-Science for arts and humanities purported to make existing e-science methodologies applicable to this new field and consider the challenges that might ensue (http://www.nesc.ac.uk/esi/themes/theme_06/community.htm). Several events brought together computer scientists and arts and humanities researchers. Further events (finished by the time of Digital Humanities 2008) will include training events for postgraduate students and architectural studies on building a national e-infrastructure for the arts and humanities.

Due to space limitations, we cannot cover all the early methodological discussions during the Theme here, but focus on two, which have been fruitful in uptake in arts and humanities: Access Grid and Ontological Engineering. A workshop discussed alternatives for video-conferencing in the arts and humanities in order to establish virtual research communities. The Access Grid has already proven to be of interest in arts and humanities research. This is not surprising, as researchers in these domains often need to collaborate with other researchers around the globe. Arts and humanities research often takes place in highly specialised domains and subdisciplines, niche subjects with expertise spread across universities. The Access Grid can provide a cheaper alternative to face-to-face meetings.

However, online collaboration technologies like the Access Grid need to be better adapted to the specific needs of humanities researchers by e.g. including tools to collaboratively edit and annotate documents. The Access Grid might be a good substitute to some face-to-face meetings, but lacks innovative means of collaboration, which can be especially important in arts and humanities research. We should aim to realise real multicast interaction, as it has been done in VNC technology or basic wiki technology. These could support new models of collaboration in which the physical organisation of the Access Grid suite can be accommodated to specific needs that would e.g. allow participants to walk around. The procedure of Access Grid sessions could also be changed, away from static meetings towards more dynamic collaborations.

Humanities scholars and performers have priorities and concerns that are often different from those of scientists and engineers (Nentwich 2003). With growing size of data resources the need arises to use recent methodological frameworks such as ontologies to increase the semantic interoperability of data. Building ontologies in the humanities is a challenge, which was the topic of the Theme workshop on 'Ontologies and Semantic Interoperability for Humanities Data'. While semantic integration has been a hot topic in business and computing

research, there are few existing examples for ontologies in the Humanities, and they are generally quite limited, lacking the richness that full-blown ontologies promise. The workshop clearly pointed at problems mapping highly contextual data as in the humanities to highly formalized conceptualization and specifications of domains.

The activities within the UK's arts and humanities e-Science community demonstrate the specific needs that have to be addressed to make e-Science work within these disciplines (Blanke and Dunn 2006). The early experimentation phase, which included the Theme events presented supra, delivered projects that were mostly trying out existing approaches in e-Science. They demonstrated the need for a new methodology to meet the requirements of humanities data that is particularly fuzzy and inconsistent, as it is not automatically produced, but is the result of human effort. It is fragile and its presentation often difficult, as e.g. data in performing arts that only exists as an event.

The second phase of arts and humanities e-Science began in September 2007 with seven 3-4 years projects that are moving away from ad hoc experimentation towards a more systematic investigation of methodologies and technologies that could provide answers to grand challenges in arts and humanities research. This second phase could be put in a nutshell as e-science methodology-led innovative research in arts and humanity.

Next to performance, music research e.g. plays an important vanguard function at adopting e-Science methodologies, mostly because many music resources are already available in digital formats. At Goldsmiths, University of London, the project 'Purcell Plus' e.g. will build upon the successful collaboration 'Online Musical Recognition and Searching (OMRAS)' (http://www.omras.org/), which has just achieved a second phase of funding by the EPSRC. With OMRAS, it will be possible to efficiently search large-scale distributed digital music collections for related passages, etc. The project uses grid technologies to index the very large distributed music resources. 'Purcell Plus' will make use of the latest explosion in digital data for music research. It uses Purcell's autograph MS of 'Fantazies and In Nomines for instrumental ensemble' and will investigate the methodology problems for using toolkits like OMRAS for musicology research. 'Purcell Plus' will adopt the new technologies emerging from music information retrieval, without the demand to change completely proven to be good methodologies in musicology. The aim is to suggest that new technologies can help existing research and open new research domains in terms of the quantity of music and new quantitative methods of evaluation.

Building on the earlier investigations into the data deluge and how to deal with it, many of the second-phase projects look into the so-called 'knowledge technologies' that help with data and text mining as well as simulations in decision support for arts and humanities research. One example is the 'Medieval Warfare on the Grid: The Case of Manzikert' project in Birmingham, which will investigate the need for medieval states to sustain armies by organising and distributing resources. A grid-based framework shall virtually reenact the Battle of Manzikert in 1071, a key historic event in Byzantine history. Agent-based modelling technologies will attempt to find out more about the reasons why the Byzantine army was so heavily defeated by the Seljurk Turks. Grid environments offer the chance to solve such complex human problems through distributed simultaneous computing.

In all the new projects, we can identify a clear trend towards investigating new methodologies for arts and humanities research, possible only because grid technologies offer unknown data and computational resources. We could see how e-Science in the arts and humanities has matured towards the development of concrete tools that systematically investigate the use of e-Science for research. Whether it is simulation of past battles or musicology using state-of-the-art information retrieval techniques, this research would have not been possible before the shift in methodology towards e-Science and e-Research.

References

Blanke, T. and S. Dunn (2006). The Arts and Humanities e-Science Initiative in the UK. *E-Science '06: Proceedings of the Second IEEE International Conference on e-Science and Grid Computing*, Amsterdam, IEEE Computer Society.

Hey, T. and A. Trefethen (2003). The data deluge: an e-Science perspective. In F. Berman, A. Hey and G. Fox (eds) *Grid Computing: Making the Global Infrastructure a Reality*. Hoboken, NJ, John Wiley & Sons.

Nentwich, M. (2003). *Cyberscience. Research in the Age of the Internet*. Vienna, Austrian Academy of Science Press.

Unsworth, J. (2006). "The Draft Report of the American Council of Learned Societies' Commission on Cyberinfrastructure for Humanities and Social Sciences." from http://www.acls.org/cyberinfrastructure/.

An OWL-based index of emblem metaphors

Peter Boot
peter.boot@huygensinstituut.knaw.nl
Huygens Institute

Intro

This paper describes an index on metaphor in Otto Vaenius' emblem book *Amoris divini emblemata* (Antwerp, 1615). The index should be interesting both for its contents (that is, for the information about the use of metaphor in the book) and as an example of modelling a complex literary phenomenon. Modelling a complex phenomenon creates the possibility to formulate complex queries on the descriptions that are based on the model. The article describes an application that uses this possibility. The application user can interrogate the metaphor data in multiple ways, ranging from canned queries to complex selections built in the application's guided query interface.

Unlike other emblem indices, the metaphor index is not meant to be a tool for resource discovery, a tool that helps emblem scholars find emblems relevant to their own research. It presents research output rather than input. The modelling techniques that it exemplifies should help a researcher formulate detailed observations or findings about his research subject – in this case, metaphor – and make these findings amenable to further processing. The result is an index, embedded in an overview or explanation of the data for the reader. I will argue that for research output data it is up to the researcher who uses these modelling techniques to integrate the presentation of data in a narrative or argument, and I describe one possible way of effecting this integration.

The paper builds on the techniques developed in (Boot 2006). The emblem book is encoded using TEI; a model of metaphor (an ontology) is formulated in OWL; the observations about the occurrence of metaphors are stored as RDF statements. An essay about the more important metaphors in this book is encoded in TEI. This creates a complex and interlinked structure that may be explored in a number of ways. The essay is hyperlinked to (1) individual emblems, (2) the presentation of individual metaphors in emblems, (3) searches in the metaphor data, and (4) concepts in the metaphor ontology. From each of these locations, further exploration is possible. Besides these ready-made queries, the application also facilitates user-defined queries on the metaphor data. The queries are formulated using the SPARQL RDF query language, but the application's guided query interface hides the actual syntax from the user.

Metaphor model

There is a number of aspects of metaphor and the texts where metaphors occur that are modelled in the metaphor index. A metaphor has a vehicle and a tenor, in the terminology of Richards (1936). When love, for its strength en endurance in adversity, is compared to a tree, the tree is the vehicle, love is the tenor. It is possible to define hierarchies, both for the comparands (that is, vehicles and tenors) and for the metaphors: we can state that 'love as a tree' (love being firmly rooted) belongs to a wider class of 'love as a plant' (love bearing fruit) metaphors. We can also state that a tree is a plant, and that it (with roots, fruit, leaves and seeds) belongs to the vegetal kingdom (Lakoff and Johnson 1980). It often happens that an emblem contains references to an object invested with metaphorical meaning elsewhere in the book. The index can record these references without necessarily indicating something they are supposed to stand for.

The index can also represent the locations in the emblem (the text and image fragments) that refer to the vehicles and tenors. The text fragments are stretches of emblem text, the image fragments are rectangular regions in the emblem pictures. The index uses the TEI-encoded text structure in order to relate occurrences of the comparands to locations in the text.

The metaphor model is formulated using the Web Ontology Language OWL (McGuinness and Van Harmelen 2004). An ontology models the kind of objects that exist in a domain, their relationships and their properties; it provides a shared understanding of a domain. On a technical level, the ontology defines the vocabulary to be used in the RDF statements in our model. The ontology thus limits the things one can say; it provides, in McCarty's words (McCarty 2005), the 'explicit, delimited conception of the world' that makes meaningful manipulation possible. The ontology is also what 'drives' the application built for consultation of the metaphor index. See for similar uses of OWL: (Ciula and Vieira 2007), (Zöllner-Weber 2005).

The paper describes the classes and the relationships between them that the OWL model contains. Some of these relationships are hierarchical ('trees belong to the vegetal kingdom'), others represent relations between objects ('emblem 6 uses the metaphor of life as a journey' or 'metaphor 123 is a metaphor for justice'). The relationships are what makes it possible to query objects by their relations to other objects: to ask for all the metaphors based in an emblem picture, to ask for all of the metaphors for love, or to combine these criteria.

Application

In order to present the metaphor index to a reader, a web application has been developed that allows readers to consult and explore the index. The application is an example of an ontology-driven application as discussed in (Guarino 1998): the data model, the application logic and the user interface are all based on the metaphor ontology.

The application was created using PHP and a MySQL database backend. RAP, the RDF API for PHP, is used for handling RDF. RDF and OWL files that contain the ontology and occurrences

are stored in an RDF model in the database. RDF triples that represent the structure of the emblem book are created from the TEI XML file that contains the digital text of the emblem book.

The application has to provide insight into three basic layers of information: our primary text (the emblems), the database-like collection of metaphor data, and a secondary text that should make these three layers into a coherent whole. The application organizes this in three perspectives: an overview perspective, an emblem perspective and an ontology perspective. Each of these perspectives offers one or more views on the data. These views are (1) a basic selection interface into the metaphor index; (2) an essay about the use and meaning of metaphor in this book; (3) a single emblem display; (4) information about metaphor use in the emblem; and (5) a display of the ontology defined for the metaphor index (built using the OWLDoc). The paper will discuss the ways in which the user can explore the metaphor data.

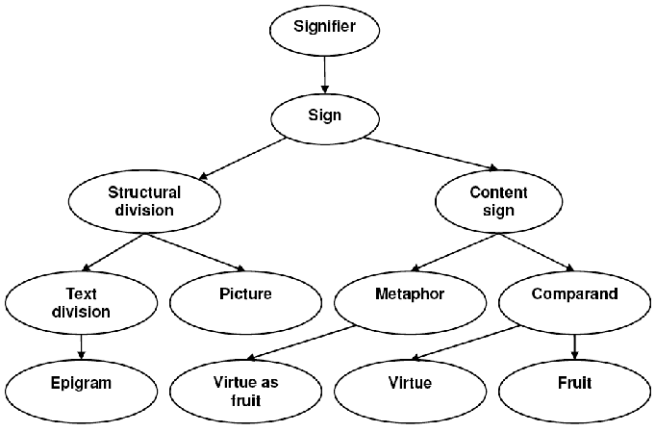


Figure 1 Part of the classes that make up the metaphor ontology. Arrows point to subclasses. The classes at the bottom level are just examples; many other could have been shown if more space were available. For simplicity, this diagram ignores class properties

Discussion

The metaphor index is experimental, among other things in its modelling of metaphor and in its use of OWL and RDF in a humanities context. If Willard McCarty is right in some respects all humanities computing is experimental. There is, however, a certain tension between the experimental nature of this index and the need to collect a body of material and create a display application. If the aim is not to support resource discovery, but solely to provide insight, do we then need this large amount of data? Is all software meant to be discarded, as McCarty quotes Perlis? The need to introduce another aspect of metaphor into the model may conflict with the need to create a body of material that it is worthwhile to explore. It is also true, however, that insight doesn't come from subtlety alone. There is no insight without numbers.

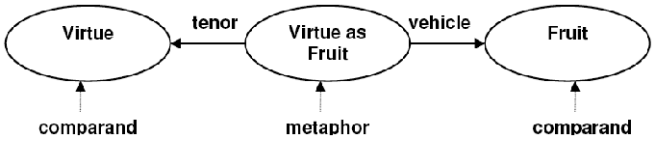


Figure 2 A metaphor and the properties relating it to the comparands

Figure 3 Objects can be queried by their relations

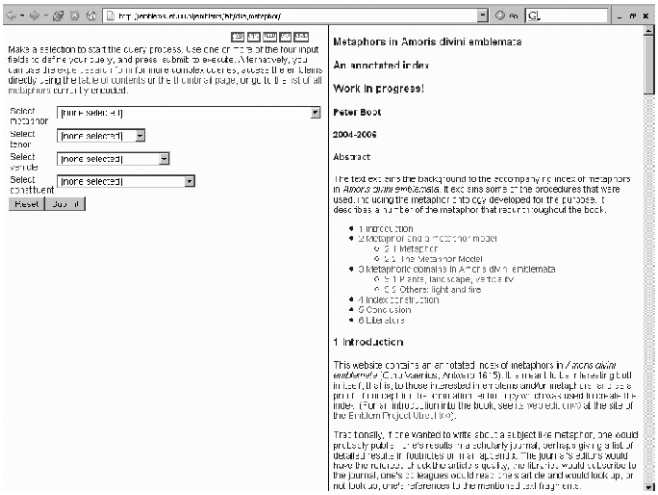


Figure 4 Overview perspective

McCarty writes about the computer as 'a rigorously disciplined means of implementing trial-and-error (…) to help the scholar refine an inevitable mismatch between a representation and reality (as he or she conceives it) to the point at which the epistemological yield of the representation has been realized'. It is true that the computer helps us be rigorous and disciplined, but perhaps for that very reason the representations that the computer helps us build may become a burden. Computing can slow us down. To clarify the conceptual structure of metaphor as it is used in the book we do not necessarily need a work of reference. The paper's concluding paragraphs will address this tension.

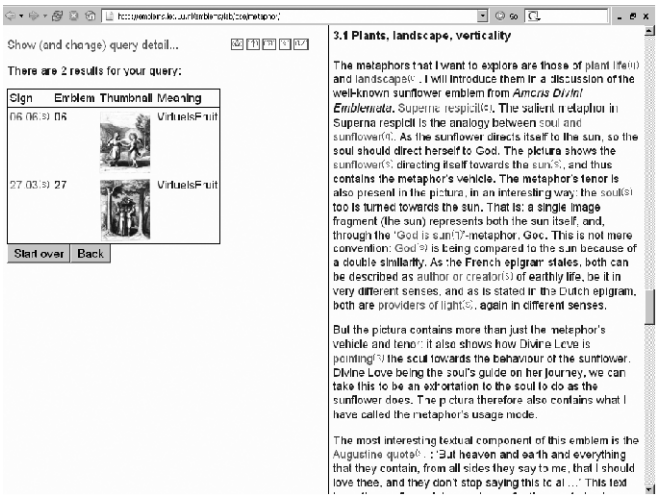


Figure 5 Clicking the hyperlink 'plant life' (top right) executes a query with hits shown in the left panel

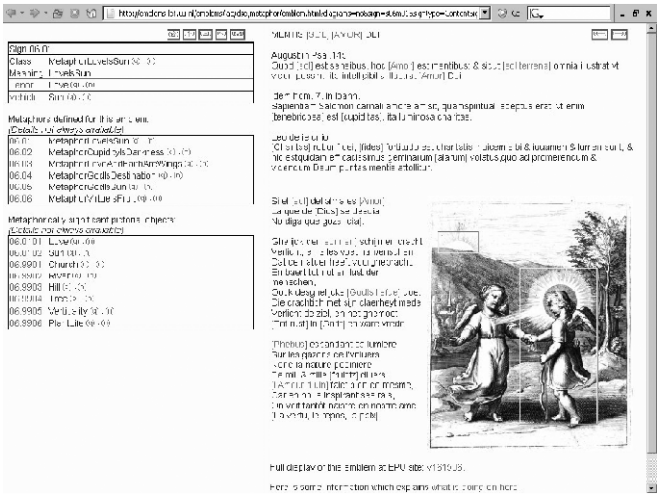


Figure 6 Emblem perspective with one metaphor highlighted in picture and text

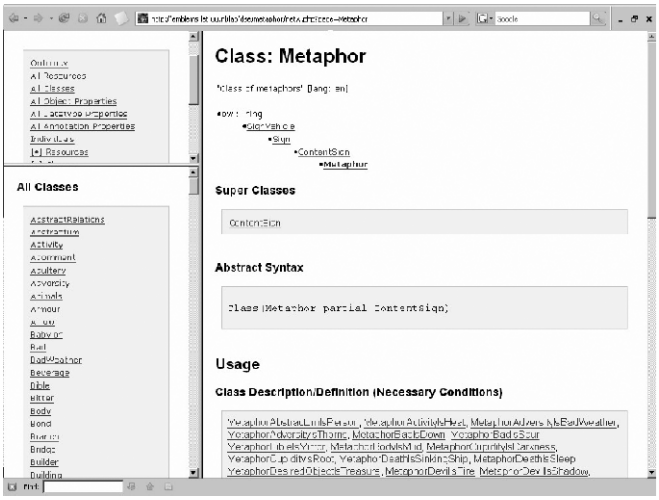


Figure 7 Ontology perspective , with display of class metaphor

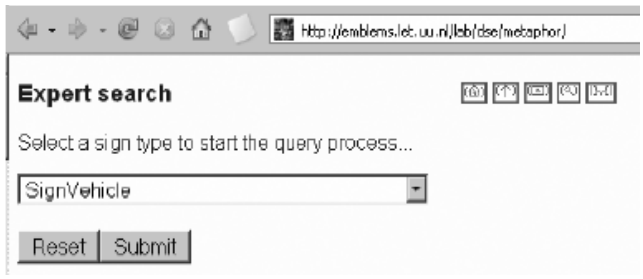


Figure 8 Expert search, start building query

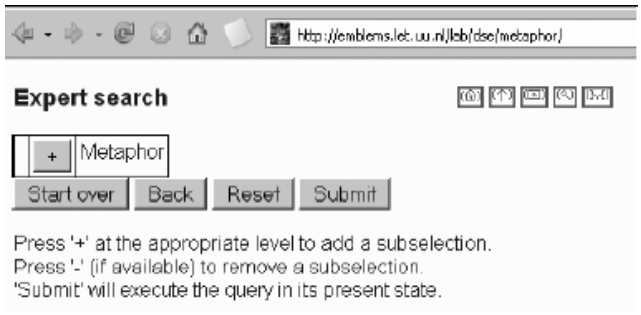


Figure 9 Expert search. Click '+' to create more criteria

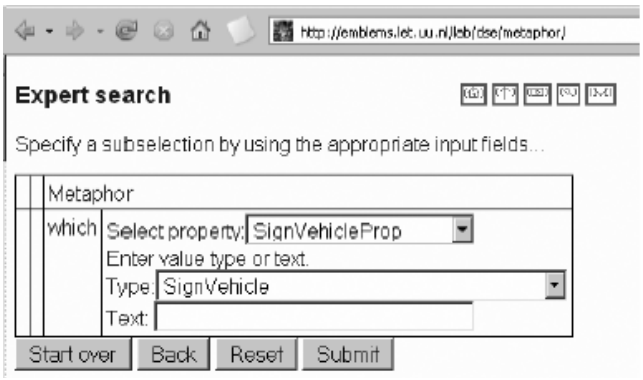


Figure 10 Expert search. Select the desired criterion

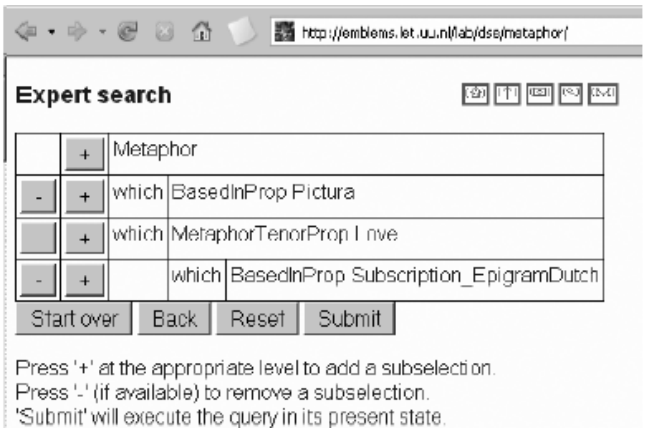


Figure 11 Expert search. Final state of query

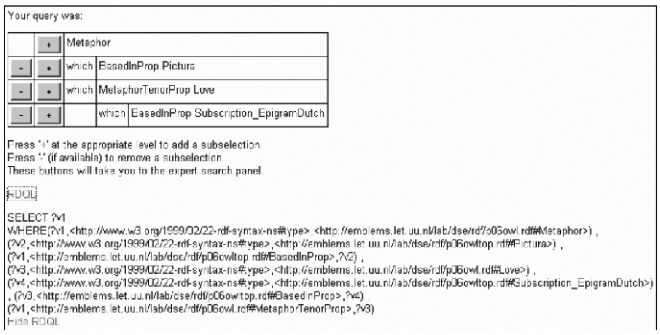


Figure 12 Expert search. Display of executed query and generated RDQL in results panel

References

Antoniou, Grigoris and Van Harmelen, Frank (2004), *A Semantic Web Primer* (Cooperative Information Systems; Cambridge (Ma); London: MIT Press).

Boot, Peter (2006), 'Decoding emblem semantics', *Literary and Linguistic Computing*, 21 supplement 1, 15-27.

Ciula, Arianna and Vieira, José Miguel (2007), 'Implementing an RDF/OWL Ontology on Henry the III Fine Rolls', paper given at OWLED 2007, Innsbruck.

Guarino, Nicola (1998), 'Formal Ontology and Information Systems', in Nicola Guarino (ed.), *Formal Ontology in Information Systems*. Proceedings of FOIS'98, Trento, Italy, 6-8 June 1998 (Amsterdam: IOS Press), 3-15.

Lakoff, George and Johnson, Mark (1980), *Metaphors we live by* (Chicago; London: University of Chicago Press).

McCarty, Willard (2005), *Humanities Computing* (Basingstoke: Palgrave Macmillan).

McGuinness, Deborah L. and Van Harmelen, Frank (2007), 'OWL Web Ontology Language. Overview. W3C Recommendation 10 February 2004', <http://www.w3.org/TR/owl-features/>, accessed 2007-02-24.

Richards, Ivor Armstrong (1936), *The Philosophy of Rhetoric* (New York, London: Oxford University Press).

Zöllner-Weber, Amelie (2005), 'Formale Repräsentation und Beschreibung von literarischen Figuren', *Jahrbuch für Computerphilologie – online*, 7.

Collaborative tool-building with Pliny: a progress report

John Bradley
john.bradley@kcl.ac.uk
King's College London, UK,

In the early days the Digital Humanities (DH) focused on the development of tools to support the individual scholar to perform original scholarship, and tools such as OCP and TACT emerged that were aimed at the individual scholar. Very little tool-building within the DH community is now aimed generally at individual scholarship. There are, I think, two reasons for this:

- First, the advent of the Graphical User Interface (GUI) made tool building (in terms of software applications that ran on the scholar's own machine) very expensive. Indeed, until recently, the technical demands it put upon developers have been beyond the resources of most tool developers within the Humanities.

- Second, the advent of the WWW has shifted the focus of much of the DH community to the web. However, as a result, tool building has mostly focused on not the *doing* of scholarly research but on the *publishing* of resources that represent the result of this.

DH's tools to support the publishing of, say, primary sources, are of course highly useful to the researcher when his/her primary research interest is the *preparation* of a digital edition. They are not directly useful to the researcher *using* digital resources. The problem (discussed in detail in Bradley 2005) is that a significant amount of the potential of digital materials to support individual research is lost in the representation in the browser, even when based on AJAX or Web 2.0 practices.

The *Pliny* project (Pliny 2006-7) attempts to draw our attention as tool builders back to the user of digital resources rather than their creator, and is built on the assumption that the *software application*, and not the browser, is perhaps the best platform to give the user full benefit of a digital resource. Pliny is not the only project that recognises this. The remarkable project Zotero (Zotero 2006-7) has developed an entire plugin to provide a substantial new set of functions that the user can do within their browser. Other tool builders have also recognised that the browser restricts the kind of interaction with their data too severely and have developed software applications that are not based on the web browser (e.g. Xaira (2005-7), WordHoard (2006-7), Juxta (2007), VLMA (2005-7)). Some of these also interact with the Internet, but they do it in ways outside of conventional browser capabilities.

Further to the issue of tool building is the wish within the DH community to create tools that work well together. This problem has often been described as one of modularity – building separate components that, when put together,

allow the user to combine them to accomplish a range of things perhaps beyond the capability of each tool separately. Furthermore, our community has a particularly powerful example of modularity in Wilhelm Ott's splendid *TuStep* software (Ott 2000). TuStep is a toolkit containing a number of separate small programs that each perform a rather abstract function, but can be assembled in many ways to perform a very large number of very different text processing tasks. However, although TuStep is a substantial example of software designed as a toolkit, the main discussion of modularity in the DH (going back as far as the CETH meetings in the 1990s) has been in terms of *collaboration* – finding ways to support the development of tools by different developers that, in fact, can co-operate. This is a very different issue from the one TuStep models for us. There is as much or more design work employed to create TuStep's framework in which the separate abstract components operate (the overall system) as there is in the design of each component itself. This approach simply does not apply when different groups are designing tools semi-independently. What is really wanted is a world where software tools such as WordHoard can be designed in ways that allow other tools (such as Juxta) to interact in a GUI, on the screen.

Why is this so difficult? Part of the problem is that traditional software development focuses on a "stack" approach. Layers of ever-more specific software are built on top of more-general layers to create a specific application, and each layer in the stack is aimed more precisely at the ultimate functions the application was meant to provide. In the end each application runs in a separate window on the user's screen and is focused specifically and exclusively on the functions the software was meant to do. Although software could be written to support interaction between different applications, it is in practice still rarely considered, and is difficult to achieve.

Pliny, then, is about two issues:

- First, Pliny focuses on digital annotation and note-taking in humanities scholarship, and shows how they can be used facilitate the development of an interpretation. This has been described in previous papers and is not presented here.

- Second, Pliny models how one could be building GUI-oriented software applications that, although developed separately, support a richer set of interactions and integration on the screen.

This presentation focuses primarily on this second theme, and is a continuation of the issues raised at last year's poster session on this subject for the DH2007 conference (Bradley 2007). It arises from a consideration of Pliny's first issue since note-taking is by its very nature an integrative activity – bringing together materials created in the context of a large range of resources and kinds of resources.

Instead of the "stack" model of software design, Pliny is constructed on top of the Eclipse framework (Eclipse 2005-7), and uses its contribution model based on Eclipse's *plugin* approach (see a description of it in Birsan 2005). This approach promotes effective collaborative, yet independent, tool building and makes possible many different kinds of interaction between separately written applications. Annotation provides an excellent motivation for this. A user may wish to annotate something in, say, WordHoard. Later, this annotation will need to be shown with annotations attached to other objects from other pieces of software. If the traditional "stack" approach to software is applied, each application would build their own annotation component inside their software, and the user would not be able to bring notes from different tools together. Instead of writing separate little annotation components inside each application, Eclipse allows objects from one application to participate as "first-class" objects in the operation of another. Annotations belong simultaneously to the application in which they were created, and to Pliny's annotation-note-taking management system.

Pliny's plugins both support the manipulation of annotations while simultaneously allowing other (properly constructed) applications to create and display annotations that Pliny manages for them. Furthermore, Pliny is able to recognise and represent references to materials in other applications within its own displays. See Figures I and II for examples of this, in conjunction with the prototype VLMA (2005-7) plugin I created from the standalone application produced by the VLMA development team. In Figure I most of the screen is managed by the VLMA application, but Pliny annotations have been introduced and combined with VLMA materials. Similarly, in figure II, most of the screen is managed by Pliny and its various annotation tools, but I have labelled areas on the screen where aspects of the VLMA application still show through.

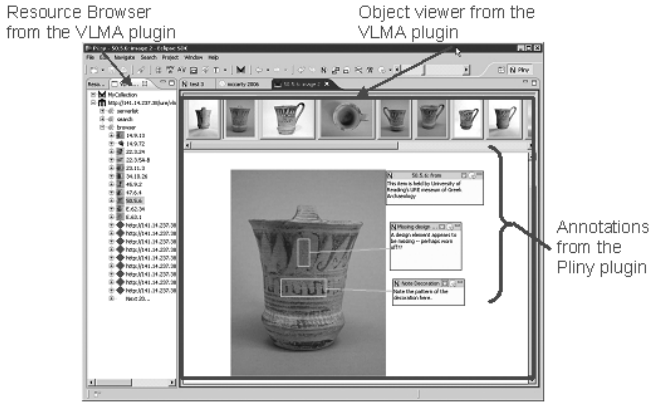


Figure I: Pliny annotations in a VLMA viewer

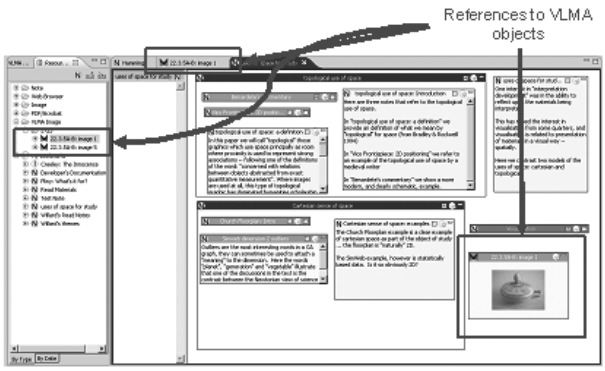


Figure II:VLMA objects in a Pliny context

This connecting of annotation to a digital object rather than merely to its display presents some new issues. What, for example, does it mean to link an annotation to a line of a KWIC display – should that annotation appear when the same KWIC display line appears in a different context generated as the result of a different query? Should it appear attached to the particular word token when the document it contains is displayed? If an annotation is attached to a headword, should it be displayed automatically in a different context when its word occurrences are displayed, or only in the context in which the headword itself is displayed? These are the kind of questions of annotation and context that can only really be explored in an integrated environment such as the one described here, and some of the discussion in this presentation will come from prototypes built to work with the RDF data application VLMA, with the beginnings of a TACT-like text analysis tool, and on a tool based on Google maps that allows one to annotate a map.

Building our tools in contexts such as Pliny's that allow for a complex interaction between components results in a much richer, and more appropriate, experience for our digital user. For the first time, perhaps, s/he will be able to experience the kind of interaction between the materials that are made available through applications that expose, rather than hide, the true potential of digital objects. Pliny provides a framework in which objects from different tools are brought into close proximity and connected by the paradigm of annotation. Perhaps there are also paradigms other than annotation that are equally interesting for object linking?

References

Birsan, Dorian (2005). "On Plug-ins and Extensible Architectures", In *Queue* (ACM), Vol 3 No 2.

Bradley, John (2005). "What you (fore)see is what you get: Thinking about usage paradigms for computer assisted text analysis" in *Text Technology* Vol. 14 No 2. pp 1-19. Online at http://texttechnology.mcmaster.ca/pdf/vol14_2/bradley14-2.pdf (Accessed November 2007).

Bradley, John (2007). "Making a contribution: modularity, integration and collaboration between tools in Pliny". In book of abstracts for the *DH2007 conference*, Urbana-Champaign, Illinois, June 2007. Online copy available at http://www.digitalhumanities.org/dh2007/abstracts/xhtml.xq?id=143 (Accessed October 2007).

Eclipse 2005-7. *Eclipse Project Website*. At http://www.eclipse.org/ (Accessed October 2007).

Juxta (2007). Project web page at http://www.nines.org/tools/juxta.html (accessed November 2007).

Ott, Wilhelm (2000). "Strategies and Tools for Textual Scholarship: the Tübingen System of Text Processing Programs (TUSTEP)" in *Literary & Linguistic Computing,* 15:1 pp. 93-108.

Pliny 2006-7. *Pliny Project Website* . At http://pliny.cch.kcl.ac.uk (Accessed October 2007).

VLMA (2005-7). *VLMA: A Virtual Lightbox for Museums and Archives*. Website at http://lkws1.rdg.ac.uk/vlma/ (Accessed October 2007).

WordHoard (2006-7). *WordHoard: An application for close reading and scholarly analysis of deeply tagged texts*. Website at http://wordhoard.northwestern.edu/userman/index.html (Accessed November 2007).

Xaira (2005-7). *Xaira Page*. Website at http://www.oucs.ox.ac.uk/rts/xaira/ (Accessed October 2007).

Zotero (2006-7). *Zotero: A personal research assistant*. Website at http://www.zotero.org/ (Accessed September 2007).

How to find Mrs. Billington? Approximate string matching applied to misspelled names

Gerhard Brey
gerhard.brey@kcl.ac.uk
King's College London, UK

Manolis Christodoulakis
m.christodoulakis@uel.ac.uk
University of East London, UK

In this paper we discuss some of the problems that arise when searching for misspelled names. We suggest a solution to overcome these and to disambiguate the names found.

Introduction

The Nineteenth-Century Serials Edition (NCSE) is a digital edition of six nineteenth-century newspaper and periodical titles. It is a collaborative project between Birkbeck College, King's College London, the British Library and Olive Software Inc. funded by the UK's Arts and Humanities Research Council. The corpus consists of about 100,000 pages that were micro-filmed, scanned in and processed using optical character recognition software (OCR) to obtain images for display in the web application as well as the full text of the publications. In the course of this processing (by Olive Software) the text of each individual issue was also automatically segmented into its constituent parts (newspaper departments, articles, advertisements, etc.).

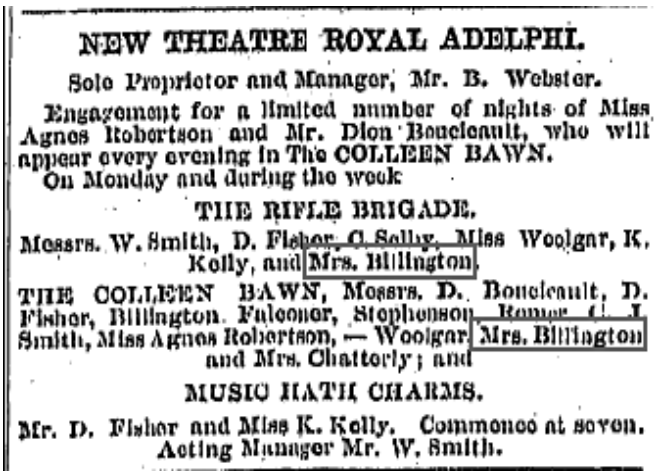
The application of text mining techniques (named entity extraction, text categorisation, etc.) allowed names, institutions and places etc. to be extracted as well as individual articles to be classified according to events, subjects and genres. Users of the digital edition can search for very specific information. The extent to which these searches are successful depends, of course, on the quality of the available data.

The problem

The quality of some of the text resulting from the OCR process varies from barely readable to illegible. This reflects the poor print quality of the original paper copies of the publications. A simple search of the scanned and processed text for a person's name, for example, would retrieve exact matches, but ignore incorrectly spelled or distorted variations.

Misspellings of ordinary text could be checked against an authoritative electronic dictionary. An equivalent reference work for names does not exist. This paper describes the solutions that are being investigated to overcome these difficulties.

This theatrical notice on page 938 of the *Leader* from 17.11.1860 highlights the limitations of OCR alone.



NEW THEATRE ROYAL ADELPHI.

Sole Proprietor and Manager, Mr. B. Webster.

Engagement for a limited number of nights of Miss Agnes Robertson and Mr. Dion Boucicault, who will appear every evening in The COLLEEN BAWN.

On Monday and during the week

THE RIFLE BRIGADE.

Messrs. W. Smith, D. Fisher, C. Selby, Miss Woolgar, K. Kelly, and Mrs. Billington.

THE COLLEEN BAWN, Messrs. D. Boucicault, D. Fisher, Billington, Falconer, Stephenson, Romer, J. Smith, Miss Agnes Robertson, — Woolgar, Mrs. Billington and Mrs. Chatterly; and

MUSIC HATH CHARMS.

Mr. D. Fisher and Miss K. Kelly. Commence at seven.

Acting Manager Mr. W. Smith.

The actress Mrs. Billington is mentioned twice. OCR recognised the name once as `Mrs. lUllinijton` and then as `Mrs. BIIMngton`. A simple search for the name Billington would therefore be unsuccessful.

By applying a combination of approximate string matching techniques and allowing for a certain amount of spelling errors (see below) our more refined approach successfully finds the two distorted spellings as `Mrs. Billington`. However, it also finds a number of other unrelated names (`Wellington` and `Rivington` among others). This additional problem is redressed by mapping misspelled names to correctly spelled names. Using a combination of string similarity and string distance algorithms we have developed an application to rank misspelled names according to their likelihood of representing a correctly spelled name.

The algorithm

As already briefly mentioned above we are applying several well known pattern matching algorithms to perform approximate pattern matching, where the pattern is a given name (a surname normally), and the "text" is a (huge) list of names, obtained from the OCR of scanned documents. The novelty of this work comes from the fact that we are utilizing application-specific characteristics to achieve better results than are possible through general-purpose pattern matching techniques.

Currently we are considering the pattern to be error-free (although our algorithms can easily be extended to deal with errors in the pattern too). Moreover, all the algorithms take as input the maximum "distance" that a name in the list may have from the pattern to be considered a match; this distance is given as a percentage. As one would expect, there is a tradeoff in distance - quality of matches: low distance threshold yields less false positives, but may miss some true matches; on the other hand, a high distance threshold has less chances of missing true matches, but will return many fake ones.

At the heart of the algorithms lies a ranking system, the purpose of which is to sort the matching names according to how well they match. (Recall that the list of matching names can be huge, and thus what is more important is really the ranking of those matches, to distinguish the good ones from random matches.) Obviously, the ranking system depends on the distance-metric in use, but there is more to be taken into account. Notice that, if a name matches our pattern with some error, *e*, there are two cases:

- either the name is a true match, and the error e is due to bad OCR, or

- the name (misspelled or not, by OCR) does not correspond to our pattern, in which case it is a bad match and should receive a lower rank.

It is therefore essential, given a possibly misspelled (due to OCR) name, *s'*, to identify the true name, *s*, that it corresponds to. Then, it is obvious that *s'* is a good match if $p = s$, and a bad match otherwise, where *p* is the pattern. To identify whether a name *s'* in our list is itself a true name, or has been mispelled we use two types of evaluation:

- We count the occurrences of *s'* in the list. A name that occurs many times, is likely to be a true name; if it had been misspelled, it is very unlikely that all its occurrences would have been misspelled in exactly the same manner, by the OCR software.

- We have compiled a list *L* of valid names; these names are then used to decide whether *s'* is a valid name ($s' \in L$) or not ($s' \notin L$).

In the case where *s'* is indeed mispelled by OCR, and is thus not a true name, one must use distance metrics to identify how closely it matches the pattern *p*. Given that the pattern is considered to be error-free, if the distance of the name from the pattern is large then it is very unlikely (but not impossible) that too many of the symbols of the name have been mispelled by OCR; instead, most probably the name does not really match the pattern.

Taking into account the nature of the errors that occur in our application, when computing the distance of a name in the list from our pattern, we consider optical similarities. That is, we drop the common tactic where one symbol is compared against another symbol, and either they match - so the distance is 0, or they don't - so the distance is 1; instead, we consider a symbol (or a group of symbols) to have a low distance from another symbol (or group of symbols) if their shapes look similar. As an example, check that "m" is optically very similar to "rn", and thus should be assigned a small distance, say 1, while "m" and "b" do not look similar to each other and therefore should have a big distance, say 10.

The results of our efforts to date have been very promising. We look forward to investigating opportunities to improve the effectiveness of the algorithm in the future.

Bibliography

NCSE website: http://www.ncse.kcl.ac.uk/index.html

Cavnar, William B. and John M. Trenkle. "N-Gram-Based Text Categorization", in: *Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval,* Las Vegas 1994, 161-175; http://citeseer.ist.psu.edu/68861.html; accessed Nov 2007.

Cavnar, William B. "Using An N-Gram-Based Document Representation With A Vector Processing Retrieval Model", in: *TREC*, 1994, 269--277; http://trec.nist.gov/pubs/trec3/papers/cavnar_ngram_94.ps.gz; accessed Nov 2007.

Gusfield, Dan. *Algorithms on strings, trees, and sequences: computer science and computational biology*, Cambridge (CUP) 1997.

Hall, Patrick A.V. and Geoff R. Dowling. "Approximate String Matching", in: *ACM Computing Surveys*, 12.4 (1980), 381--402; http://doi.acm.org/10.1145/356827.356830; accessed November 2007

Jurafsky, Daniel Saul and James H. Martin. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*, Upper Saddle River, NJ (Prentice Hall) 2000, Chapter 5.

Lapedriza, Àgata and Jordi Vitrià. "Open N-Grams and Discriminant Features in Text World: An Empirical Study"; http://www.cvc.uab.es/~jordi/AgataCCIA2004.pdf; accessed November 2007.

Navarro, Gonzalo. "A guided tour to approximate string matching", in: *ACM Computing Surveys*, 33.1 (2001), 31--88; http://doi.acm.org/10.1145/375360.375365; accessed November 2007

Navarro, Gonzalo and Mathieu Raffinot. *Flexible pattern matching in strings: practical on-line search algorithms for texts and biological sequences*, Cambridge (CUP) 2002.

Oakes, Michael P. *Statistics for corpus linguistics*, Edinburgh (Edinburgh Univ. Press) 1998, (Edinburgh textbooks in empirical linguistics, XVI), Chapter 3.4

Degrees of Connection: the close interlinkages of Orlando

Susan Brown
sbrown@uoguelph.ca
University of Guelph, Canada

Patricia Clements
patricia.clements@ualberta.ca
University of Alberta, Canada

Isobel Grundy
Isobel.Grundy@UAlberta.ca
University of Alberta, Canada

Stan Ruecker
sruecker@ualberta.ca
University of Alberta, Canada

Jeffery Antoniuk
jefferya@ualberta.ca
University of Alberta, Canada

Sharon Balazs
Sharon.Balazs@ualberta.ca
University of Alberta, Canada

Orlando: Women's Writing in the British Isles from the Beginnings to the Present is a literary-historical textbase comprising more than 1,200 core entries on the lives and writing careers of British women writers, male writers, and international women writers; 13,000+ free-standing chronology entries providing context; 12,000+ bibliographical listings; and more than 2 million tags embedded in 6.5 million words of born-digital text. The XML tagset allows users to interrogate everything from writers' relationships with publishers to involvement in political activities or their narrative techniques.

The current interface allows users to access entries by name or via various criteria associated with authors; to create chronologies by searching on tags and/or contents of dated materials; and to search the textbase for tags, attributes, and text, or a combination. The XML serves primarily to structure the materials; to allow users to draw on particular tags to bring together results sets (of one or more paragraphs incorporating the actual 'hit') according to particular interests; and to provide automatic hyperlinking of names, places, organizations, and titles.

Recognizing both that many in our target user community of literary students and scholars dislike tag searches, and that our current interface has not fully plumbed the potential of *Orlando*'s experimentation in structured text, we are exploring what other kinds of enquiry and interfaces the textbase can support. We report here on some investigations into new ways of probing and representing the links created by the markup.

The current interface leverages the markup to provide contexts for hyperlinks. Each author entry includes a "Links" screen that provides hyperlinks to mentions of that author elsewhere in the textbase. These links are sorted into groups based on the semantic tags of which they are children, so that users can choose, for instance, from the more than 300 links on the George Eliot Links screen, between a link to Eliot in the Elizabeth Stuart Phelps entry that occurs in the context of Family or Intimate Relationships, and a link to Eliot in Simone de Beauvoir's entry that occurs in the discussion of features of de Beauvoir's writing. Contextualizing Links screens are provided not only for writers who have entries, but also for any person who is mentioned more than once in the textbase, and also for titles of texts, organizations, and places. It thus provides a unique means for users to pursue links in a less directed and more informed way than that provided by many interfaces.

Building on this work, we have been investigating how *Orlando* might support queries into relationships and networking, and present not just a single relationship but the results of investigating an entire field of interwoven relationships of the kind that strongly interest literary historians. Rather than beginning from a known set of networks or interconnections, how might we exploit our markup to analyze interconnections, reveal new links, or determine the points of densest interrelationship? Interface design in particular, if we start to think about visualizing relationships rather than delivering them entirely in textual form, poses considerable challenges.

We started with the question of the degrees of separation between different people mentioned in disparate contexts within the textbase. Our hyperlinking tags allow us to conceptualize links between people not only in terms of direct contact, that is person-to-person linkages, but also in terms of linkages through other people, places, organizations, or texts that they have in common. Drawing on graph theory, we use the hyperlinking tags as key indices of linkages. Two hyperlinks coinciding within a single document—an entry or a chronology event--were treated as vertices that form an edge, and an algorithm was used to find the shortest path between source and destination. So, for instance, you can get from Chaucer to Virginia Woolf in a single step in twelve different ways: eleven other writer entries (including Woolf's own) bring their names together, as does the following event:

1 November 1907: The British Museum's reading room reopened after being cleaned and redecorated; the dome was embellished with the names of canonical male writers, beginning with Chaucer and ending with Browning.

> Virginia Woolf's *A Room of One's Own* describes the experience of standing under the dome "as if one were a thought in the huge bald forehead which is so splendidly encircled by a band of famous names." Julia Hege in *Jacob's Room* complains that they did not leave room for an Eliot or a Brontë.

It takes more steps to get from some writers to others: five, for instance, to get from Frances Harper to Ella Baker. But this is very much the exception rather than the rule.

Calculated according to the method described here, we have a vast number of links: the textbase contains 74,204 vertices with an average of 102 edges each (some, such as London at 101,936, have considerably more than others), meaning that there are 7.5 million links in a corpus of 6.5 million words. Working just with authors who have entries, we calculated the number of steps between them all, excluding some of the commonest links: the Bible, Shakespeare, England, and London. Nevertheless, the vast majority of authors (on average 88%) were linked by a single step (such as the example of Chaucer and Woolf, in which the link occurs within the same source document) or two steps (in which there is one intermediate document between the source and destination names). Moreover, there is a striking similarity in the distribution of the number of steps required to get from one person to another, regardless of whether one moves via personal names, places, organizations, or titles. 10.6% of entries are directly linked, that is the two authors are mentioned in the same source entry or event. Depending on the hyperlinking tag used, one can get to the destination author with just one intermediate step, or two degrees of separation, in 72.2% to 79.6% of cases. Instances of greater numbers of steps decline sharply, so that there are 5 degrees of separation in only 0.6% of name linkage pages, and none at all for places. Six degrees of separation does not exist in *Orlando* between authors with entries, although there are a few "islands", in the order of from 1.6% to 3.2%, depending on the link involved, of authors who do not link to others.

These results raise a number of questions. As Albert-László Barabási reported of social networks generally, one isn't dealing with degrees of separation so much as degrees of proximity. However, in this case, dealing not with actual social relations but the partial representation in *Orlando* of a network of social relations from the past, what do particular patterns such as these mean? What do the outliers—people such as Ella Baker or Frances Harper who are not densely interlinked with others—and islands signify? They are at least partly related to the brevity of some entries, which can result either from paucity of information, or decisions about depth of treatment, or both. But might they sometimes also represent distance from literary and social establishments? Furthermore, some linkages are more meaningful, in a literary historical sense, than others. For instance, the *Oxford Dictionary of National Biography* is a common link because it is frequently cited by title, not because it indicates a historical link between people. Such incidental links can't be weeded out automatically. So we are investigating the possibility of using the relative number of single- or double-step links between two authors to determine how linked they 'really' are. For instance, Elizabeth Gaskell is connected to William Makepeace Thackeray, Charles Dickens, and George Eliot by 25, 35, and 53 single-step links, respectively, but to Margaret Atwood, Gertrude Stein, and Toni Morrison by 2, 1, and 1. Such contrasts suggest the likely utility of such an approach to distinguishing meaningful from incidental associations.

The biggest question invited by these inquiries into linkages is: how might new modes of inquiry into, or representation of, literary history, emerge from such investigations? One way to address this question is through interfaces. We have developed a prototype web application for querying degrees of separation in *Orlando*, for which we are developing an interface. Relationships or associations are conventionally represented by a network diagram, where the entities are shown as nodes and the relationships as lines connecting the nodes. Depending on the content, these kinds of figures are also referred to as directed graphs, link-node diagrams, entity-relationship (ER) diagrams, and topic maps. Such diagrams scale poorly, since the proliferation of items results in a tangle of intersecting lines. Many layout algorithms position the nodes to reduce the number of crisscrossing lines, resulting in images misleading to people who assume that location is meaningful.

In the case of *Orlando*, two additional complexities must be addressed. First, many inter-linkages are dense: there are often 50 distinct routes between two people. A conventional ER diagram of this data would be too complex to be useful as an interactive tool, unless we can allow the user to simplify the diagram. Second, the *Orlando* data differs from the kind of data that would support "distant reading" (Moretti 1), so our readers will need to access the text that the diagram references. How, then, connect the diagram to a reading view? We will present our preliminary responses to these challenges in an interface for degree of separation queries and results. We are also experimenting with the Mandala browser (Cheypesh et al. 2006) for XML structures as a means of exploring embedded relationships. The current Mandala prototype cannot accommodate the amount of data and number of tags in *Orlando*, so we will present the results of experimenting with a subset of the hyperlinking tags as another means of visualizing the dense network of associations in *Orlando*'s representations of literary history.

References

Barabási, Albert-Lászlo. *Linked: The New Science of Networks*. Cambridge, MA: Perseus Publishing, 2002.

Brown, Susan, Patricia Clements, and Isobel Grundy, ed. *Orlando: Women's Writing in the British Isles from the Beginnings to the Present.* Cambridge: Cambridge Online, 2006.

Cheypesh, Oksana, Constanza Pacher, Sandra Gabriele, Stéfan Sinclair, Drew Paulin and Stan Ruecker. "Centering the mind and calming the heart: mandalas as interfaces." Paper presented at the Society for Digital Humanities (SDH/SEMI) conference. York University, Toronto. May 29-31, 2006.

Mandala Rich Prospect Browser. Dir. Stéfan Sinclair and Stan Ruecker. http://mandala.humviz.org Accessed 22 November 2007.

Moretti, Franco. *Graphs, Maps, Trees: Abstract Models for a Literary Theory.* London: Verso, 2005.

The impact of digital interfaces on virtual gender images

Sandra Buchmüller
sandra.buchmueller@telekom.de
Deutsche Telekom Laboratories, Germany

Gesche Joost
gesche.joost@telekom.de
Deutsche Telekom Laboratories, Germany

Rosan Chow
rosan.chow@telekom.dt
Deutsche Telekom Laboratories, Germany

This paper documents an exploratory research in progress, investigating the relationship between the quality of digital interfaces, their technical conditions and the interfacial mediation of the users' body. Here, we focus on the bodily dimension of gender. For this reason, we analysed two online role playing games with different representation modes (text-based versus image-based) asking which impact the digital interface and their technical conditions have on gender performances.

Following sociological investigations (Bahl, 1998/ Goffman, 2001/ Lübke, 2005/ Müller, 1996), the bodily aspect of gender plays a central role in communication settings due to its social, cultural meaning which nowadays strongly is mediated by information technology. For this reason, we focus on the interfaces. We claim that their representation mode, their design and software constraints have a crucial impact on the virtual patterns of gender referring to individual performance, spatial movement and communication.

This interfacial analysis is just a part of an interdisciplinary inquiry about the interrelation between gender, design and ICT. It is allocated in the overlapping field of sociology, gender studies, design research and information technology.

In this respect, we regard the body and its gender as culturally constructed interface of social interactions and advocate for reflecting it within the process software development and interface design.

Introduction

Today's communication is highly influenced by information technology, which substitutes more and more physical body representations in face-to-face communication. Disembodied experiences have become a part of ordinary life as self performances and interactions are often mediated by designed hardware and software interfaces.

In this context, designers who make ICT accessible via their product and screen designs can be regarded as mediators between technological requirements and user needs. They usually regard interfaces from the point of view of formal-aesthetic (Apple Computer, Inc, 1992; McKey, 1999; Schneiderman/ Plaisant, 2005), or of usability (Krug, 2006; Nielsen/ Loranger, 2006) and interaction-design (Cooper/ Reimann, 2003; Preece/ Rogers/ Sharp, 2002). But designers not only make ICT usable, but also culturally significant. In this respect, they also deal with the cultural constructions and implications of gender which have to be reflected by their screen designs. The interfacial conditions decide about the bodily representations in ICT interaction.

By referring to interactionist (Goffman, 2001), constructivist (Butler, 2006/ Teubner, Wetterer, 1999/ Trettin, 1997/ West, Zimmermann, 1991, 1995) theories and sociological investigations of virtual environments (Eisenrieder, 2003/ Lübke, 2005/ Turkle, 1999), we claim that the bodily aspect of gender is an essential reference point of individual performance, communication, even spatial movements not as a physical property of the body, but due to its cultural meaning and connotative potential. It is supposed to be the most guiding information referring to interaction contexts (Goffman, 2001/ Lübke, 2005). It has a crucial impact on the behavior: Being polite, e.g. opening the door for someone is often a decision made in dependence of the counterpart's gender (Müller, 1996). Not knowing about it causes behavioral confusion (Bahl, 1998/ Lübke, 2005).

Research Perspectives & Research Questions

In contrast to a sociological point of view, a conventional design perspective or technological requirements, we add a completely new aspect to design and technologically driven inquiries in two respects:

- By taking the body as a benchmark for analyzing gender representations of digital interfaces.

- By investigating the virtual body and its gender representations from an interfacial point of view.

Our investigations and reflections are guided by the following questions:

- Which gender images do exist in virtual spaces?

- Which impact do the digital interface and their technical conditions have on gender performances?

Furthermore, we are particularly interested in knowing about the impact of design on preserving traditional and stereotype gender images as well as on modifying or deconstructing them.

Objects of Investigation

Objects of investigations are two online role playing games, so called Multi User Dungeons (MUDs). They are especially suitable for this purpose because they directly refer to bodily representations in form of virtual characters. We choose two MUDs with different representation modes in order to compare the effects of opposite interfaces on the virtual embodiment of gender: LambdaMOO, a popular text-based online role playing game (see Image 1 LM), is contrasted with Second Life, the currently most popular and populated graphical MUD (see Image 1 SL). Examining their interfaces promise to get concrete insights into the interfacial influence on virtual gender images.

Image 1 LM: Interface

Image 1 SL: Interface

Methodology

We use the methods of content analysis and participatory observation – the first in order to explore the interfacial offer of options and tools to create virtual gender representations and the second in order to know, how it feels developing and wearing the respective gendered skin. The analysis and observations are guided and structured using the different dimensions of the body as a matrix of investigating which are empirically generated from the observations themselves. Within these bodily categories, different aspects of gender are identified:

Body presence

- modes of existence or being there

Personality / individuality

- forms of personal performance

- forms of non-verbal communication (facial expressions, gestures, vocal intonations and accentuation)

- modes of emotional expressions

Patterns of gender

- categories or models of gender

Actions and spatial Movements

- modes of behavior and actions

- modes of spatial orientation and navigation

Patterns of social behavior

- modes of communication and interaction

- community standards, behavioral guidelines and rules

Research Results

The main findings show, how interfaces and their specific technical conditions can expand or restrict the performance of gender. Moreover, they demonstrate how the conventional bipolar gender model of western culture can be re- and deconstructed by the interfacial properties and capacities.

The table below gives a summarizing overview about how the different bodily aspects are treated by the respective interface of LambdaMOO or Second Life. Interfacial elements which explicitly address gender aspects are indicated in ***bold italics***. Reference to images are marked "see *Image # LM* (LambdaMOO)/*SL*(Second Life)".

Body aspect	LambdaMOO (text-based)	Second Life (image-based)
Body presence ***(addressing more or less explicitly cultural associations of gender)***	Nickname: ***Renaming is possible at any time*** (See *Image 2.1 LM* and *Image 2.2 LM*)	Nickname: *name is predetermined by the interface (surname is selectable from a list, the availability of the combination of fore and sure name has to be checked by the system); name can't be changed after being registered* (See *Image 2.1 SL* and *Image 2.2 SL*)
		Avatar: *after registration, a character has to be selected of a default set of* ***6 male and female avatars*** (See *Image 3 SL*) *which can be individually modified by the 'Appearance Editor'*
Personality/ individuality	Self-descriptions command	- Appearance Editor: *just offers the* ***gender categories 'male and female'****, but allows to create* ***transgender images*** (See *Image 7.1 SL* and *Image 7.2 SL*) - Profile Window
	Emote-command	Set of gestures: *differentiated into common* ***and male/ female gestures corresponding to the avatar's gender*** (See *Image 9.1 SL, Image 9.2 SL, Image 10.1 SL* and *Image 10.2 SL*)
Actions and spatial movements	Emote-command	Set of gestures: differentiated into common and ***male/ female gestures corresponding to the avatar's gender***
	- Moving-around commands: (cardinal points + up/down) - Look-around commands	- Navigation tool: ***the avatar's movements are programmed corresponding to gender stereotypes: female avatars walk with swinging hips while male avatars walk bowlegged*** (See *Image 8.1 SL*) - Camera-view tool - Sit down / Stand up command: ***the avatar's movements are programmed corresponding to gender stereotypes: female avatars sit close-legged, while male avatars sit bowlegged*** (See *Image 8.2 SL*) - Teleporting
	- Examine-object commands	- Take-object command
Patterns of gender	***10 gender categories: Neuter, male, female, either, Spivak ("To adopt the Spivak gender means to abjure the gendering of the body ,to refuse to be cast as male, female or transsexual."Thomas, 2003)), splat, plural, egoistical, royal, 2nd*** (See *Image 4.1 LM* and *Image 4.2 LM*)	***Male, female: Changing the gender category is possible at any time by the 'Appearance Editor' (****See Image 6 SL****); naked male avatars are sexless(****See Image 5.1 SL****); female avatars have breast but also no genital; masculinity can be emphasized by editing the genital area of the avatar's trousers (****See Image 5.2 SL* and *Image 5.3****) femininity by the size of breast and their position to each other (together or more distant)***
		Set of male / female gestures (See *Image 10.1 SL* and *Image 10.2 SL*): *Female gesture set with 18* ***versus male gesture set with 12 categories, may support the stereotype of females being more emotionally expressive***
Virtual patterns of social behavior	Chat-command: say	Chat-commands: speak / shout ***Some gestures are accompanied by vocal expressions which are differentiated into male and female voices***
	Long-distance-communication command	Instant Messenger
	Behavioral guidelines	Community Standards: 'Big Six'& Policy ***includes the topic 'harassment' which mainly affects females personas***
		Abuse Reporter Tool *Reporting violations directly to the creators of Second Life 'Linden Lab'*

Image 2.1 LM: Rename command

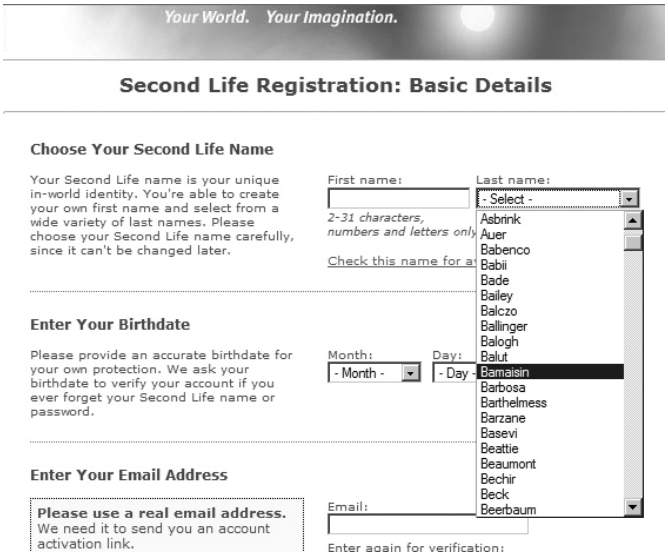


Image 2.1 SL: Choose surname

Image 2.2 LM: Nickname Magenta_Guest

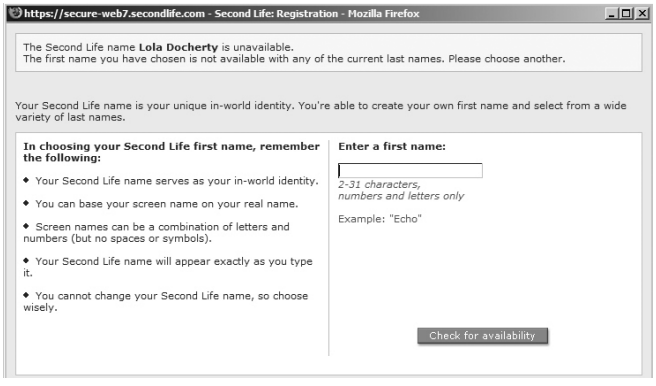


Image 2.2 SL: Name check

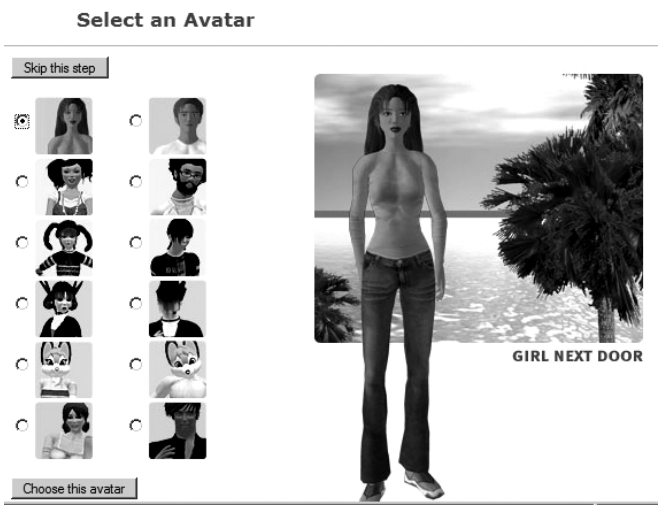


Image 3 SL: First set of avatars

Image 4.1 LM: Gender categories

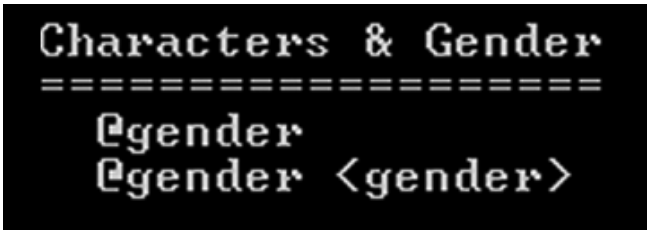


Image 4.2 LM: Re-gender

Image 5.1 SL: Male avatar without genitals

Image 5.2 SL: Edit masculinity - small

Image 5.3 SL: Edit masculinity - big

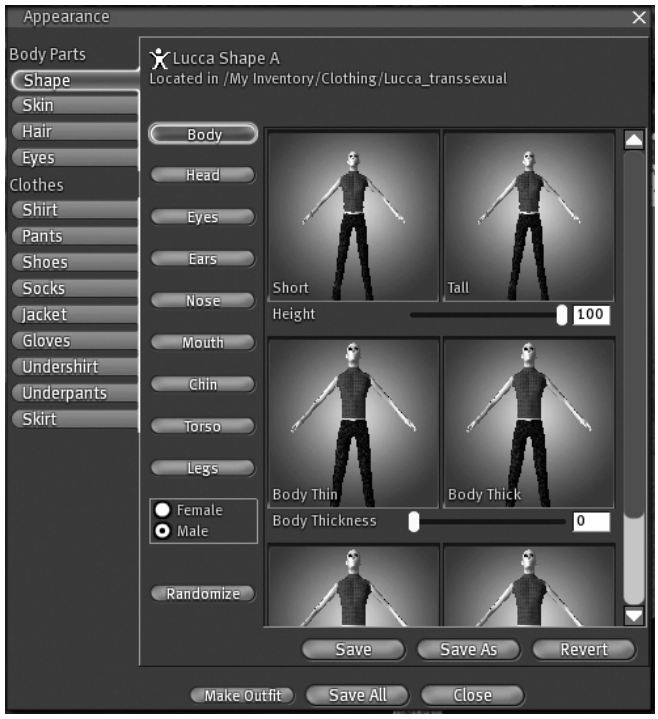


Image 6 SL:Appearance editor

Image 7.1 SL:Transgender

Image 7.2 SL: Gender swap from female to male

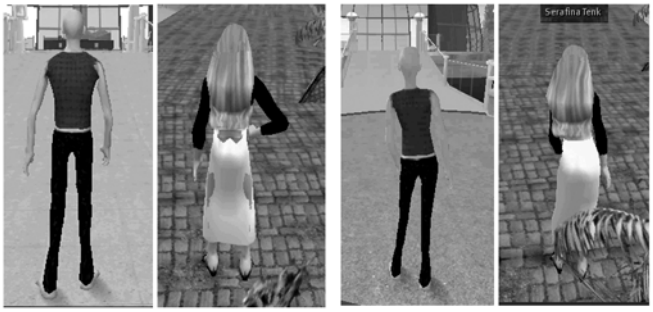


Image 8.1 SL:Walking avatars

Image 8.2 SL: Sitting bowlegged/close legged

Image 9.1 SL: Common gestures - female avatar

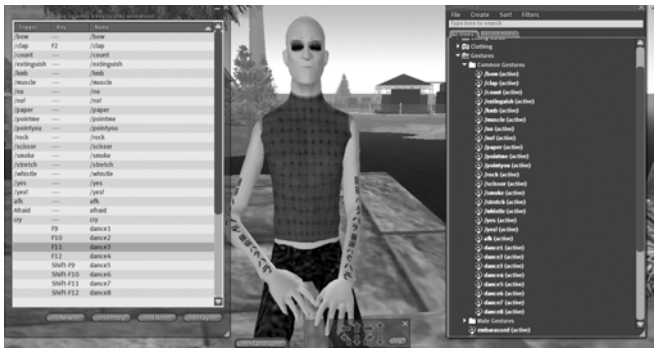


Image 9.2 SL: Common gestures - male avatar

Image 10.1 SL: Female gestures - female avatar

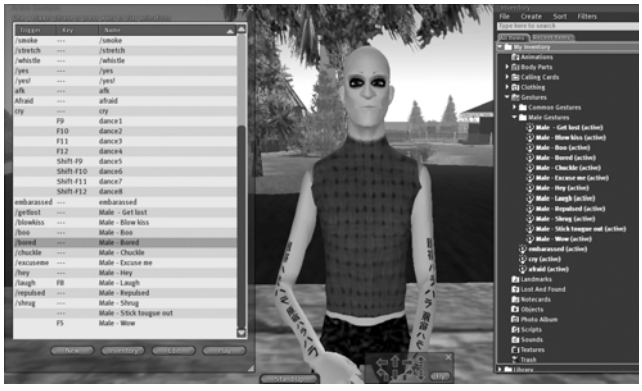


Image 10.2 SL: Male gestures - male avatar

The comparison demonstrates that both interfaces use nearly the same patterns of embodiment. Referring to the bodily aspect of gender, Second Life is a poor and conservative space. Gender modifications besides the common model seem not to be intended by the interface: In case of gender switches the gender specific gestures set does not shift correspondently (see images 11.1 SL and11.2 SL).

Image 11.1 SL: Female gestures - male avatar

Image 11.2 SL: Male gestures - female avatar

In case of a male-to-female modification, the avatars clothes do not fit any more (see image 12 SL).

Image 12 SL: Female to male swap - unfitting clothes

In contrast to that, LambdaMOO questions the analogue gender model by violating and expanding it up to 10 categories.

Outlook

This inquiry belongs to a design research project which generally deals with the relation between gender and design. It aims at the development of a gender sensitive design approach investigating the potential of gender modification and deconstruction by design.

References

Apple Computer, Inc. (1992): *Macintosh Human Interface Guidelines* (Apple Technical Library), Cupertino

Bahl, Anke (1998): MUD & MUSH. Identität und Geschlecht im Internet Ergebnisse einer qualitativen Studie. In Beinzger, Dagmar/ Eder, Sabine/Luca, Renate/ Röllecke, Renate (Hrsg.): *Im Weiberspace - Mädchen und Frauen in der Medienlandschaft,* Schriften zur Medienpädagogik Nr.26, Gesellschaft für Medienpädagogik und Kommunikationskultur, Bielefeld, p. 138 – 151

Bahl, Anke (1996): Spielraum für Rollentäuscher. Muds: Rollenspielen im Internet: http://unitopia.uni-stuttgart.de/texte/ct.html (also in: c't, 1996, Heft 8)

Becker, Barbara (2000): Elektronische Kommunikationsmedien als neue „Technologien des Selbst"? Überlegungen zur Inszenierung virtueller Identitäten in elektronischen Kommunikationsmedien. In Huber, Eva (Hrsg.): *Technologien des Selbst. Zur Konstruktion des Subjekts,* Basel/ Frankfurt a. M., P. 17 – 30

Bratteteig, Tone (2002): Bridging gender issues to technology design. In Floyd, C. et al.: *Feminist Challenges in the Information Age*. Opladen: Leske + Budrich, p. 91-106.

Butler, J. (2006): *Das Unbehagen der Geschlechter,* 1991, Suhrkamp Frankfurt a. M.

Clegg, S. and Mayfield, W. (1999): Gendered by design. In *Design Issues* 15 (3), P. 3-16

Cooper, Alan and Reimann, Robert (2003): *About Face 2.0. The Essentials of Interaction Design*, Indianapolis

Eisenrieder, Veronika (2003): *Von Enten, Vampiren und Marsmenschen - Von Männlein, Weiblein und dem "Anderen". Soziologische Annäherung an Identität, Geschlecht und Körper in den Weiten des Cyberspace*, München

Goffman, Erving (2001); *Interaktion und Geschlecht,* 2. Auflage, Frankfurt a. M.

Kleinen, Barbara (1997): Körper und Internet - Was sich in einem MUD über Grenzen lernen lässt: http://tal.cs.tu-berlin.de/~finut/mud.html (or in Bath, Corinna and Kleinen, Barbara (eds): *Frauen in der Informationsgesellschaft: Fliegen oder Spinnen im Netz?* Mössingen-Talheim, p. 42-52.

Krug, Steve (2006): *Don't make me think! Web Usability. Das intuitive Web*, Heidelberg

Lin, Yuwei (2005): Inclusion, diversity and gender equality: Gender Dimensions of the Free/Libre Open Source Software Development. Online: opensource.mit.edu/papers/lin3_gender.pdf (9.5.2007)

Lübke, Valeska (2005): *CyberGender. Geschlecht und Körper im Internet,* Königstein

McKey, Everett N. (1999): *Developing User Interfaces for Microsoft Windows,* Washington

Moss, Gloria, Gunn, Ron and Kubacki, Krysztof (2006): Successes and failures of the mirroring principle: the case of angling and beauty websites. In *International Journal of Consumer Studies*, Vol. 31 (3), p 248-257.

Müller, Jörg (1996): Virtuelle Körper. Aspekte sozialer Körperlichkeit im Cyberspace unter: http://duplox.wz-berlin.de/texte/koerper/#toc4 (oder WZB Discussion Paper FS II 96-105, Wissenschaftszentrum Berlin

Nielsen, Jakob and Loranger, Hoa (2006): *Web Usability,* München

Oudshoorn, N., Rommes, E. and Stienstra, M. (2004): Configuring the user as everybody: Gender and design cultures in information and communication technologies. In *Science, Technologie and Human Values* 29 (1), P. 30-63

Preece, Jennifer, Rogers, Yvonne and Sharp, Helen (2002): *Interaction Design. Beyond human-computer interaction*, New York

Rex, Felis alias Richards, Rob: http://www.LambdaMOO.info/

Rommes, E. (2000): Gendered User Representations. In Balka, E. and Smith, R. (ed.): *Women, Work and Computerization. Charting a Course to the Future*. Dodrecht, Boston: Kluwer Academic Pub, p. 137-145.

Second Life Blog: April 2007 Key Metrics Released http://blog.secondlife.com/2007/05/10/april-2007-key-metrics-released/, http://s3.amazonaws.com/static-secondlife-com/economy/stats_200705.xls

Teubner, U. and A. Wetterer (1999): Soziale Konstruktion transparent gemacht. In Lober, J. (Editor): *Gender Paradoxien*, Leske & Budrich: Opladen, p. 9-29.

Thomas, Sue (2003): www.barcelonareview.com/35/e_st.htm, issue 35, March - April

Trettin, K. (1997): Probleme des Geschlechterkonstruktivismus. In: G. Völger, Editor: Sie und Er, Rautenstrauch-Joest-Museum, Köln

Turkle, Sherry (1986): *Die Wunschmaschine. Der Computer als zweites Ich*, Reinbek bei Hamburg

Turkle, Sherry (1999): *Leben im Netz. Identität in Zeiten des Internet*, Reinbek bei Hamburg

West, C., Zimmermann, D.H. (1991): Doing Gender. In Lorber, J. and Farell, S. A. (eds): *The Social Construction of Gender,* Newbury Park/ London, p. 13 – 37

West, C., Zimmerman, D.H. (1995): Doing Difference. *Gender & Society*, 9, p. 8-37.

Towards a model for dynamic text editions

Dino Buzzetti
buzzetti@philo.unibo.it
University of Bologna, Italy

Malte Rehbein
malte.rehbein@nuigalway.ie
National University of Ireland, Galway, Ireland

Creating digital editions so far is devoted for the most part to visualisation of the text. The move from mental to machine processing, as envisaged in the Semantic Web initiative, has not yet become a priority for the editorial practice in a digital environment. This drawback seems to reside in the almost exclusive attention paid until now to markup at the expense of textual data models. The move from "the database as edition" [Thaller, 1991: 156-59] to the "edition as a database" [Buzzetti et al., 1992] seems to survive only in a few examples. As a way forward we might regard digital editions to care more about processing textual information rather than just being satisfied with its visualisation.

Here we shall concentrate on a recent case study [Rehbein, forthcoming], trying to focus on the kind of logical relationship that is established there between the markup and a database managing contextual and procedural information about the text. The relationship between the markup and a data model for textual information seems to constitute the clue to the representation of textual mobility. From an analysis of this kind of relationship we shall tentatively try to elicit a dynamic model to represent textual phenomena such as variation and interpretation.

I.

The case study uses the digital edition of a manuscript containing legal texts from the late medieval town Göttingen. The text shows that this law was everything else but unchangeable. With it, the city council reacted permanently on economical, political or social changes, thus adopting the law to a changing environment. The text is consequently characterised by its many revisions made by the scribes either by changing existing text or creating new versions of it. What has come to us is, thus, a multi-layered text, reflecting the evolutionary development of the law.

In order to visualise and to process the text and its changes, not only the textual expression but, what is more, its context has to be regarded and described: when was the law changed, what was the motivation for this and what were the consequences? Answers to these questions are in fact required in order to reconstruct the different layers of the text and thereby the evolution of the law. Regarding the text nowadays, it is however not always obvious how to date the alterations. It is sometimes even not clear to reveal their chronological order.

A simple example shall prove this assumption. Consider the sentence which is taken from the Göttingen bylaws about beer brewing

we ock vorschote ~~100~~ marck, de darf 3 warve bruwen

together with 150 as a replacement for 100 and 2 as a replacement for 3. (The meaning of the sentence in Low Middle German is: one, who pays 100 (150) marks as taxes, is allowed to brew beer 3 (2) times a year.) Without additional information, the four following readings are allowed, all representing different stages of the textual development:

R1: we ock vorschote 100 marck, de darf 3 warve bruwen

R2: we ock vorschote 100 marck, de darf 2 warve bruwen

R3: we ock vorschote 150 marck, de darf 3 warve bruwen

R4: we ock vorschote 150 marck, de darf 2 warve bruwen

With some more information (mainly palaeographical) but still limited knowledge, three facts become clear: firstly, that R1 is the oldest version of the text, secondly that R4 is its most recent and thirdly that either R2 or R3 had existed as text layers or none of them but not both. But what was, however, the development of this sentence? Was it the path directly from R1 to R4? Or do we have to consider R1 > R2 > R4 or R1 > R3 > R4? In order to answer these questions we need to know about the context of the text, something that can not be found in the text itself. It is the external, procedural and contextual knowledge that has to be linked to the textual expression in order to fully analyse and edit the text.

Textual mobility in this example means that, to a certain extent, the textual expression itself, its sequence of graphemes, can be regarded as invariant and objective, the external knowledge about its context cannot. It is essential in our case study not only to distinguish between the expression and the context of the text but what is more to allow flexibility in the definition and reading of (possible) text layers. It became soon clear, that for both, visualising and processing a dynamic text, a new understanding of an edition is needed, and, as a consequence, the mark-up strategy has to be reconsidered. This new understanding would "promote" the reader of an edition to its user, by making him part of it in a way that his external knowledge, his contextual setting would have influence on the representation of the text. Or in other words: dynamic text requires dynamic representation.

The way chosen in this study is to regard textual expression and context (external knowledge) separately. The expression is represented by mark-up, encoding the information about the text itself. Regarding this stand-alone, the different units of the text (its oldest version, its later alterations or annotations) could indeed be visualised but not be brought into a meaningful relationship to each other. The latter is realised by a database providing structured external information about the text, mainly what specific "role" a certain part of the text "plays" in the context of interest. Only managing and processing both, markup and database, will allow to reconstruct the different stages of the text and consequently to represent the town law in its evolutionary development.

Using the linkage mechanism between mark-up and database, the whole set of information is processable. In order to create a scholarly edition of the text, we can automatically produce a document that fulfils TEI conformity to allow the use of the widely available tools for transformation, further processing and possibly interchange.

II.

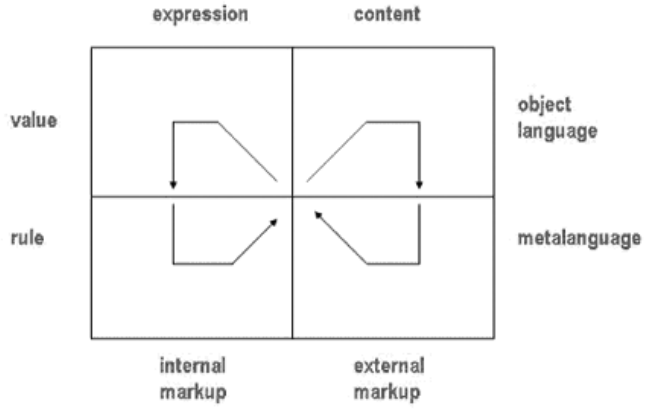
The case study just examined shows that in order to render an edition processable we have to relate the management system of the relevant data model to the markup embedded in the text. But we cannot provide a complete declarative model of the mapping of syntactic markup structures onto semantic content structures. The markup cannot contain a complete content model, just as a content model cannot contain a complete and totally definite expression of the text. To prove this fact we have to show that a markup description is equivalent to a second-order object language self-reflexive description and recall that a second-order logical theory cannot be complete. So the mapping cannot be complete, but for the same reason it can be categorical; in other words, all the models of the text could be isomorphic. So we can look for general laws, but they can provide only a dynamic procedural model.

Let us briefly outline the steps that lead to this result. In a significant contribution to the understanding of "the meaning of the markup in a document," [Sperberg-McQueen, Huitfeldt, and Renear, 2000: 231] expound it as "being constituted," and "not only described," by "the set of inferences about the document which are licensed by the markup." This view has inspired the BECHAMEL Markup Semantics Project, a ground breaking attempt to specify mechanisms "for bridging [...] syntactic relationships [...] with the distinctive semantic relationships that they represent" [Dubin and Birnbaum, 2004], and to investigate in a systematic way the "mapping [of] syntactic markup structures [on]to instances of objects, properties, and relations" [Dubin, 2003] that could be processed trough an appropriate data model. Following [Dubin and Birnbaum, 2004], "that markup can communicate the same meaning in different ways using very different syntax", we must conclude that "there are many ways of expressing the same content, just as there are many ways of assigning a content to the same expression" [Buzzetti, forthcoming].

The relationship between expression and content is then an open undetermined relationship that can be formalized by taking into account the "performative mood" of markup [Renear, 2001: 419]. For, a markup element, or any textual mark for that matter, is ambivalent: it can be seen as part of the

text, or as a metalinguistic description/ indication of a certain textual feature. Linguistically, markup behaves as punctuation, or as any other diacritical mark, i.e. as the expression of a reflexive metalinguistic feature of the text. Formally, markup behaves just as Spencer-Brown's symbols do in his formal calculus of indications [1969]: a symbol in that calculus can express both an operation and its value [Varela, 1979: 110-111].

Markup adds structure to the text, but it is ambivalent. It can be seen as the result of a restructuring operation on the expression of the text (as a textual variant) or as an instruction to a restructuring operation on the content of the text (as an interpretational variant). By way of its ambivalence it can act as a conversion mechanism between textual and interpretational variants [Buzzetti and McGann, 2006: 66] [Buzzetti, forthcoming].



Markup originates a loop connecting the structure of the text's expression to the structure of the text's content. An implementation of the *markup loop* would considerably enhance the functionality of text representation and processing in a digital edition. To achieve implementation, markup information could be integrated into the object (or datatype) 'string' on which an application system operates. Extended strings, as a datatype introduced by Manfred Thaller [1996, 2006], look as a suitable candidate for the implementation of the markup loop.

Markup originates a loop connecting the structure of the text's expression to the structure of the text's content. An implementation of the markup loop would considerably enhance the functionality of text representation and processing in a digital edition. To achieve implementation, markup information could be integrated into the object (or datatype) 'string' on which an application system operates. Extended strings, as a datatype introduced by Manfred Thaller [1996, 2006], look as a suitable candidate for the implementation of the markup loop.

Bibliography

[Buzzetti, 1992] Buzzetti, Dino, Paolo Pari e Andrea Tabarroni. 'Libri e maestri a Bologna nel xiv secolo: Un'edizione come database,' *Schede umanistiche*, n.s., 6:2 (1992), 163-169.

[Buzzetti, 2002] Buzzetti, Dino. 'Digital Representation and the Text Model,' *New Literary History*, 33:1 (2002), 61-87.

[Buzzetti, 2004] Buzzetti, Dino. 'Diacritical Ambiguity and Markup,' in D. Buzzetti, G. Pancaldi, and H. Short (eds.), *Augmenting Comprehension: Digital Tools and the History of Ideas*, London-Oxford, Office for Humanities Communication, 2004, pp. 175-188: URL = <http://137.204.176.111/dbuzzetti/pubblicazioni/ambiguity.pdf>

[Buzzetti and McGann, 2006] Buzzetti, Dino, and Jerome McGann. 'Critical Editing in a Digital Horizon,' in *Electronic Textual Editing*, ed. Lou Burnard, Katherine O'Brien O'Keeffe, and John Unsworth, New York, The Modern Language Association of America, 2006, pp. 51-71.

[Buzzetti, forthcoming] Buzzetti, Dino. 'Digital Editions and Text Processing,' in *Text Editing in a Digital Environment*, Proceedings of the AHRC ICT Methods Network Expert Seminar (London, King's College, 24 March 2006), ed. Marilyn Deegan and Kathryn Sutherland (Digital Research in the Arts and Humanities Series), Aldershot, Ashgate, forthcoming.

[Dubin, 2003] Dubin, David. 'Object mapping for markup semantics,' *Proceedings of Extreme Markup Languages 2003*, Montréal, Québec, 2003: URL = <http://www.idealliance.org/papers/extreme/proceedings//html/2003/Dubin01/EML2003Dubin01.html>

[Dubin and Birnbaum, 2004] Dubin, David, and David J. Birnbaum. 'Interpretation Beyond Markup,' *Proceedings of Extreme Markup Languages 2004*, Montréal, Québec, 2004: URL = <http://www.idealliance.org/papers/extreme/proceedings//html/2004/Dubin01/EML2004Dubin01.html>

[McGann, 1991] McGann, Jerome. *The Textual Condition*, Princeton, NJ, Princeton University Press, 1991.

[McGann, 1999] McGann, Jerome. 'What Is Text? Position statement,' *ACH-ALLC'99 Conference Proceedings*, Charlottesville, VA, University of Virginia, 1999: URL = <http://www.iath.virginia.edu/ach-allc.99/proceedings/hockey-renear2.html>

[Rehbein, forthcoming] Rehbein, Malte. Reconstructing the textual evolution of a medieval manuscript.

[Rehbein, unpublished] Rehbein, Malte. Göttinger Burspraken im 15. Jahrhundert. Entstehung – Entwicklung – Edition. PhD thesis, Univ. Göttingen.

[Renear, 2001] Renear, Allen. 'The descriptive/procedural distinction is flawed,' *Markup Languages: Theory and Practice*, 2:4 (2001), 411–420.

[Sperberg-McQueen, Huitfeldt and Renear, 2000] Sperberg-McQueen, C. M., Claus Huitfeldt, and Allen H. Renear. 'Meaning and Interpretation of Markup,' *Markup Languages: Theory and Practice*, 2:3 (2000), 215-234.

[Thaller, 1991] Thaller, Manfred. 'The Historical Workstation Project,' *Computers and the Humanities*, 25 (1991), 149-62.

[Thaller, 1996] Thaller, Manfred. 'Text as a Data Type,' *ALLC-ACH'96: Conference Abstracts,* Bergen, University of Bergen, 1996.

[Thaller, 2006] Thaller, Manfred. 'Strings, Texts and Meaning.' *Digital Humanities 2006: Conference Abstracts*, Paris, CATI - Université Paris-Sorbonne, 2006, 212-214.

Reflecting on a Dual Publication: Henry III Fine Rolls Print and Web

Arianna Ciula
arianna.ciula@kcl.ac.uk
King's College London, UK

Tamara Lopez
tamara.lopez@kcl.ac.uk
King's College London, UK

A collaboration between the National Archives in the UK, the History and Centre for Computing in the Humanities departments at King's College London, the Henry III Fine Rolls project (http://www.frh3.org.uk) has produced both a digital and a print edition (the latter in collaboration with publisher Boydell & Brewer) [1] of the primary sources known as the Fine Rolls. This dual undertaking has raised questions about the different presentational formats of the two resources and presented challenges for the historians and digital humanities researchers involved in the project, and, to a certain extent, for the publisher too.

This paper will examine how the two resources evolved: the areas in which common presentational choices served both media, and areas in which different presentational choices and production methodologies were necessary. In so doing, this paper aims to build a solid foundation for further research into the reading practices and integrated usage of hybrid scholarly editions like the Henry III Fine Rolls.

Presentation as interpretation

In Material Culture studies and, in particular, in studies of the book, the presentational format of text is considered to be of fundamental importance for the study of production, social reading and use. Therefore, description of and speculation about the physical organisation of the text is essential of understanding the meaning of the artefact that bears that text. Similarly, in Human Computer Interaction studies and in the Digital Humanities, the presentation of a text is considered to be an integral outgrowth of the data modelling process; a representation of the text but also to some degree an actualisation of the interpretative statements about the text. Indeed, to the eyes of the reader, the presentational features of both a printed book and a digital written object will not only reveal the assumptions and beliefs of its creators, but affect future analysis of the work.

Dual publication: digital and print

On the practical side, within the Henry III Fine Rolls project, different solutions of formatting for the two media have been negotiated and implemented.

The print edition mainly represents a careful pruning of the digital material, especially as pertains to the complex structure of the indexes.

The indexes were built upon the marked-up fine rolls texts and generated from an ontological framework (Ciula, Spence, Vieira, & Poupeau; 2007). The latter, developed through careful analysis by scholars and digital humanities researchers, constitutes a sort of an *a posteriori* system that represents familial networks, professional relationships, geo-political structures, thematic clusters of subjects, and in general various types of associations between the 13th century documents (the so called Fine Rolls) and the the roles played by places and people in connection with them.

The ontology is used to produce a series of pre-coordinate axes (the indices) that the reader can follow to explore the texts. The flexibility of the ontology allows the texts to be fairly exhaustively indexed, just as the presentational capabilities of the digital medium allow for the display and navigation of indexes that are correspondingly large.

By contrast, the print edition had to follow the refined conventions of a well established scholarly tradition in publishing editions in general and calendar [2] editions in particular, both in terms of formatting and, more importantly for us, in terms of content selection/creation and modelling.

Though the indices within the printed edition are also pre-coordinate axes along which to explore the text, the way in which they are produced is perceived to be a nuanced and intuitive aspect of the scholarship and one that revealed itself to be less tolerant to change. This, coupled with the presentational constraints of the printed medium result in indices that present information succinctly and with a minimum of conceptual repetition. Similarly, the first print volume of around 560 pages gives absolute prominence -something that can be stated much more strongly in a linear publication than in a digital one- to a long and detailed historical introduction, followed by a section on the adopted editorial strategies.

However, the two artefacts of the project also share many points in common, either because the digital medium had to mirror the tradition of its more authoritative predecessor, or for more practical -nevertheless not to be dismissed- reasons of work-flow and foreseen usage. An interesting example of the latter is the adopted layout of footnotes, where the print format was modelled on the base of the digital layout and, although it was a completely unusual arrangement, was accepted as suitable by the publisher.

On the base of the work done so far and on the feedback on the use of the book and the website, the presentational format will be refined further for future print volumes to come and for the additional material to be included in the digital edition before the end of the project.

One reading process

On the methodological side, we believe that further research into the usage and reading process of these parallel publications could lead towards a better understanding of scholarly needs and therefore a better modelling of such a dual product that is becoming a more and more common deliverable in digital humanities projects.

As this paper will exemplify, the presentation of data needs to be tailored to take into account the more or less fine conventions of two different media which have different traditions, different life cycles, different patterns of use and, possibly, different users.

However, although very different in nature, these two publications are not necessarily perceived and – more importantly- used as separate resources with rigid boundaries between them. For a scholar interested in the fine rolls, the reading of the edition and the seeking of information related to it (persons, places, subjects and any other interesting clue to its historical study in a broader sense) is a global process that does not stop when the book is closed or the the browser shut. We believe that, when supported by a deep interest in the material, the connection between the two publications is created in a rather fluid manner.

The reality of the reading process and information seeking, as it happens, is influenced by the products it investigates, but ultimately has a form of its own that is different from the objects of analysis. It is dynamic and heterogeneous, it leaves on the integration between different types of evidence, no matter what their format is, including other kind of external sources. Indeed, the library or archive is the most likely environment where a scholar of the fine rolls would find herself browsing the print or digital edition, eventually the original primary sources or their digital images, plus any range of secondary sources.

Studying the integration of print and digital

The data behind the two publications are drawn from the same informational substrate, but are separated to create two presentational artefacts. As established, *reading* is expected to be the primary activity performed using both and a stated design goal for the project is that the two artefacts will form a rich body of materials with which to conduct historical research. The heterogeneity of the materials, however, suggests that working with texts will of necessity also involve periods of *information seeking*: moments while reading that give rise to questions which the material at hand cannot answer and the subsequent process embarked upon in order to answer them. Our working hypothesis is that to fill these information gaps (Wilson, 1999), the reader will turn to particular texts in the alternative medium to find answers, moving between the website and the books, fluctuating between states of reading and seeking.

Thus, the analytical stream in this paper will move from the practices of creating two types of resources to establishing an analytical framework for evaluating their use. Situating the project materials and domain experts within the literature of information behaviour research, we will identify and develop a model for evaluating how well the features of the website and the book support information seeking activities that bridge (Wilson, 1999) reading within the individual media.

Conclusions

Based on our experience in creating a hybrid edition for the Henry III Fine Rolls project, the challenges and adopted solutions for the two types of published resources are a starting point from which to reflect on the integrated production of a dual object. At the same time, continuing work begun elsewhere in the digital humanities (Buchanan, Cunningham, Blandford, Rimmer, & Warwick; 2006) to adapt methodologies used in Information Science and Book Studies, a rationale and method for the design of an analysis of their use and, in particular, of the interaction between scholars and the website/books can be outlined.

Notes

[1] The first volume was published in September 2007 (Dryburgh et al. 2007).

[2] Calendar stays here for an an English summary of the Latin records.

Bibliography

Buzetti, Dino and Jerome McGann (2005) "Critical Editing in a Digital Horizon". In Burnard, O'Keeffe, and Unsworth eds. *Electronic Textual Editing*.

<http://www.tei-c.org/Activities/ETE/Preview/mcgann.xml>.

Buchanan, G.; Cunningham, S.; Blandford, A.; Rimmer, J. & Warwick, C. (2005), 'Information seeking by humanities scholars', *Research and Advanced Technology for Digital Libraries, 9th European Conference, ECDL*, 18--23.

Ciula, A. Spence, P., Vieira, J.M., Poupeau, G. (2007) *Expressing Complex Associations in Medieval Historical Documents: The Henry III Fine Rolls Project*. Paper presented at Digital Humanities 2007, Urbana-Champaign, 4-8 June, 2007. <http://www.digitalhumanities.org/dh2007/abstracts/xhtml.xq?id=196>

Dervin, B. (1983) "An overview of sense-making research: Concepts, methods, and results to date", *International Communication Association Annual Meeting, Dallas, TX*, 1983.

Drucker, Johanna (2003). "The Virtual Codex from Page Space to E-space." *The Book Arts Web* <http://www.philobiblon.com/drucker/>

Dryburgh, Paul and Beth Hartland eds. Arianna Ciula and José Miguel Vieira tech. Eds. (2007) *Calendar of the Fine Rolls of the Reign of Henry III [1216-1248]*, vol. I: 1216-1224, Woodbridge: Boydell & Brewer.

Lavagnino, John (2007). *Being Digital, or Analogue, or Neither*. Paper presented at Digital Humanities 2007, Urbana-Champaign, 4-8 June, 2007. <http://www.digitalhumanities.org/dh2007/abstracts/xhtml.xq?id=219>

McGann, Jerome. (1997) "The Rational of Hypertext." In Kathryn Sutherland ed. *Electronic text: investigations in method and theory*. Clarendon Press: Oxford: 19-46.

Siemens, Ray, Elaine Toms, Stéfan Sinclair, Geoffrey Rockwell, and Lynne Siemens. "The Humanities Scholar in the Twenty-first Century: How Research is Done and What Support is Needed." *ALLC/ACH 2004 Conference Abstracts*. Göteborg: Göteborg University, 2004. <http://www.hum.gu.se/allcach2004/AP/html/prop139.html>

Wilson, T. (1999) "Models in information behaviour research." *Journal of Documentation* 55(3), 249-270.

Performance as digital text: capturing signals and secret messages in a media-rich experience

Jama S. Coartney
jcoartney@virginia.edu
University of Virginia Library, USA

Susan L. Wiesner
slywiesner@virginia.edu
University of Virginia Library, USA

Argument

As libraries increasingly undertake digitisation projects, it behooves us to consider the collection/capture, organisation, preservation, and dissemination of all forms of documentation. By implication, then, these forms of documentation go beyond written text, long considered the staple of library collections. While several libraries have funded projects which acknowledge the need to digitise other forms of text, in graphic and audio formats, very few have extended the digital projects to include film, much less performed texts. As more performing arts incorporate born-digital elements, use digital tools to create media-rich performance experiences, and look to the possibility for digital preservation of the performance text, the collection, organisation, preservation, and dissemination of the performance event and its artefacts must be considered. The ARTeFACT project, underway at the Digital Media Lab at the University of Virginia Library, strives to provide a basis for the modeling of a collection of performance texts. As the collected texts document the creative process both prior to and during the performance experience, and, further, as an integral component of the performance text includes the streaming of data signals to generate audio/visual media elements, this paper problematises the capture and preservation of those data signals as artefacts contained in the collection of the media-rich performance event.

Premise

In a report developed by a working group at the New York Public Library, the following participant thoughts are included:

> Although digital technologies can incorporate filmic ways of perceiving [the performing arts], that is the tip of the iceberg. It is important for us to anticipate that there are other forms we can use for documentation rather than limiting ourselves to the tradition of a camera in front of the stage. Documentation within a digital environment far exceeds the filmic way of looking at a performance' (Ashley 2005 NYPL Working Group 4, p.5)

> How can new technology both support the information we think is valuable, but also put it in a format that the next generation is going to understand and make use of?' (Mitoma 2005 NYPL Working Group 4, p.6)

These quotes, and many others, serve to point out current issues with the inclusion of the documentation of movement-based activities in the library repository. Two important library tasks, those of the organization and dissemination of text, require the development of standards for metadata. This requirement speaks towards the need for enabling content-based searching and dissemination of moving-image collections. However, the work being performed to provide metadata schemes of benefit to moving image collections most often refers to (a) a filmed dramatic event, e.g. a movie, and/or (b) metadata describing the film itself. Very little research has been completed in which the moving image goes beyond a cinematic film, much less is considered as one text within a multi-modal narrative.

In an attempt to address these issues, the authors developed the ARTeFACT project in hopes of creating a proof-of-concept in the University of Virginia Library. Not content, however, to study the description of extant, filmic and written texts, the project authors chose to begin with describing during the process of the creation of the media-rich, digital collection, including the description of a live performance event. The decision to document a performance event begged another set of answers to questions of issues involved with the collection of texts in a multiplicity of media formats and the preservation of the artefacts created through a performance event.

Adding to this layer of complexity was the additional decision to create not just a multi-media performance, but to create one in which a portion of the media was born-digital during the event itself. Created from signals transmitted from sensor devices (developed and worn by students), the born-digital elements attain a heightened significance in the description of the performance texts. After all, how does one capture the data stream for inclusion in the media-rich digital collection?

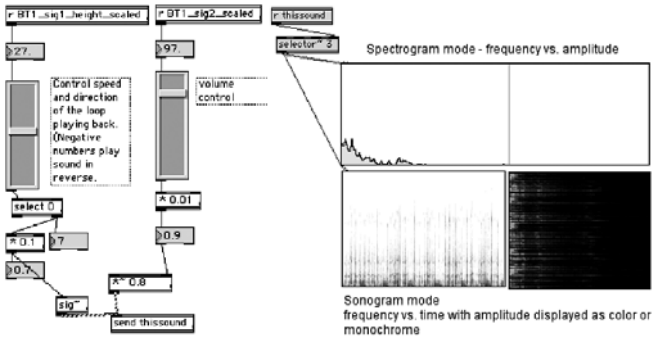
Methodology

The ARTeFACT project Alpha includes six teams of students in an Introductory Engineering Class. Each team was asked to design and build an orthotic device that, when worn, causes the wearer to emulate the challenges of walking with a physical disability (included are stroke 'drop foot' and paralysis, CP hypertonia, Ricketts, etc.) During the course of the semester, the student teams captured pre-event process in a variety of digital formats: still and video images of the prototypes from cameras and cell phones, PDF files, CAD drawings, PowerPoint files and video documentation of those presentations. The digital files were collected in a local SAKAI implementation.

In addition to these six teams, two other teams were assigned the task of developing wireless measurement devices which, when attached to each orthotic device, measures the impact

of the orthotic device on the gait of the wearer. The sensor then transmits the measurement data to a computer that feeds the data into a Cycling74's software application: Max/MSP/Jitter. Jitter, a program designed to take advantage of data streams, then creates real-time data visualization as output. The resultant audio visual montage plays as the backdrop to a multi-media event that includes student 'dancers' performing while wearing the orthotic devices.

The sensors generate the data signals from Bluetooth devices (each capable of generating up to eight independent signals) as well as an EMG (electro-myogram) wireless system. At any given time there may be as many as seven signalling devices sending as many as 50 unique data streams into the computer for processing. As the sensor data from the performers arrives into Max/MSP/Jitter, it is routed to various audio and video instruments within the application, processed to generate the data visualization, then sent out of the computer via external monitor and sound ports. The screenshot below displays visual representations of both the input (top spectrogram) and output (bottom: sonogram) of a real-time data stream. The data can be visualized in multiple ways; however, the data stream as we wish to capture it is not a static image, but rather a series of samples of data over time.



There are several options for capturing these signals, two of which are: writing the data directly to disk as the performance progresses and/or the use of external audio and video mixing boards that in the course of capturing can display the mix.

Adding to the complexity, the performance draws on a wide variety of supporting, media rich, source material created during the course of the semester. A subset of this material is extrapolated for use in the final performance. These elements are combined, processed, morphed, and reformatted to fit the genre of the presentation and although they may bear some similarity to the original material, they are not direct derivatives of the source and thus become unique elements in the production. Further, in addition to the capture of data streams, the totality of the performance event must be collected. For this, traditional means of capturing the performance event have been determined to be the simplest of the challenges faced. Therefore, a video camera and a microphone pointed at the stage will suffice to fill this minimum requirement for recording the event.

Conclusion

The complexity of capturing a performance in which many of the performance elements themselves are created in real time, processed, and used to generate audio visual feedback is challenging. The inclusion of data elements in the artefact collection begs questions with regard to the means of capturing the data without impacting the performance. So, too, does it require that we question what data to include: Does it make sense to capture the entire data stream or only the elements used at a specific instance in time to generate the performance; what are the implications of developing a sub-system within the main performance that captures this information? When creating a collection based on a performance as digital text, and before any work may be done to validate metadata schemes, we must answer these questions. We must consider how we capture the signals and interpret the secret messages generated as part of the media-rich experience.

Sample bibliography

Adshead-Lansdale, J. (ed.) 1999, *Dancing Texts: intertextuality and interpretation* London: Dance Books.

Goellner, E. W. & Murphy, J. S. 1995, *Bodies of the Text* New Brunswick, NJ: Rutgers University Press.

Kholief, M., Maly, K. & Shen, S. 2003, 'Event-Based Retrieval from a Digital Library containing Medical Streams' in *Proceedings of the 2003 Joint Conference on Digital Libraries* (Online) Available at http://ieeexplore.ieee.org/Xplore/login.jsp?url=/iel5/8569/27127/01204867.pdf

New York Public Library for the Performing Arts, Jerome Robbins Dance Division 2005, Report from Working Group 4 *Dance Documentation Needs Analysis* Meeting 2, NYPL: New York.

Reichert, L. 2007, 'Intelligent Library System Goes Live as Satellite Data Streams in Real-Time' (Online) Available at http://www.lockheedmartin.com/news/press_releases/2001/IntelligentLibrarySystemGoesLiveAsS.html

Function word analysis and questions of interpretation in early modern tragedy

Louisa Connors
Louisa.Connors@newcastle.edu.au
University of Newcastle, Australia

The use of computational methods of stylistic analysis to consider issues of categorization and authorship is now a widely accepted practice. The use of computational techniques to analyze style has been less well accepted by traditional humanists. The assumption that most humanists make about stylistics in general, and about computational stylistics in particular, is that it is "concerned with the formal and linguistic properties of the text as an isolated item in the work" (Clark 2005). There are, however, two other points of emphasis that are brought to bear on a text through a cognitive approach. These are: "that which refers to the points of contact between a text, other texts and their readers/listeners", and "that which positions the text and the consideration of its formal and psychological elements within a socio-cultural and historical context" (Clark 2005). Urszula Clark (2005) argues that these apparently independent strands of analysis or interpretive practice are an "integrated, indissolvable package" (Clark 2005).

Computational stylistics of the kind undertaken in this study attempts to link statistical findings with this integrated indissolvable package; it highlights general trends and features that can be used for comparative purposes and also provides us with evidence of the peculiarities and creative adaptations of an individual user. In this case, the individual user is Elizabeth Cary, the author of the earliest extant original play in English by a woman, *The Tragedy of Mariam, The Fair Queen of Jewry* (1613). As well as *Mariam*, the set of texts in the sample includes the other 11 closet tragedies associated with the "Sidney Circle", and 48 tragedies written for the public stage. All plays in the study were written between 1580 and 1640.[1] The only other female authored text in the group, Mary Sidney's *Antonius*, is a translation, as is Thomas Kyd's *Cornelia*[2]. Alexander Witherspoon (1924), describes the Sidnean closet tragedies as "strikingly alike, and strikingly unlike any other dramas in English" (179). He attributes this to the extent to which the closet writers draw on the work of French playwright Robert Garnier as a model for their own writing. In contrast to other plays of the period closet tragedies have not attracted much in the way of favourable critical attention. They are, as Jonas Barish (1993) suggests, "odd creatures" (19), and Mariam is described as one of oddest.

Mariam, as it turns out, is the closet play that is *most* like a play written for the public stage, in terms of the use of function words. But this isn't immediately obvious. Some textual preparation was carried out prior to the analysis. Homographs were not tagged, but contracted forms throughout the texts were expanded so that their constituents appeared as separate words. The plays were divided into 2,000 word segments and tagged texts were then run through a frequency count using *Intelligent Archive* (IA). A total of 563 two-thousand word segments were analysed, 104 of which were from closet plays, and 459 from plays written for the public stage. A discriminant analysis on the basis of the frequency scores of function words demonstrates that there are significant differences between the two groups of plays. Table 1 shows the classification results for a discriminant analysis using the full set of function words. In this test, 561 of the 563 segments were classified correctly. One segment from each group was misclassified. Thus 99.6% of cross-validated grouped cases were correctly classified on the basis of function words alone. The test also showed that only 38 of the 241 function word variables were needed to successfully discriminate between the groups.

Table 1

Classification results for discriminant analysis using the full set of function words in 60 tragedies (1580-1640) closet/non-closet value correctly assigned

		Closet/Stage	Predicted Group Membership		Total
			Closet	Public	
Original	Count	Closet	103	1	104
		Public	1	458	459
	%	Closet	99.0	1.0	100.0
		Public	.2	99.8	100.0
Cross-validated (a)	Count	Closet	103	1	104
		Public	1	458	459
	%	Closet		1.0	100.0
		Public		99.8	100.0

a Cross validation is done only for those cases in the analysis. In cross validation, each case is classified by the functions derived from all cases other than that case.

b 99.6% of original grouped cases correctly classified.

c 99.6% of cross-validated grouped cases correctly classified.

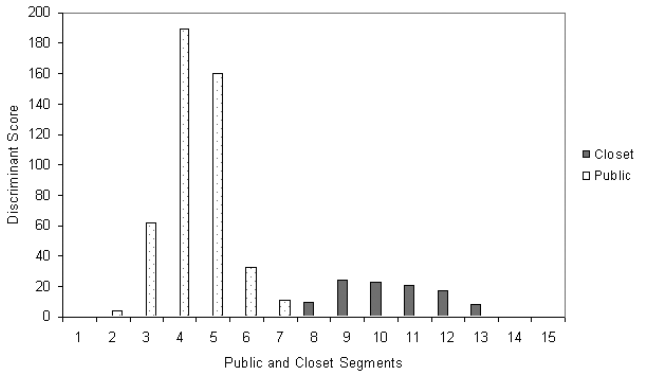


Figure 1. Discriminant scores for correctly identified public and closet play segments from 60 tragedies (1580-1640) in 2000 word segments on the basis of 38 most discriminating function words

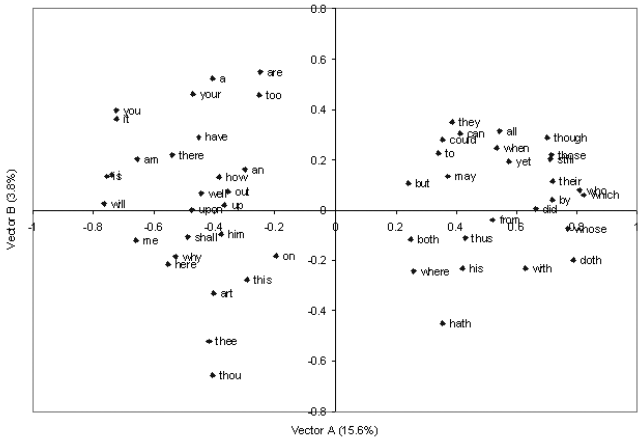


Figure 2. Principal component analysis for 60 tragedies (1580-1640) in 4000 word segments for 54 most discriminating function words selected from 100 most frequently occurring function words - word plot for first two eigenvectors

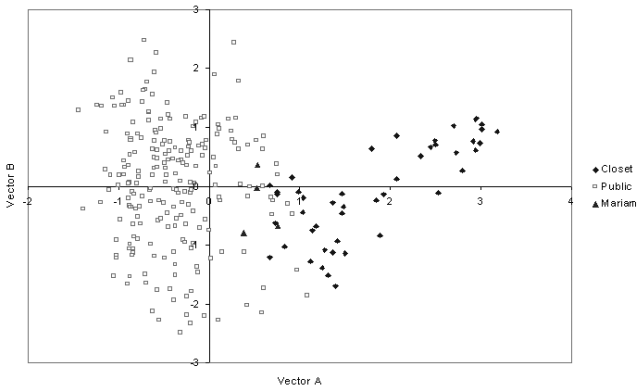


Figure 3. Principal component analysis for 60 tragedies (1580-1640) in 4000 word segments for 54 most discriminating function words selected from 100 most frequently occurring function words

A principal component analysis (PCA) gives additional information about the differences and similarities between the two sets. An Independent-samples T-test was used to identify the variables most responsible for the differences. This process picked out 54 variables that were significant at the level of 0.0001 and these 54 variables were used for the PCA. PCA looks to define factors that can be described as most responsible for differences between groups. For this text, the texts were broken into 4000 word segments to ensure that frequencies remained high enough for reliable analysis. This produced 268 segments in total (50 closet segments and 218 public play segments). Figure 2 plots the 54 most discriminating words-types for the first two eigenvectors (based on factor loadings for each variable) and shows which words behave most like or unlike each other in the sample.

In Figure 3 the eigenvalues from the component matrix have been multiplied through the standardised frequencies for each of the 4,000 word segments to show which segments behave most like or unlike each other on the first two principal components. The scores that produce Figure 3 are "the sum of the variable counts for each text segment, after each count is multiplied by the appropriate coefficient" (Burrows and Craig 1994 68). High counts on word variables at the western end of Figure 2 and low counts on word variables at the eastern end bring the text segments at the western end of Figure 3 to their positions on the graph. The reverse is true for text segments on the eastern side of Figure 3. We can see that the far western section of Figure 3 is populated predominantly with public play segments, and that the eastern side of the *y*-axis is populated exclusively with segments from stage plays. It is clear that in the case of these most discriminating variables, the segments from *Mariam* are the closet segments most intermingled with the segments written for the public stage.

Looking at Figure 2 there is evidence of an "emphasis on direct personal exchange" in western section of the graph. In the opposite section of the graph there is evidence of a more disquisitory style of language that is "less personal", with "markers of plurality, past time, and a more connected syntax" (Burrows and Craig 1994 70). It may be that the results reflect the kinds of observations that critics have long made about early modern closet tragedy and tragedy written for the public stage, suggesting that "word-counts serve as crude but explicit markers of the subtle stylistic patterns to which we respond when we read well" (Burrows and Craig 1994 70). It may also be the case, however, that we can link these statistical results with more interpretive work.

Returning to *Mariam*, the function words which most distinguish the *Mariam* segments from the rest of the segments of both closet and public plays, are the auxiliary verbs *did* (8.4) and *had* (6.9). In the case of *did* over 500 of the segments have scores of between -1 and 2. Six of the eight of the *Mariam* segments are extremely high (the two middle segments of *Mariam* have fairly average z-scores for *did*). The lowest score occurs in segment 4, when Mariam is conspicuously absent from the action. A very similar pattern emerges for *had*. In conventional grammars *do* and other auxiliaries including *be* and *have* are viewed as meaningless morphemes that serve a grammatical purpose. Langacker argues that serving a specifiable grammatical function is not inherently incompatible with being a meaningful element (1987 30).

Cognitive linguistics suggests that function word schemas interact with each other to produce what Talmy calls a "dotting" of semantic space (1983 226), and that they "play a basic conceptual structuring role" (Talmy 88 51). In this framework, auxiliaries are viewed as profiling a process – they determine which entity is profiled by a clause and impose a particular construal. Langacker argues, for example, that *do* always conveys some notion of activity or some kind of volitionality or control on the part of the subject. *Have* designates a subjectively construed relation of anteriority and current relevance to a temporal reference point (Langacker 1991 239). Talmy argues that *have* can be understood in terms of force dynamics patterns; it "expresses indirect causation either without an intermediate volitional entity…or… with

such an entity" (1988 65). Talmy goes further to suggest that the "concepts" of force dynamics are "extended by languages to their semantic treatment of *psychological* elements and interactions" (1988 69).

Bringing the tools of cognitive linguistics to bear on the results of computational analysis of texts can provide a framework that validates the counting of morphemes like did and had. The same framework may also shed light on questions of interpretation. This approach appears to provide a disciplined way of identifying and analyzing the linguistic features that are foregrounded in a text, while supporting their interpretation as part of an integrated, indissolvable package.

Notes

1 Thomas Kyd wrote for both the public and the private stage. Kyd's Cornelia and The Spanish Tragedy are included in the study.

2 John Burrows (2002) explores some of the issues around translation and whether it can be "assumed that poets stamp their stylistic signatures as firmly on translation as their original work" (679). Burrows found that Drydan was able to "conceal his hand", but in other cases it appeared that a "stylisitic signature", even in the case of a translation, remained detectable (696).

References

Barish, J. (1993). Language for the Study: Language for the Stage. In A. L. Magnusson & C. E. McGee (Eds.), *The ElizabethanTheatre XII* (pp. 19-43). Toronto: P.D. Meany.

Burrows, J. (2002). The Englishing of Jevenal: Computational stylistics and translated Texts. Style, 36, 677-750.

Burrows, J. F., & Craig, D. H. (1994). Lyrical Drama and the "Turbid Mountebanks": Styles of Dialogue in Romantic and Renaissance Tragedy. *Computers and the Humanities, 28*, 63-86.

Cooper, M. M. (1998). Implicature and *The Taming of the Shrew.* In J. Culpeper, M. H. Short & P. Verdonk (Eds.), *Exploring the Language of Drama: From Text to Context* (pp. 54-66). London: Routledge.

Clark, U. (2005). *Social cognition and the future of stylistics, or "What is cognitive stylistics and why are people saying such good things about it?!*" Paper presented at the PALA 25: Stylistics and Social Cognition.

Connors, L. (2006). An Unregulated Woman: A computational stylistic analysis of Elizabeth Cary's *The Tragedy of Mariam, The Faire Queene of Jewry. Literary and Linguistic Computing, 21* (Supplementary Issue), 55-66.

Sweetser, E. (1990). *From Etymology to Pragmatics: Metaphorical and cultural aspects of semantic change.* Cambridge: Cambridge University Press.

Langacker, R. W. (1999). Losing control: grammaticization, subjectification, and transparency. In A. Blank & P. Koch (Eds.), *Historical Semantics and Cognition* (Vol. 13, pp. 147-175). Berlin: Mouton de Gruyter.

Talmy, L. (1983). How language structures space. In H. Pick & L. Acredolo (Eds.), *Spatial orientation: Theory, research, and application.* New York: Plenum Press.

Talmy, L. (1988). Force Dynamics in Language and Cognition. *Cognitive Science, 12.* 49-100.

Witherspoon, A. M. (1924: rpt. 1968). *The Influence of Robert Garnier on Elizabethan Drama.*

Deconstructing Machine Learning: A Challenge for Digital Humanities

Charles Cooney
cmcooney@diderot.uchicago.edu
University of Chicago, USA

Russell Horton
russ@diderot.uchicago.edu
University of Chicago, USA

Mark Olsen
mark@barkov.uchicago.edu
University of Chicago, USA

Glenn Roe
glenn@diderot.uchicago.edu
University of Chicago, USA

Robert Voyer
rlvoyer@diderot.uchicago.edu
University of Chicago, USA

Machine learning tools, including document classification and clustering techniques, are particularly promising for digital humanities because they offer the potential of using machines to discover meaningful patterns in large repositories of text. Given the rapidly increasing size and availability of digital libraries, it is clear that machine learning systems will, of necessity, become widely deployed in a variety of specific tasks that aim to make these vast collections intelligible. While effective and powerful, machine learning algorithms and techniques are not a panacea to be adopted and employed uncritically for humanistic research. As we build and begin to use them, we in the digital humanities must focus not only on the performance of specific algorithms and applications, but also on the theoretical and methodological underpinnings of these systems.

At the heart of most machine learning tools is some variety of classifier. Classification, however, is a fraught endeavor in our poststructuralist world. Grouping things or ideas into categories is crucial and important, but doing so without understanding and being aware of one's own assumptions is a dubious activity, not necessarily for moral, but for intellectual reasons. After all, hierarchies and orders of knowledge have been shown to be both historically contingent and reflections of prevailing power structures. Bowker and Star state that "classification systems in general... reflect the conflicting, contradictory motives of the sociotechnical situations that gave rise to them" (64). As machine learning techniques become more widely applied to all forms of electronic text, from the WWW to the emerging global digital library, an awareness of the politics of classification and the ordering of knowledge will become ever more important. We would therefore like to present a paper outlining our concerns about these techniques and their underlying technical/intellectual assumptions based on our experience using them for experimental research.

In many ways, machine learning relies on approaches that seem antithetical to humanistic text analysis and reading and to more general poststructuralist sensibilities. The most powerful and effective techniques rely on the abilities of systems to classify documents and parts of documents, often in binary oppositions (spam/not spam, male/female, etc). Features of documents employed in machine learning applications tend to be restricted to small subsets of available words, expressions or other textual attributes. Clustering of documents based on relatively small feature sets into a small and often arbitrary number of groups similarly tends to focus on broad patterns. Lost in all of these operations are the marginal and exceptional, rendered hidden and invisible as it were, in classification schemes and feature selection.

Feature set selection is the first necessary step in many text mining tasks. Ian Witten notes that in "many practical situations there are far too many attributes for learning schemes to handle, and some of them -- perhaps the overwhelming majority -- are clearly irrelevant or redundant" (286-7). In our work, we routinely reduce the number of features (words, lemmas, bigrams, etc) using a variety of techniques, most frequently by filtering out features which occur in a small subset of documents or instances. This selection process is further required to avoid "overfitting" a learner to the training data. One could build an effective classifier and train it using features that are unique to particular documents, but doing so would limit the general applicability of the tool. Attempting to classify French novels by gender of author while retaining the names of characters (as in Sand's novel, *Conseulo*) or other distinctive elements is very effective, but says little about gendered writing in 19th century France (Argamon et. al., *Discourse*). Indeed, many classification tasks may be successful using a tiny subset of all of the words in a corpus. In examining American and non-American Black Drama, we achieved over 90% accuracy in classifying over nearly 700 plays using a feature set of only 60 surface words (Argamon et. al., *Gender, Race*). Using a vector space similarity function to detect articles in the *Encyclopédie* which borrow significantly from the *Dictionnaire de Trévoux,* we routinely get impressive performance by selecting fewer than 1,000 of the 400,000 unique forms in the two documents (Allen et. al.). The requirement of greatly reductive feature set selection for practical text mining and the ability of the systems to perform effective classifications based on even smaller subsets suggests that there is a significant distance from the texts at which machine learning must operate in order to be effective.

Given the reductive nature of the features used in text mining tasks, even the most successful classification task tends to highlight the lowest common denominators, which at best may be of little textual interest and at worst extremely misleading, encouraging stereotypical conclusions. Using a decision tree to classify modern and ancient geography articles in the *Encyclopédie*, we found "selon" (according to) to be the primary distinction, reflecting citation of ancient

sources ("selon Pline"). Classification of Black Drama by gender of author and gender of speaker can be very effective (80% or more accuracy), but the features identified by the classifiers may privilege particular stereotypes. The unhappy relationship of Black American men with the criminal justice system or the importance of family matters to women are both certainly themes raised in these plays. Of course, men talk more of wives than women and only women tend to call other women "hussies," so it is hardly surprising that male and female authors/characters speak of different things in somewhat different ways. However, the operation of classifiers is predicated on detecting patterns of word usage which most distinguish groups and may bring to the forefront literary and linguistic elements which play a relatively minor role in the texts themselves. We have found similar results in other classification tasks, including gender mining in French literary works and *Encyclopédie* classifications.

Machine learning systems are best, in terms of various measures of accuracy, at binomial classification tasks, the dreaded "binary oppositions" of male/female, black/white and so forth, which have been the focus of much critical discussion in the humanities. Given the ability of statistical learners to find very thin slices of difference, it may be that any operation of any binary opposition may be tested and confirmed. If we ask for gender classification, the systems will do just that, return gender classifications. This suggests that certain types of hypothesis testing, particularly in regard to binary classifications, may show a successful result simply based on the framing of the question. It is furthermore unclear as to just what a successful classification means. If we identify gender or race of authors or characters, for example, at a better than 80% rate and generate a list of features most associated with both sides of the opposition, what does this tell us about the failed 20%? Are these errors to be corrected, presumably by improving classifiers or clustering models or should we further investigate these as interesting marginal instances? What may be considered a failure in computer science could be an interesting anomaly in the humanities.

Machine learning offers great promise to humanistic textual scholarship and the development of digital libraries. Using systems to sift through the ever increasing amounts of electronic texts to detect meaningful patterns offers the ability to frame new kinds of questions. But these technologies bring with them a set of assumptions and operations that should be subject to careful critical scrutiny. We in the digital humanities must do this critical work, relying on our understanding of epistemology and our technical skills to open the black box and shine light on what is inside. Deconstruction in the digital library should be a reading strategy not only for the texts found therein, but also of the systems being developed to manage, control and make the contents of electronic resources accessible and intelligible.

Bibliography

Allen, Timothy, Stéphane Douard, Charles Cooney, Russell Horton, Robert Morrissey, Mark Olsen, Glenn Roe, and Robert Voyer. "Plundering Philosophers: Identifying Sources of the *Encyclopédie* using the Vector Space Model" in preparation for *Text Technology.*

Argamon, Shlomo, Russell Horton, Mark Olsen, and Sterling Stuart Stein. "Gender, Race, and Nationality in Black Drama, 1850-2000: Mining Differences in Language Use in Authors and their Characters." *DH07*, Urbana-Champaign, Illinois. June 4-8, 2007.

Argamon, Shlomo, Jean-Baptiste Goulain, Russell Horton, and Mark Olsen. "Discourse, Power, and *Ecriture Féminine*: Text Mining Gender Difference in 18th and 19th Century French Literature." *DH07*, Urbana-Champaign, Illinois. June 4-8, 2007.

Bowker, Geoffrey C. and Susan Leigh Star. *Sorting Things Out: Classification and its Consequences*. Cambridge, MA: MIT Press, 1999.

Introna, L. and H. Nissenbaum. "Shaping the Web: Why the Politics of Search Engines Matters." *The Information Society*, 16(3): 1-17, 2000.

Witten, Ian H. and Eibe Frank. *Data Mining: Practical Machine Learning Tools and Techniques,* 2nd ed. San Francisco, CA: Morgan Kaufmann, 2005.

Feature Creep: Evaluating Feature Sets for Text Mining Literary Corpora.

Charles Cooney
cmcooney@diderot.uchicago.edu
University of Chicago, USA

Russell Horton
russ@diderot.uchicago.edu
University of Chicago, USA

Mark Olsen
mark@barkov.uchicago.edu
University of Chicago, USA

Glenn Roe
glenn@diderot.uchicago.edu
University of Chicago, USA

Robert Voyer
rlvoyer@diderot.uchicago.edu
University of Chicago, USA

Machine learning offers the tantalizing possibility of discovering meaningful patterns across large corpora of literary texts. While classifiers can generate potentially provocative hints which may lead to new critical interpretations, the features sets identified by these approaches tend not to be able to represent the complexity of an entire group of texts and often are too reductive to be intellectually satisfying. This paper describes our current work exploring ways to balance performance of machine learning systems and critical interpretation. Careful consideration of feature set design and selection may provide literary critics the cues that provoke readings of divergent classes of texts. In particular, we are looking at lexical groupings of feature sets that hold the promise to reveal the predominant "idea chunklets" of a corpus.

Text data mining and machine learning applications are dependent on the design and selection of feature sets. Feature sets in text mining may be described as structured data extracted or otherwise computed from running or unstructured text in documents which serves as the raw data for a particular classification task. Such features typically include words, lemmas, n-grams, parts of speech, phrases, named entities, or other actionable information computed from the contents of documents and/or associated metadata. Feature set selection is generally required in text mining in order to reduce the dimensionality of the "native feature space", which can range in the tens or hundreds thousands of words in even modest size corpora [Yang]. Not only do many widely used machine learning approaches become inefficient when using such high dimensional feature spaces, but such extensive feature sets may actually reduce performance of a classifier. [Witten 2005: 286-7] Li and Sun report that "many *irrelevant* terms have a detrimental effect on categorization accuracy due to *overfitting*" as well as tasks which "have many relevant but *redundant* features... also hurt categorization accuracy". [Li 2007]

Feature set design and selection in computer or information science is evaluated primarily in terms of classifier performance or measures of recall and precision in search tasks. But the features identified as most salient to a classification task may be of more interest than performance rates to textual scholars in the humanities as well as other disciplines. In a recent study of American Supreme Court documents, McIntosh [2007] refers to this approach as "Comparative Categorical Feature Analysis". In our previous work, we have used a variety of classifiers and functions implemented in PhiloMine [1] to examine the most distinctive words in comparisons of wide range of classes, such as author and character genders, time periods, author race and ethnicity, in a number of different text corpora [Argamon et. al. 2007a and 2007b]. This work suggests that using different kinds of features -- surface form words compared to bilemmas and bigrams -- allows for similar classifier performance on a selected task, while identifying intellectually distinct types of differences between the selected groups of documents.

Argamon, Horton et al. [Argamon 2007a] demonstrated that learners run on a corpus of plays by Black authors successfully classified texts by nationality of author, either American or non-American, at rates ranging between 85% and 92%. The features tend to describe gross literary distinctions between the two discourses reasonably well. Examination of the top 30 features is instructive.

American: ya', momma, gon', jones, sho, mississippi, dude, hallway, nothin, georgia, yo', naw, alabama, git, outta, y', downtown, colored, lawd, mon, punk, whiskey, county, tryin', runnin', jive, buddy, gal, gonna, funky

Non-American: na, learnt, don, goat, rubbish, eh, chief, elders, compound, custom, rude, blasted, quarrel, chop, wives, professor, goats, pat, corruption, cattle, hmm, priest, hunger, palace, forbid, warriors, princess, gods, abroad, politicians

Compared side by side, these lists of terms have a direct intuitive appeal. The American terms suggest a body of plays that deal with the Deep South (the state names), perhaps the migration of African-Americans to northern cities (hallway and downtown), and also contain idiomatic and slang speech (ya', gon', git, jive) and the language of racial distinction (colored). The non-American terms reveal, as one might expect, a completely different universe of traditional societies (chief, elders, custom) and life under colonial rule (professor, corruption, politicians). Yet a drawback to these features is that they have a stereotypical feel. Moreover, these lists of single terms reduce the many linguistically complex and varied works in a corpus to a distilled series of terms. While a group of words, in the form of a concordance, can show something quite concrete about a particular author's oeuvre or an individual play, it is difficult to come to a nuanced understanding of an entire

corpus through such a list, no matter how long. Intellectually, lists of single terms do not scale up to provide an adequate abstract picture of the concerns and ideas represented in a body of works.

Performing the same classification task using bilemmas (bigrams of word lemmas with function words removed) reveals both slightly better performance than surface words (89.6% cross validated) and a rather more specific set of highly ranked features. Running one's eye down this list is revealing:

American: yo_mama, white_folk, black_folk, ole_lady, st_louis, uncle_tom, rise_cross, color_folk,front_porch, jim_crow, sing_blue black_male, new_orleans, black_boy, cross_door, black_community, james_brown,

Non-American: palm_wine, market_place, dip_hand, cannot_afford, high_priest, piece_land, join_hand,bring_ water, cock_crow, voice_people, hope_nothing, pour_ libation, own_country, people_land, return_home

American (not present in non-American): color_boy, color_girl, jive_ass, folk_live

Here, we see many specific instances of African-American experience, community, and locations. Using bigrams instead of bilemmas delivers almost exactly the same classifier performance. However, not all works classified correctly using bilemmas are classified correctly using bigrams. Langston Hughes, *The Black Nativity,* for example, is correctly identified as American when using bigrams but incorrectly classified when using bilemmas. The most salient bigrams in the classification task are comparable, but not the same as bilemmas. The lemmas of "civil rights" and "human rights" do not appear in the top 200 bilemmas for either American or non-American features, but appear in bigrams, with "civil rights" as the 124th most predictive American feature and "human rights" as 111th among non-American features.

As the example of *The Black Nativity* illustrates, we have found that different feature sets give different results because, of course, using different feature sets means fundamentally changing the lexically based standards the classifier relies on to make its decision. Our tests have shown us that, for the scholar interested in examining feature sets, there is therefore no single, definitive feature set that provides a "best view" of the texts or the ideas in them. We will continue exploring feature set selection on a range of corpora representing different genres and eras, including Black Drama, French Women Writers, and a collection of American poetry. Keeping in mind the need to balance performance and intelligibility, we would like to see which combinations of features work best on poetry, for example, compared to dramatic writing. Text classifiers will always be judged primarily on how well they group similar text objects. Nevertheless, we think they can also be useful as discovery tools, allowing critics to find sets of ideas that are common to particular classes of texts.

Notes

1. See http://philologic.uchicago.edu/philomine/. We have largely completed work on PhiloMine2, which allows the user to perform classification tasks using a variety of features and filtering on entire documents or parts of documents. The features include words, lemmas, bigrams, bilemmas, trigrams and trilemmas which can be used in various combinations.

References

[Argamon 2007a] Argamon, Shlomo, Russell Horton, Mark Olsen, and Sterling Stuart Stein. "Gender, Race, and Nationality in Black Drama, 1850-2000: Mining Differences in Language Use in Authors and their Characters." *DH07*, Urbana-Champaign, Illinois. June 4-8, 2007.

[Argamon 2007b] Argamon, Shlomo, Jean-Baptiste Goulain, Russell Horton, and Mark Olsen. "Discourse, Power, and *Ecriture Féminine*: Text Mining Gender Difference in 18th and 19th Century French Literature." *DH07*, Urbana-Champaign, Illinois. June 4-8, 2007.

[Li 2007] Li, Jingyang and Maosong Sun, "Scalable Term Selection for Text Categorization", *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning* (EMNLP-CoNLL), pp. 774-782.

[McIntosh 2007] McIntosh, Wayne, "The Digital Docket: Categorical Feature Analysis and Legal Meme Tracking in the Supreme Court Corpus", *Chicago Colloquium on Digital Humanities and Computer Science,* Northwestern University, October 21-22, 2007

[Witten 2005] Witten, Ian H. and Eibe Frank. *Data Mining: Practical Machine Learning Tools and Techniques*, 2nd ed. San Francisco, CA: Morgan Kaufmann, 2005.

Yang, Yiming and Jan Pedersen, "A Comparative Study on Feature Selection in Text Categorization" In D. H. Fisher (ed.), *Proceedings of ICML-97, 14th International Conference on Machine Learning,* pp. 412-420, Nashville, US. Morgan Kaufmann Publishers, San Francisco, US.

Hidden Roads and Twisted Paths: Intertextual Discovery using Clusters, Classifications, and Similarities

Charles Cooney
cmcooney@diderot.uchicago.edu
University of Chicago, USA

Russell Horton
russ@diderot.uchicago.edu
University of Chicago, USA

Mark Olsen
mark@barkov.uchicago.edu
University of Chicago, USA

Glenn Roe
glenn@diderot.uchicago.edu
University of Chicago, USA

Robert Voyer
rlvoyer@diderot.uchicago.edu
University of Chicago, USA

While information retrieval (IR) and text analysis in the humanities may share many common algorithms and technologies, they diverge markedly in their primary objects of analysis, use of results, and objectives. IR is designed to find documents containing textual information bearing on a specified subject, frequently by constructing abstract representations of documents or "distancing" the reader from texts. Humanistic text analysis, on the other hand, is aimed primarily at enhancing understanding of textual information as a complement of "close reading" of relatively short passages. Interpreting a poem, a novel or a philosophical treatise in the context of large digital corpora is, we would argue, a constant interplay between the estranging technologies of machine learning and the direct reading of passages in a work. A significant element of interpreting or understanding a passage in a primary text is based on linking its specific elements to parts of other works, often within a particular temporal framework. To paraphrase Harold Bloom, 'understanding' in humanistic textual scholarship, 'is the art of knowing the hidden roads that go from text to text'. [1] Indeed, situating a passage of a text in a wider context of an author's *oeuvre*, time period, or even larger intellectual tradition, is one of the hallmarks of textual hermeneutics.

While finding the "hidden roads and twisted paths" between texts has always been subject to the limitations of human reading and recollection, machine learning and text mining offer the tantalizing prospect of making that search easier. Computers can sift through ever-growing collections of primary documents to help readers find meaningful patterns, guiding research and mitigating the frailties of human memory. We believe that a combination of supervised and unsupervised machine learning approaches can be integrated to propose various kinds of passages of potential interest based on the passage a reader is examining at a given moment, overcoming some limitations of traditional IR tools. One-dimensional measures of similarity, such as the single numerical score generated by a vector space model, fail to account for the diverse ways texts interact. Traditional 'ranked relevancy retrieval' models assume a single measure of relevance that can be expressed as an ordered list of documents, whereas real textual objects are composed of smaller divisions that each are relevant to other text objects in various complex ways. Our goal is to utilize the many machine learning tools available to create more sophisticated models of intertextual relation than the monolithic notion of "similarity."

We plan to adapt various ideas from information retrieval and machine learning for our use. Measures of document similarity form the basis of modern information retrieval systems, which use a variety of techniques to compare input search strings to document instances. In 1995, Singhal and Salton[2] proposed that vector space similarity measures may also help to identify related documents without human intervention such as human-embedded hypertext links. Work by James Allan[3] suggests the possibility of categorizing automatically generated links using innate characteristics of the text objects, for example by asserting the asymmetric relation "summary and expansion" based on the relative sizes of objects judged to be similar. Because we will operate not with a single similarity metric but with the results of multiple classifiers and clusterers, we can expand on this technique, categorizing intertextual links by feeding all of our data into a final voting mechanism such as a decision tree.

Our experiments with intertextual discovery began with using vector space calculations to try to identify that most direct of intertextual relationships, plagiarism or borrowing. Using the interactive vector space function in PhiloMine[4], we compared the 77,000 articles of the 18th century *Encyclopédie* of Diderot and d'Alembert to the 77,000 entries in a reference work contemporary to it, the *Dictionnaire universel françois et latin*, published by the Jesuits in the small town of Trévoux. The Jesuit defenders of the *Dictionnaire de Trévoux,* as it was popularly known, loudly accused the Encyclopédists of extensive plagiarism, a charge Diderot vigorously refuted, but which has never been systematically investigated by scholars. Our procedure is to compare all articles beginning with a specific letter in the *Encyclopédie* to all *Trévoux* articles beginning with the same letter. For each article in the *Encyclopédie*, the system displays each *Trévoux* article that scores above a user-designated similarity threshold. Human readers manually inspect possible matches, noting those that were probably plagiarized. Having completed 18 of 26 letters, we have found more than 2,000 articles (over 5% of those consulted) in the *Encyclopédie* were "borrowed" from the *Dictionnaire de Trévoux*, a surprisingly large proportion given the well-known antagonism between the Jesuits and Encyclopédists.[5]

The *Encyclopédie* experiment has shown us strengths and weaknesses of the vector space model on one specific kind of textual relationship, borrowing, and has spurred us to devise an expanded approach to find additional types of intertextual links. Vector space proves to be very effective at finding textual similarities indicative of borrowing, even in cases where significant differences occur between passages. However, vector space matching is not as effective at finding articles borrowed from the *Trévoux* that became parts of larger *Encyclopédie* articles, suggesting that we might profit from shrinking the size of our comparison objects, with paragraph-level objects being one obvious choice. In manually sifting through proposed borrowings, we also noticed articles that weren't linked by direct borrowing, but in other ways such as shared topic, tangential topic, expansion on an idea, differing take on the same theme, etc. We believe that some of these qualities may be captured by other machine learning techniques. Experiments with document clustering using packages such as CLUTO have shown promise in identifying text objects of similar topic, and we have had success using naive Bayesian classifiers to label texts by topic, authorial style and time period. Different feature sets also offer different insights, with part-of-speech tagging reducing features to a bare, structural minimum and N-gram features providing a more semantic perspective. Using clustering and classifiers operating on a variety of featuresets should improve the quality of proposed intertextual links as well as a providing a way to assign different types of relationships, rather than simply labeling two text objects as broadly similar.

To test our hypothesis, we will conduct experiments linking *Encyclopédie* articles to running text in other contemporaneous French literature and reference materials using the various techniques we have described, with an emphasis on intelligently synthesizing the results of various machine learning techniques to validate and characterize proposed linkages. We will create vector representations of surface form, lemma, and ngram feature sets for the *Encyclopédie* and the object texts as a pre-processing step before subjecting the data to clustering and categorization of several varieties. Models trained on the *Encyclopédie* will be used to classify and cluster running text, so that for each segment of text we will have a number of classifications and scores that show how related it is to various *Encyclopédie* articles and classes of articles. A decision tree will be trained to take into account all of the classifications and relatedness measures we have available, along with innate characteristics of each text object such as length, and determine whether a link should exist between two give text objects, and if so what kind of link. We believe a decision tree model is a good choice because such models excel at generating transparent classification procedures from low dimensionality data.

The toolbox that we have inherited or appropriated from information retrieval needs to be extended to address humanistic issues of intertextuality that are irreducible to single numerical scores or ranked lists of documents. Humanists know that texts, and parts of texts, participate in complex relationships of various kinds, far more nuanced than the reductionist concept of "similarity" that IR has generally adopted. Fortunately, we have a wide variety of machine learning tools at our disposal which can quantify different kinds of relatedness. By taking a broad view of all these measures, while looking narrowly at smaller segments of texts such as paragraphs, we endeavor to design a system that can propose specific kinds of lower-level intertextual relationships that more accurately reflect the richness and complexity of humanities texts. This kind of tool is necessary to aid the scholar in bridging the gap between the distant view required to manipulate our massive modern text repositories, and the traditional close, contextual reading that forms the backbone of humanistic textual study.

Notes

1. "Criticism is the art of knowing the hidden roads that go from poem to poem", Harold Bloom, "Interchapter: A Manifesto for Antithetical Criticism" in, *The Anxiety of Influence; A Theory of Poetry*, (Oxford University Press, New York, 1973)

2. Singhal, A. and Salton, G. "Automatic Text Browsing Using Vector Space Model" in *Proceedings of the Dual-Use Technologies and Applications Conference,* May 1995, 318-324.

3. Allan, James. "Automatic Hypertext Linking" in *Proc. 7th ACM Conference on Hypertext,* Washington DC, 1996, 42-52.

4. PhiloMine is the text mining extensions to PhiloLogic which the ARTFL Project released in Spring 2007. Documentation, source code, and many examples are available at http://philologic.uchicago.edu/philomine/ A word on current work (bigrams, better normalization, etc).

5. An article describing this work is in preparation for submission to *Text Technology.*

A novel way for the comparative analysis of adaptations based on vocabulary rich text segments: the assessment of Dan Brown's *The Da Vinci Code* and its translations

Maria Csernoch
mariacsernoch@hotmail.com
University of Debrecen, Hungary

In this work the appearance of the newly introduced words of *The Da Vinci Code* and its different translations have been analyzed and compared. The concept "newly introduced words" refers to the words whose first appearance is detected at a certain point of the text. In general the number of the newly introduced words follows a monotonic decay, however, there are segments of the texts where this descending is reversed and a sudden increase is detectable (Csernoch, 20066a, 2006b, 2007) – these text slices are referred to as vocabulary rich text slices. The question arises whether the detectable, unexpectedly high number of newly introduced words of the original work is traceable in the different translations of the text or not.

Before advancing on the project let us define the concept of translation. In this context the definition of Hatim and Mason (1993) is accepted, that is any adaptation of a literary work is considered as translation. Beyond the foreign language translations of *The Da Vinci Code* two more adaptations, the lemmatized and the condensed versions of the original work were analyzed.

The comparison of the newly introduced word types and lemmas of the different translations allows us to trace how precisely translation(s) follow(s) the changes in the vocabulary of the original text. Whether the vocabulary rich sections of the original text appear similarly rich in the translations, or the translations swallow them up, or the translations are richer at certain sections than the original work. In addition, could the changes in the newly introduced words, as an additional parameter to those already applied, be used to give a hint of the quality of a translation?

To carry out the analysis a previously introduced method was applied (Csernoch, 2006a, 2006b). The text was divided into constant-length intervals, blocks. The number of the newly introduced words was mapped to each block. Those text segments were considered vocabulary rich, in which the number of the newly introduced words significantly exceeds that predicted by a first-order statistical model.

The original *The Da Vinci Code*

In the original *The Da Vinci Code* eleven significant, vocabulary rich, text slices were detected. This number is similar to what was found in other, previously analyzed works (Csernoch, 2007). However, a further analysis of these text slices was applied to reveal their nature. Four parameters,

– the distribution of these text slices within the text,

– their length – the number of blocks in the significant text slice,

– their intensity – the local maximum of the relative number of the newly introduced words, and

– their content were determined.

The majority of the significant text slices of *The Da Vinci Code* both in lengths and intensity turned out to be unusually small, containing descriptions of events, characters and places. None of them stands for stylistic changes which usually (Baayen, 2001) trigger extremely vocabulary rich text slices. The distribution of them is uneven, they mainly appear in the first half of the text. This means that the second half of the novel hardly introduces any more vocabulary items that predicted by a first-order statistical model.

The analysis of the lemmatized texts

To see whether the word types, with all their suffixes, are responsible for any losses of vocabulary rich text slices, the lemmatization, as an adaptation of the text, was carried out. As it was found earlier (Csernoch, 2006b), the lemmatization of the text produced minor, if any, differences in the analysis of the word types in an English text. To the English non-lemmatized and lemmatized *The Da Vinci Code* their Hungarian correspondences were compared. Unlike the English texts, at the beginning of the Hungarian texts the number of newly introduced word types was so high that up to the first two hundred blocks (20,000 tokens), some of the text slices with significantly high number of newly introduced word types might be swallowed up.

The foreign language translations of *The Da Vinci Code*

While the absolute numbers of the newly introduced words of texts in different languages cannot, their relative numbers can be compared using the method outlined in Csernoch (2006a). Relying on the advantages of the method three different translations were compared to the original text, the Hungarian, the German, and the French. If the vocabulary rich text slices are found at the same positions both in the original and in the translated text, the translation can then be considered as exact in respect of vocabulary richness.

In the Hungarian translation the lengths and the intensities of the vocabulary rich text slices were not altered, however their distribution was more even than in the English text, and their number increased to sixteen. While the vocabulary rich text slices of the English text were all found in the Hungarian text, further such text slices were identified in the second half of the Hungarian text. The comparison revealed that the Hungarian text is richer in vocabulary than the English text.

The German translation replicated the vocabulary rich text slices of the original English text, and provided five more, which, similarly to the Hungarian translation, means that these text slices are richer in vocabulary than the corresponding text slices in the English text.

In the French translation only nine significant text slices were found. All of these text slices were only one block long, and their intensities were also surprisingly small. This means that the distribution of the newly introduced words in the French translation hardly differs from that predicted by the model. Furthermore, they are quite different in content from the vocabulary rich text slices in the other languages. Thus, there are hardly any concentrated, vocabulary rich text slices in the French text.

The condensed versions of *The Da Vinci Code*

Finally, the condensed versions of the English and Hungarian texts were compared to the corresponding full-length text and to each other. Several questions are to be answered in this context. By the nature of this adaptation it is obvious that the length of the text is curtailed to some extent. However, this parameter does not tell much about the nature of the condensation. We do not know from this parameter whether the condensation is only a cropping of the original text – certain text segments are left out while others are untouched – or the whole text is modified to some extent. If the text is modified, the percentage of the remaining tokens, word types, lemmas, and hapax legomena are parameters which tell us more about the condensation. To get a further insight it is worth considering how the vocabulary rich text segments of the original text are transferred into the condensed text. This last parameter might be a great help in deciding from which first order adaptation a second order adaptation of a text – in this case the condensed Hungarian text – is derived.

Both the English and the Hungarian condensed texts are 45% of the original texts in length. The number of word types is 64 and 55%, the number of lemmas are 64 and 61%, while the number of hapax legomena is 70 and 56% of the English and Hungarian full-length texts, respectively. These parameters indicate that the Hungarian condensed text bore more serious damage than the English did. The number of vocabulary rich text segments dropped to six – somewhat more than the half of original number – in the English text. On the other hand, the number of these text segments in the Hungarian text dropped to one-third, which is a notable difference compared to the full-length Hungarian text. Both in the condensed English and Hungarian texts the vocabulary rich segments were concentrated to the first half of the texts representing the same events, none of the segments unique to the full-length Hungarian text appeared in the condensed Hungarian text. The cumulative length of the vocabulary rich text slices dropped to 51% in the English and to 43% in the Hungarian text. Again, the Hungarian text seemed to be damaged to a greater extent. All the analyzed parameters thus clearly indicate that for the condensed Hungarian version the condensed English text was the direct source.

Bibliography

Baayen, R. H. (2001) *Word Frequency Distributions*. Kluwer Academic Publishers, Dordrecht, Netherlands

Csernoch, M. (2006a) The introduction of word types and lemmas in novels, short stories and their translations. http://www.allc-ach2006.colloques.paris-sorbone.fr/DHs.pdf. Digital Humanities 2006. The First International Conference of the Alliance of Digital Humanities Organisations. (5-9 July 2006, Paris)

Csernoch, M. (2006b) Frequency-based Dynamic Models for the Analysis of English and Hungarian Literary Works and Coursebooks for English as a Second Language. Teaching Mathematics and Computer Science. Debrecen, Hungary

Csernoch, M. (2007) *Seasonalities in the Introduction of Word-types in Literary Works.* Publicationes Universitatis Miskolcinensis, Sectio Philosophica, Tomus XI. – Fasciculus 3. Miskolc 2006-2007.

Hatim, B. and Mason, I. (1993) *Discourse and the Translator.* Longman Inc., New York.

Converting St Paul: A new TEI P5 edition of The Conversion of St Paul using stand-off linking

James C. Cummings
James.Cummings@oucs.ox.ac.uk
University of Oxford, UK

In researching the textual phenomena and scribal practices of late-medieval drama I have been creating an electronic edition of The Conversion of St Paul. This late-medieval verse play survives only in Bodleian MS Digby 133. My edition attempts to be a useful scholarly work, which where possible leverages existing resources to create an agile interoperable resource. In undertaking this work I have been able to explore a number of issues related to the creation of such editions which may have pedagogic benefit as a case study to others. These include shortcuts to speed up the creation of the initial edition, the generation of additional resources based solely on a single XML encoded text, the use of new mechanisms in the TEI P5 Guidelines, and the exploitation of stand-off markup to experiment with seamlessly incorporating external resources. The TEI P5 Guidelines have been followed in production of this edition, and specifically I have used a number of features which are new additions to TEI P5 and which others may not yet be familiar. It was developed in tandem with early drafts of the Guidelines, in part to test some of the features we were adding.

These include:

• A customised view of the TEI expressed as a TEI ODD file. This allows generation not only of constrained TEI Schemas and DTDs but also project specific documentation through the TEI's ODD processor 'Roma'.

• A manuscript description of MS Digby 133, and the manuscript item of this play, using the TEI's new module for manuscript description metadata.

• Consistent and in-depth use of the new <choice> structure to provide alternative textual information at individual points in the text. Specifically, in this highly-abbreviated medieval text, this has been used to provide both abbreviations and expansions which then can be toggled in the rendered version. Regularisation of medieval spelling has been handled with stand-off markup, but could equally have been incorporated into <choice>

• Inside abbreviations and expansions, the new elements <am> (abbreviation marker) and <ex> (expanded text) have been used. This allows the marking and subsequent display of abbreviation marks in a diplomatic edition view and italicised rendering of the supplied text in the expanded view.

• The edition also records information about the digital photographic surrogate provided by the Bodleian using the new <facsimile> element. While this will also allow linking from the text to particular zones of the images, this has not yet been undertaken..

• The edition also uses various new URI-based pointing mechanisms and datatype constraints new to TEI P5.

To speed up the creation of the edition I took an out of copyright printed version of the text (Furnivall 1896) which was scanned and passed through optical character recognition. This was then carefully corrected and proofread letter-by-letter against freely available (though restrictively licensed) images made available online by the Bodleian Library, Oxford (at http://image.ox.ac.uk/show?collection=bodleian&manuscript=msdigby133). I used OpenOffice to create the initial edition, with specialised formatting used to indicate various textual and editorial phenomena such as expanded material, superscript abbreviation markers, stage directions, and notes. The up-scaling of the markup through using this presentational markup was achieved when I converted it, using XSLT, to very basic TEI XML. While this is a quick method of data entry familiar to many projects, it only tends to work successfully in non-collaborative projects where the consistency of the application of formatting can be more easily controlled as any inconsistencies can lead to significant manual correction of the generated XML.

Another of the issues I was interested in exploring in this edition was the use of stand-off markup to create interoperable secondary resources and how this might effect the nature of scholarly editing. While I could have stored much of this information I needed in the edition itself, I wanted to experiment with storing it in external files and linking them together by pointing into the digital objects. The motivation for this comes from a desire to explore notions of interoperability since stand-off markup methodology usually leaves the base text untouched and stores additional information in separate files. As greater numbers of good scholarly academic resources increasingly become available in XML, the pointing into a number of resources, and combining these together to form an additional greater resource is becoming more common. Stand-off markup was used here partly to experiment with the idea of creating an interoperable flexible resource, that is an 'agile edition'. For example, an edition can be combined with associated images, a glossary or word list, or other external resources such as dictionaries. In the case of this edition, I generated a word list (encoded using the TEI dictionaries module) using XSLT. The word list included any distinct orthographic variants in the edition. This was based on a 'deep-equals' comparison which compared not only the spelling of words, but all of their descendant elements, and thus captured differences in abbreviation/expansion inside individual words. The location of individual instances of orthographic variance in the edition could easily have been stored along with the entry in the word list. However, since part of the point was to experiment with handling stand-off markup, I stored these in a

third file whose only purpose was to record <link> elements pointing both to an entry in the word list and every single instance of this word in the edition. This linking was done using automatically-generated xml:id attributes on each word and word list entry. This enables a number of usability features. The clicking on any individual word in the edition takes you to its corresponding entry in the word list. From any entry in the word list you can similarly get back to any other individual instance of that word in the edition. Moreover the word list entry also contains an optionally displayed concordance of that word to allow easy comparison of its use in context.

In addition to using resources created by myself, it was a desired aim of this investigation into stand-off markup to use external resources. The most appropriate freely-available resource in this case is the Middle English Dictionary (MED), created by the University of Michigan. As this scholarly edition was being created in my spare time, I did not want to exhaustively check orthographic words in my edition against the MED and link directly to the correct senses. While that is certainly possible, and should be the recommended text-critical practice, it would take a significant amount of time and be prone to error. Instead I desired to pass a regularised form of the word to the MED headword search engine, and retrieve the results and incorporate them dynamically into the display of the entry for that word in the word list. However, this proved impossible to do from the MED website because their output, despite claiming to be XHTML, was not well-formed. Luckily, they were willing to supply me with an underlying XML file which provided not only headwords, but also their various different orthographic forms and the MED id number to which I could link directly. Thus, I was able to achieve the same effect as transcluding the MED search results by reimplementing the functionality of their search directly in my XSLT and thus providing pre-generated links in each entry to possible headwords in the MED. While successful for my resource, in terms of true interoperability it is really a failure, one which helps to highlight some of the problems encountered when pointing into resources over which you have no control.

The proposed paper will describe the process of creation of the edition, the benefits and drawbacks of using stand-off markup in this manner, its linking to external resources, and how the same processes might be used in either legacy data migration or the creation of new editions. One of the concluding arguments of the paper is that the advent of new technologies which make the longpromised ability for the interoperability of resources that much easier, also encourages (and is dependent upon) us making our own existing materials accessible in a compatible manner.

Bibliography

Baker, Donald C., John L. Murphy, and Louis B. Hall, Jr. (eds) *Late Medieval Religious Plays of Bodleian MSS Digby 133 and E Museo 160*, EETS, 283 (Oxford: Oxford Univ. Press, 1982)

Eggert, Paul. 'Text-encoding, theories of the text and the "work-site"'. *Literary and Linguistic Computing*, (2005), 20:4, 425-435

Furnivall, F. J. (ed.) *The Digby Plays*, (New Shakespeare Society Publications, 1882) Re-issued for EETS Extra Series LXX, (London: EETS, 1896)

Robinson, P. 'The one and the many text', *Literary and Linguistic Computing*, (2000), 15:1, 5-14.

Robinson, P. 'Where We Are with Electronic Scholarly Editions, and Where We Want to Be', *Jahrbuch für Computerphilologie*, 5 (2003), 123-143.

Robinson, P. 'Current issues in making digital editions of medieval texts or, do electronic scholarly editions have a future?', *Digital Medievalist*, (2005), 1:1, Retrieved 1 Nov. 2007 <http://www.digitalmedievalist.org/journal/1.1/robinson/>

Sperberg-McQueen, C. M. *Textual Criticism and the Text Encoding Initiative*. Annual Convention of the Modern Language Association, 1994. reprinted in Finneran, R. J. (Ed.) (1996) The Literary Text in the Digital Age. Ann Arbor: University of Michigan Press. 37-62

TEI Consortium, eds. *TEI P5: Guidelines for Electronic Text Encoding and Interchange*. Retrieved 1 Nov. 2007 <http://www.tei-c.org/release/doc/tei-p5-doc/en/html/index.html>

Unsworth, J. *Electronic Textual Editing and the TEI*. Annual Convention of the Modern Language Association, 2002. Retrieved 1 Nov. 2007. <http://www3.isrl.uiuc.edu/~unsworth/mla-cse.2002.html>.

Vanhoutte, E. 'Prose fiction and modern manuscripts: limitations and possibilities of text-encoding for electronic editions'. In J. Unsworth, K. O'Keefe, and L. Burnard (Eds). *Electronic Textual Editing*. (New York: Modern Language Association of America, 2006), p. 161-180.

ENRICHing Manuscript Descriptions with TEI P5

James C. Cummings
James.Cummings@oucs.ox.ac.uk
University of Oxford, UK

ENRICH is a pan-European eContent+ project led by the Czech National Library, starting 1 December 2007, which is creating a base for the European digital library of cultural heritage (manuscript, incunabula, early printed books, and archival papers) by the integration of existing but scattered electronic content within the Manuscriptorium digital library through the use of the metadata enrichment and coordination between heterogeneous metadata and data standards. The consortium brings together critical mass of content, because the project groups together the three richest owners of digitized manuscripts among national libraries in Europe (Czech Republic, Iceland and Serbia as associated partners); ENRICH partner libraries possess almost 85% currently digitized manuscripts in the national libraries in Europe, which will be enhanced by substantial amount of data from university libraries and other types of institutions in terms of several hundreds of thousands of digitized pages plus hundreds of thousands of pages digitized by the partners through the European Library data collections. When the project has finished the consortium will make available more than an estimated 5,076,000 digitized pages.

ENRICH <http://enrich.manuscriptorium.com/> builds upon the existing Manuscriptorium platform <http://www.manuscriptorium.com/> and is adapting it to the needs of those organizations holding repositories of manuscripts. The principle of integration is centralisation of the metadata (descriptive evidence records) within the Manuscriptorium digital library and distribution of data (other connected digital documents) among other resources within the virtual net environment. That is, the metadata of the manuscript descriptions is aggregated in a single place and format to assist with searching these disparate repositories, but the images and original metadata records remain with the resource-holding institutions.

ENRICH target groups are content owners/holders, Libraries, museums and archives, researchers & students, policy makers and general interest users. The project allows them to search and access documents which would otherwise be hardly accessible by providing free access to almost all digitized manuscripts in Europe. Besides images it also offers access to TEI-structured historical full texts, research resources, other types of illustrative data (audio and video files) or large images of historical maps. The ENRICH consortium is closely cooperating with TEL (The European Library) and will become a component part of the European Digital Library when this becomes reality.

Institutions involved in the ENRICH project include:

- National Library of Czech Republic (Czech Republic)
- Cross Czech a.s. (Czech Republic)
- AiP Beroun s r.o. (Czech Republic)
- Oxford University Computing Services (United Kingdom)
- Københavns Universitet - Nordisk Foskningsinstitut (Denmark)
- Biblioteca Nazionale Centrale di Firenze National Library in Florence (Italy)
- Università degli Studi di Firenze - Centro per la comunicazione e l integrazione dei media Media Integration and Communicaiton Centre Firenze(Italy)
- Institute of mathematics and informatics (Lithuania)
- University Library Vilnius (Lithuania)
- SYSTRAN S.A. (France)
- University Library Wroclaw (Poland)
- Stofnun Árna Magnússonar í íslenskum fræðum (Iceland)
- Computer Science for the Humanities - Universität zu Köln (Germany)
- St. Pölten Diocese Archive (Austria)
- The National and University Library of Iceland (Iceland)
- Biblioteca Nacional de Espana - The National Library of Spain (Spain)
- The Budapest University of Technology and Economics (Hungary)
- Poznan Supercomputing and Networking Center (Poland)

Manuscriptorium is currently searchable via OAI-PMH from the TEL portal, this means that any ENRICH full or associated partner automatically enriches the European Digital Library. The main quality of ENRICH and Manuscriptorium is the application of homogeneous and seamless access to widespread resources including access to images from the distributed environment under a single common interface. ENRICH supports several levels of communication with remote digital resources, ranging from OAI harvesting of partner libraries to full integration of their resources into Manuscriptorium. Furthermore, ENRICH has developed free tools to assist in producing XML structured digitized historical documents, and these are already available and ready to use. However, the existing XML schema reflects the results of the now-dated EU MASTER project for TEI-based manuscript descriptions. It also incorporates other modern content standards, especially in the imaging area. It is this schema that is being updated to reflect the developments in manuscript description available in TEI P5. The internal Manuscriptorium format is based on METS containerization of the schema and related parallel descriptions which enables a flexible approach needed by the disparate practices of researchers in this field.

The Research Technology Services section of the Oxford University Computing Services is leading the crucial work-package on the standardization of shared metadata. In addition to a general introduction to the project it is the progress of this work-package which the proposed paper will discuss. This work-package is creating a formal TEI specification, based on the TEI P5 Guidelines, for the schema to be used to describe

manuscripts managed within Manuscriptorium. The use of this specification enables automatic generation of reference documentation in different languages and the creation of a formal DTD or Schemas as well as formal DTD or Schemas. A suite of tools is also being developed to convert automatically existing sets of manuscript descriptions, where this is feasible, and to provide simple methods of making them conformant to the new standard where it is not. These tools are being developed and validated against the large existing base of adopters of the Master standard and will be distributed free of charge by the TEI.

The proposed paper will report on the development and validation of a TEI conformant specification for the existing Manuscriptorium schema using the TEI P5 specification language (ODD). This involved a detailed examination of the current schema and documentation developed for the existing Manuscriptorium repository and its replacement by a formal TEI specification. This specification will continue to be further enhanced in light of the needs identified by project participants and the wider MASTER community to form the basis of a new schema and documentation suite. The paper will report on the project's production of translations for the documentation in at least English, French, Czech, German, as well as schemas to implement it both as DTD and RELAXNG. The production of these schemas and the translated documentation are produced automatically from the TEI ODD file.

A meeting of representatives from other European institutions who have previously used the MASTER schema has been organized where the differences in how they have used the schema will have been explored, along with their current practice for manuscript description. The aim is to validate both the coverage of the new specification and the feasibility and ease of automatic conversion towards it. The outputs from this activity will include a report on any significant divergence of practice amongst the sample data sets investigated. The ODD specification will continue to be revised as necessary based on the knowledge gained from the consultation with other MASTER users. This will help to create an enriched version of the preliminary specifications produced. Finally, software tools are being developed to assist in conversion of sets of records produced for earlier MASTER specifications, and perhaps some others, to the new TEI P5 conformant schema. These tools are being tested against the collection of datasets gained from the participants of the meeting with other MASTER users, but also more widely within the TEI community. OUCS is also preparing tutorial material and discussion papers on the best practice to assist other institutions with migration existing MASTER material to the new standard. In this subtask ENRICH is cooperating with a broader TEI-managed effort towards the creation of TEI P5 migration documentation and resources.

An OAI-PMH harvester is being implemented and incorporated into Manuscriptorium. The first step is to ensure that OAI/PMH metadata is available for harvesting from all the resources managed within Manuscriptorium. Appropriate software tools to perform this harvesting are also being developed. Eventually, the internal environment of Manuscriptorium will be enhanced through implementation of METS containerization of the Manuscriptorium Scheme. This will involve an assessment of the respective roles of the TEI Header for manuscript description and of a METSconformant resource description and will enable different kinds of access to the resources within the Manuscriptorium. This will help to demonstrate for others the interoperability of these two important standards, and in particular where their facilities are complementary.

Improvement and generalization of Unicode treatment in Manuscriptorium is the final part of the OUCS-led work package. As Manuscriptorium is basically an XML system, all the data managed is necessarily represented in Unicode. This could cause problems for materials using non-standard character encodings, for example where manuscript descriptions quote from ancient scripts and include glyphs not yet part of Unicode. The TEI recommendations for the representation of nonstandard scripts are being used within ENRICH project which is producing a suite of non-standard character and glyph descriptions appropriate to the project's needs.

The proposed paper is intended as a report on the work done in the conversion and rationalization of manuscript metadata across a large number of archives with disparate practices. While it will introduce the project to delegates at Digital Humanities 2008, it will concentrate on reporting the the problems and successes encountered in the course of these aspects of project. Although the overall project will not be finished by the time of the conference, the majority of the work in developing a suite of conversion tools will be complete by this time and the paper will focus on this work. As such, although it will detail work done by the author, it will rely on work by project partners who although not listed as authors here will be briefly acknowledged in the paper where appropriate.

Editio ex machina - Digital Scholarly Editions out of the Box

Alexander Czmiel
alexander@czmiel.de
Berlin-Brandenburg Academy of Sciences and Humanities, Germany

For the most part, digital scholarly editions have been historically grown constructs. In most cases, they are oriented toward print editions, especially if they are retro-digitized. But even "digital-born" editions often follow the same conventions as printed books. A further problem is that editors start to work on a scholarly edition without enough previous in-depth analysis about how to structure information for electronic research. In the course of editing and the collection of data, the requirements and wishes of how the edited texts should be presented frequently changes, especially when the editor does not have a chance to "see" the edition before it is complete. Usually all the data is collected, the text is edited and the last step is to think about the different presentation formats.

One of the first steps in the production of a digital scholarly edition should be to analyze what kind of information might be of interest to a broad audience, and how this should be structured so that it can be searched and displayed appropriately. The crucial point in the process of designing the data structure should be that different scholars have different intellectual requirements from resources. They are not always happy with how editors organize scholarly editions.

Michael Sperberg-McQueen demanded in 1994, "Electronic scholarly editions should be accessible to the broadest audience possible."[1] However, current scholarly editions are produced for a certain audience with very specialized research interests. The great potential of a digital scholarly edition is that it can be designed flexibly enough to meet the demands of people with many different viewpoints and interests. So why not let the audience make the decision as to which information is relevant and which is not? Some information might be of more interest, other information of less interest or could be entirely ignorable.

Imagine a critical scholarly edition about medicine in ancient Greece provided by editors with a philological background. A philologist has different needs relating to this edition than, e.g., a medical historian, who might not be able to read Greek and might be only interested in reading about ancient medical practices. Who knows what kinds of annotations within the text are of interest for the recipients?

The digital world provides us with many possibilities to improve scholarly editions, such as almost unlimited space to give complete information, more flexibility in organizing and presenting information, querying information, instant feedback and a range of other features.

We have to think about how we can use these benefits to establish a (perhaps formalized) workflow to give scholars the chance to validate their work while it is in progress and to present their work in progress in an appropriate way for discussion within a community. This would not just validate the technical form of the work, but also improve the quality of the content, often due to ongoing feedback, and improve the intellectual benefits.

On the other hand, digital editions should be brought online in an easy way without weeks of planning and months of developing software that may fit the basic needs of the editors but in most cases is just re-inventing the wheel. If needed, digitized images should be included easily and must be citeable. As it might be about work in progress, all stages of work must be saved and documented. Therefore, a versioning system is needed that allows referencing of all work steps.

Finally, it is necessary that the scholar is able to check his or her work viewed in various media by the click of a button - for example, how the edition looks like on screen or printed, or even with different layouts or website designs.

What is the potential of such a system? It offers an easy way to present the current state of a work in progress. Scholars can check their work at any time. But this is not useful without modifications for the special needs of certain digital editions and special navigation issues. Nevertheless, it can be used as base system extensible by own scripts, which implement the needs of a concrete project.

And last but not least, such a system should offer the possibility of re-using data that is based on the same standard for other projects. This is especially true for more formalized data, such as biographical or bibliographical information, which could be used across different projects that concern the same issue or the same period.

This paper will give a practical example of what an architecture that follows the aforementioned demands would look like. This architecture gives scholars the possibility of producing a scholarly edition using open standards. The texts can be encoded directly in XML or using WYSIWYG-like methods, such as possible with the oXygen XML editor or WORD XML exports.

The "Scalable Architecture for Digital Editions" (SADE) developed at the Berlin-Brandenburg Academy of Sciences and Humanities is a modular and freely scalable system that can be adapted by different projects that follow the guidelines of the Text Encoding Initiative (TEI)[2] or easily use other XML standards as input formats. Using the TEI is more convenient, as less work in the modification of the existing XQuery and XSLT scripts needs to be done.

Scalability of SADE relates to the server side as well as to the browser side. Browser-side scalability is equivalent to the users' needs or research interests. The user is able to arrange the information output as he or she likes. Information can be switched on or off. The technological base for this functionality is AJAX or the so-called Web 2.0 technologies.

Server-side scalability is everything that has to do with querying the database and transforming the query results into HTML or PDF. As eXist[3], the database we use, is a native XML database, the whole work can be done by XML-related technologies such as XQuery and XSLT. These scripts can be adapted with less effort to most projects' needs.

For the connection between text and digitized facsimile, SADE uses Digilib[4], an open-source software tool jointly developed by the Max-Planck-Insitute for the History of Science, the University of Bern and others. Digilib is not just a tool for displaying images, but rather a tool that provides basic image editing functions and the capability of marking certain points in the image for citation. The versioning system at the moment is still a work in progress, but will be available by conference time.

Documents can be queried in several ways - on the one hand with a standard full text search in texts written in Latin or Greek letters, and on the other hand by using a more complex interface to query structured elements, such as paragraphs, names, places, the apparatus, etc. These query options are provided by SADE. Furthermore, all XML documents are available not rendered in raw XML format, and so can be integrated in different projects rendered in a different way.

SADE could be the next step in improving the acceptance of digital editions. Texts are accessible in several ways. The editor decisions are transparent and comprehensible at every stage of work, which is most important for the intellectual integrity of a scholarly edition. The digitized facsimiles can be referenced and therefore be discussed scientifically. The database back end is easy to handle and easy to adapt to most projects' needs.

A working example, which is work in progress and extended continually, is the website of the "Corpus Medicorum Graecorum / Latinorum"[5] long term academy project at the Berlin-Brandenburg Academy of Sciences and Humanities[6]. This website exemplifies how a recipient can organize which information is relevant for his or her information retrieval. Other examples can be found at http://pom.bbaw.de.

References

[1] Cited in http://www.iath.virginia.edu/~jmu2m/mla-cse.2002.html

[2] http://www.tei-c.org/index.xml

[3] http://www.exist-db.org/

[4] http://digilib.berlios.de/

[5] http://pom.bbaw.de/cmg/

[6] http://www.bbaw.de/

Bibliography

Burnard, L.; Bauman, S. (ed.) (2007): *TEI P5: Guidelines for Electronic Text Encoding and Interchange*. <http://www.tei-c.org/release/doc/tei-p5-doc/en/html/index.html>

Buzzetti, Dino; McGann, Jerome (2005): Critical Editing in a Digital Horizon. In: John Unsworth, Katharine O'Brien O'Keeffe u. Lou Burnard (Hg.): *Electronic Textual Editing*. 37–50.

Czmiel, A.; Fritze, C.; Neumann, G. (2007): Mehr XML – Die digitalen Projekte an der Berlin-Brandenburgischen Akademie der Wissenschaften. In: *Jahrbuch für Computerphilologie -online*, 101-125. <http://computerphilologie.tu-darmstadt.de/jg06/czfrineu.html>

Dahlström, Mats (2004): How Reproductive is a Scholarly Edition? In: *Literary and Linguistic Computing* 19/1, S. 17–33.

Robinson, Peter (2002): What is a Critical Digital Edition? In: *Variants – Journal of the European Society for Textual Scholarship* 1, 43–62.

Shillingsburg, Peter L. (1996): Principles for Electronic Archives, Scholarly Editions, and Tutorials. In: Richard Finneran (ed.): *The Literary Text in the Digital Age*. 23–35.

Talia: a Research and Publishing Environment for Philosophy Scholars

Stefano David
sdavid@delicias.dia.fi.upm.es
DEIT - Università Politecnica delle Marche, Italy

Michele Nucci
mik.nucci@gmail.com
DEIT - Università Politecnica delle Marche, Italy

Francesco Piazza
f.piazza@univpm.it
DEIT - Università Politecnica delle Marche, Italy

Introduction

Talia is a distributed semantic digital library and publishing system, which is specifically designed for the needs of the scholarly research in the humanities. The system is designed for the diverging needs of heterogeneous communities. It is strictly based on Semantic Web technology[1] and computational ontologies for the organisation of knowledge and will help the definition of a state-ofthe-art research and publishing environment for the humanities in general and for philosophy in particular.

Talia is developed within the Discovery project, which aims at easing the work of philosophy scholars and researchers, making a federation of distributed digital archives (nodes) each of them dedicated to a specific philosopher: Pre-Socratic, Diogenes Laertius, Leibniz, Nietzsche, Wittgenstein, Spinoza, and others.

Use Case and Requirements

One of the use cases that Talia should satisfy is called *The Manuscripts Comparer*. The archive contains digital reproductions of handwritten manuscripts by one philosopher. The scholar wants to browse them like books, navigating by chapters and pages. The archive also contains transcriptions of the different paragraphs, on which to click to see the correct transcription.

Moreover, the scholar is interested in comparing different versions of the same thought, placed by the philosopher in different paragraphs of his works, so the scholar needs to find all those different versions. Finally, he/she can write his own remarks on those paragraphs, so that they are published in the archive, after a peer review process.

From this use case, we can devise some technical requirements that drive the design of Talia:

1. *Making metadata remotely available.* The metadata content of the archives shall be made available through interfaces similar to those of URIQA[2], which will allow automated agents and clients to connect to a Talia node and ask for metadata about a resource of interest to retrieve its description and related metadata (e.g., the author, the resource type, etc.). This type of services enable numerous types of digital contents reuse.

2. *Querying the system using standard query-languages.* It shall possible to directly query the distributed archives using the SPARQL Semantic Web Query Language. SPARQL provides a powerful and standard way to query RDF repositories but also enables merging of remote datasets.

3. *Transformation of encoded digital contents into RDF.* The semantic software layer shall provide tools to transform textually encoded material, for example the transcription of manuscripts in TEI[3] format, into RDF.

4. *Managing structural metadata.* "Structural" metadata describes the structure of a document (e.g., its format, its publisher, etc.). It shall be necessary to provide tools to manage these kinds of metadata.

5. *Import and exporting RDF data.* It shall be important to include facilities in Talia to import and export RDF data with standard and well know formats like RDF/XML and N-Triples.

Talia should also provide facilities to enrich the content of the archives with metadata, whose types and formats depend on the kind of source and on the user community which works on the data. For example, in some archives, different versions of the same paragraph may exist in different documents. In order to allow the users to follow the evolution of that particular philosophical thought, the relations among these different versions must be captured by the system.

An Overview of the Talia System

Talia is a distributed semantic digital library system which combines the features of digital archives management with an on-line peer-review system.

The Talia platform stores digital objects, called *sources*, which are identified by their unique and stable URI. Each source represents either a work of a philosopher or a fragment of it (e.g., a paragraph), and can have one or more data files (e.g., images or textual documents) connected to it. The system also stores information about the sources, which can never be removed once published and are maintained in a fixed state. Combined with other long-term preservation techniques, Talia allows the scholars to reference their works and gives the research community immediate access to new content.

The most innovative aspect of Talia is that for the first time, at least in the field of humanities, all the interfaces exposed publicly will be based on proven Semantic Web standards enabling data interoperability within the Talia federation, and eases the data interchange with other systems. Two additional features of Talia are the highly customisable Graphic User Interface (GUI) and the dynamic contextualisation.

Figure 1: An Example of the Talia User Interface.

The web interface framework, based on modular elements called *widgets*, allows to build GUIs according to requirements coming from heterogeneous communities. Widgets can be packaged independently and used as building blocks for the application's user interface. In particular, Talia provides semantic widgets that interact with an RDF Knowledge Base. To customise the site's appearance, it also would be possible to add custom HTML rendering templates. Figure 1 shows a screenshot of a Talia's GUI.

The *dynamic contextualisation* provides a means for data exchange among different and distributed digital libraries based on the Talia framework. The dynamic contextualisation allows a Talia library (node) to share parts of its RDF graph in a peer-to-peer network. By using this mechanism, a node may notify another node that a semantic link between them exists, and the other node may use this information to update its own RDF graph and create a bidirectional connection.

Talia uses computational ontologies and Semantic Web technology to help the creation of a state-of-the-art research and publishing environment. Talia will provide an innovative and adaptable system to enable and ease data interoperability and new paradigms for information enrichment, dataretrieval, and navigation.

Computational Ontologies

Ontologies have become popular in computer science as a means for the organisation of information. This connotation of *ontology* differs from the traditional use and meaning it has in philosophy, where ontologies are considered as A system of categories accounting for a certain vision of the world [2]

In computer science, the concept of (computational) ontology evolved from the one first provided by Gruber [3], who defined an ontology as a specification of a conceptualisation4 to a more precise one, extracted from Guarino's definitions [4]: A computational ontology is A formal, partial specification of a shared conceptualisation of a world (domain).

Intuitively, a computational ontology is a set of assumptions that define the structure of a given domain of interest (e.g., philosophy), allowing different people to use the same concepts to describe that domain. Talia will use computational ontologies to organise information about writings of a philosopher or documents (manuscripts, essays, theses, and so on) of authors concerning that philosopher.

Related Work

Talia is directly related to the Hyper Platform which was used for the HyperNietzsche archive [1], a specialised solution, designed for specific needs of the Nietzsche communities. HyperNietzsche has a fixed graphical user interface, it does not use Semantic Web technology and it is not adaptable for different communities with heterogeneous needs.

Talia shares some properties with other semantic digital library systems like JeromeDL [5], BRICKS [6], and Fedora [7]. However, these projects are mostly focused on the back-end technology and none of them offers a flexible and highly customisable research and publishing system like Talia.

Conclusion and Future Work

Talia is a novel semantic digital web library system, which aims to improve scholarly research in the humanities, in particular in the field of philosophy. Thanks to the use of Semantic Web technology, Talia represents a very adaptable state-of-the art research and publishing system.

The ontologies used in Talia are currently being developed by Discovery's content partners, who are in charge of organising their contributions. These ontologies will be part of the final core Talia application. At the moment, only a first public demo version is available[5].

Although Talia is currently intended only for philosophy scholars, it should be straightforward to adopt it for humanities, with the help of suitably developed ontologies. Using dynamic Ruby as programming language and the RubyOnRails framework, Talia provides an ideal framework for the rapid development of customised semantic digital libraries for the humanities.

Acknowledgements

This work has been supported by Discovery, an ECP 2005 CULT 038206 project under the EC eContentplus programme.

We thank Daniel Hahn and Michele Barbera for the fruitful contributions to this paper.

Notes

1 http://www.w3.org/2001/sw/

2 http://sw.nokia.com/uriqa/URIQA.html

3 http://www.tei-c.org/

4 A more up-to-date definition and additional readings can be found at http://tomgruber.org/writing/ontologydefinition-2007.htm

5 http://demo.talia.discovery-project.eu

References

[1] D'Iorio, P.: Nietzsche on new paths: The Hypernietzsche project and open scholarship on the web. In Fornari, C., Franzese, S., eds.: *Friedrich Nietzsche. Edizioni e interpretazioni.* Edizioni ETS, Pisa (2007)

[2] Calvanese, D., Guarino, N.: Ontologies and Description Logics. *Intelligenza Artificiale - The Journal of the Italian Association for Artificial Intelligence,* 3(1/2) (2006)

[3] Gruber, T.R.: A translation approach to portable ontology specifications. *Knowledge Acquisition* 5(2) (1993) 199-220

[4] Guarino, N.: Formal Ontology and Information Systems. In Guarino, N., ed.: *1st International Conference on Formal Ontologies in Information Systems,* IOS Press (1998) 3-15

[5] Kurk S., Woroniecki T., Gzella A., Dabrowski M. McDaniel B.: Anatomy of a social semantic library. In: *European Semantic Web Conference. Volume Semantic Digital Library Tutorial.* (2007).

[6] Risse T., Knezevic P., Meghini C., Hecht R., Basile F.: The bricks infrastructure an overview. In: *The International Conference EVA*, Moscow, 2005.

[7] Fedora Development Team: Fedora Open Source Repository software. White Paper.

Bootstrapping Classical Greek Morphology

Helma Dik
helmadik@mac.com
University of Chicago, USA

Richard Whaling
rwhaling@uchicago.edu
University of Chicago, USA

In this paper we report on an incremental approach to automated tagging of Greek morphology using a range of already existing tools and data. We describe how we engineered a system that combines the many freely available resources into a useful whole for the purpose of building a searchable database of morphologically tagged classical Greek.

The current state of the art in electronic tools for classical Greek morphology is represented by Morpheus, the morphological analyzer developed by Gregory Crane (Crane 1991). It provides all possible parses for a given surface form, and the lemmas from which these are derived. The rich morphology of Greek, however, results in multiple parses for more than 50% of the words (http://grade-devel.uchicago.edu/morphstats.html). There are fully tagged corpora available for pre-classical Greek (early Greek epic, developed for the 'Chicago Homer', http://www.library.northwestern.edu/homer/) and for New Testament Greek, but not for the classical period.

Disambiguating more than half of our 3 million-word corpus by hand is not feasible, so we turned to other methods. The central element of our approach has been Helmut Schmid's TreeTagger (Schmid 1994, 1995). TreeTagger is a Markov-model based morphological tagger that has been successfully applied to a wide variety of languages. Given training data of 20,000 words and a lexicon of 350,000 words, TreeTagger achieved accuracy on a German news corpus of 97.5%. When TreeTagger encounters a form, it will look it up in three places: first, it has a lexicon of known forms and their tags. Second, it builds from that lexicon a suffix and prefix lexicon that attempts to serve as a a morphology of the language, so as to parse unknown words. In the absence of a known form or recognizable suffix or prefix, or when there are multiple ambiguous parses, it will estimate the tag probabilistically, based on the tags of the previous *n* (typically two) words; this stochastic model of syntax is stored as a decision tree extracted from the tagged training data.

Since classical Greek presents, prima facie, more of a challenge than German, given that it has a richer morphology, and is a non-projective language with a complex syntax, we were initially unsure whether a Markov model would be capable of performing on Greek to any degree of accuracy. A particular complicating factor for Greek is the very large tagset: our full lexicon contains more than 1,400 tags, making it difficult for TreeTagger to build a decision tree from small datasets. Czech

(Hajič 1998) is comparable in the number of tags, but has lower rates of non-projectivity (compare Bamman and Crane 2006:72 on Latin).

Thus, for a first experiment, we built a comprehensive lexicon consisting of all surface forms occurring in Homer and Hesiod annotated with the parses occurring in the hand-disambiguated corpus--a subset of all grammatically possible parses--so that TreeTagger only had about 740 different possible tags to consider. Given this comprehensive lexicon and the Iliad and Odyssey as training data (200,000 words), we achieved 96.6% accuracy for Hesiod and the Homeric Hymns (see http://grade-devel.uchicago.edu/tagging.html).

The experiment established that a trigram Markov model was in fact capable of modeling Greek grammar remarkably well. The good results can be attributed in part to the formulaic nature of epic poetry and the large size of the training data, but they established the excellent potential of TreeTagger for Greek. This high degree of accuracy compares well with state-of-the-art taggers for such disparate languages as Arabic, Korean, and Czech (Smith et al., 2005).

Unfortunately, the Homeric data form a corpus that is of little use for classical Greek. In order to start analyzing classical Greek, we therefore used a hand-tagged Greek New Testament as our training data (160,000 words). New Testament Greek postdates the classical period by some four hundred years, and, not surprisingly, our initial accuracy on a 2,000 word sample of Lysias (4th century BCE oratory) was only 84% for morphological tagging, and performance on lemmas was weak. Computational linguists are familiar with the statistic that turning to out-of-domain data results in a ten percent loss of accuracy, so this result was not entirely unexpected.

At this point one could have decided to hand-tag an appropriate classical corpus and discard the out-of-domain data. Instead, we decided to integrate the output of Morpheus, thereby drastically raising the number of recognized forms and possible parses. While we had found that Morpheus alone produced too many ambiguous results to be practical as a parser, as a lexical resource for TreeTagger it is exemplary. TreeTagger's accuracy on the Lysias sample rose to 88%, with much improved recognition of lemmas. Certain common Attic constructs, unfortunately, were missed wholesale, but the decision tree from the New Testament demonstrated a grasp of the fundamentals.

While we are also working on improving accuracy by further refining the tagging system, so far we have seen the most prospects for improvement in augmenting our New Testament data with samples from classical Greek: When trained on our Lysias sample alone, TreeTagger performed at 96.8% accuracy when tested on that same text, but only performed at 88% on a new sample. In other words, 2,000 words of in-domain data performed no better or worse than 150,000 words of Biblical Greek combined with the Morpheus lexicon. We next used a combined training set of the tagged New Testament and the hand-tagged Lysias sample. In this case, the TreeTagger was capable of augmenting the basic decision tree it had already extracted from the NT alone with Attic-specific constructions. Ironically, this system only performed at 96.2% when turned back on the training data, but achieved 91% accuracy on the new sample (http://grade-devel.uchicago.edu/Lys2.html for results on the second sample). This is a substantial improvement given the addition of only 2,000 words of text, or less than 2% of the total training corpus. In the longer term, we aim at hand-disambiguating 40,000 words, double that of Schmid (1995), but comparable to Smith et al. (2005).

We conclude that automated tagging of classical Greek to a high level of accuracy can be achieved with quite limited human effort toward hand-disambiguation of in-domain data, thanks to the possibility of combining existing morphological data and machine learning, which together bootstrap a highly accurate morphological analysis. In our presentation we will report on our various approaches to improving these results still further, such as using a 6th order Markov model, enhancing the grammatical specificity of the tagset, and the results of several more iterations of our bootstrap procedure.

References

Bamman, David, and Gregory Crane (2006). The design and use of a Latin dependency treebank. In J. Hajič and J. Nivre (eds.), *Proceedings of the Fifth International Workshop on Treebanks and Linguistic Theories* (TLT) 2006, pp. 67-78. http://ufal.mff.cuni.cz/tlt2006/pdf/110.pdf

Crane, Gregory (1991). Generating and parsing classical Greek. *Literary and Linguistic Computing,* 6(4): 243-245, 1991.

Hajič, Jan (1998). Building a syntactically annotated corpus: The Prague Dependency Treebank. In Eva Hajičova (ed.), *Issues of Valency and Meaning,* pp. 106-132.

Schmid, Helmut (1994). Probabilistic part-of-speech tagging using decision trees. In *International Conference on New Methods in Language Processing,* pp. 44-49. http://www.ims.uni-stuttgart.de/ftp/pub/corpora/tree-tagger1.pdf

Schmid, Helmut (1995). Improvements in part-of-speech tagging with an application to German. In *Proceedings of the ACL SIGDAT-Workshop*. http://www.ims.uni-stuttgart.de/ftp/pub/corpora/tree-tagger2.pdf

Smith, Noah A., David A. Smith, and Roy W. Tromble (2005). Context-based morphological disambiguation with random fields. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pp. 475-482. http://www.cs.jhu.edu/~dasmith/sst_emnlp_2005.pdf

Mining Classical Greek Gender

Helma Dik
helmadik@mac.com
University of Chicago, USA

Richard Whaling
rwhaling@uchicago.edu
University of Chicago, USA

This paper examines gendered language in classical Greek drama. Recent years have seen the emergence of data mining in the Humanities, and questions of gender have been asked from the start. Work by Argamon and others has studied gender in English, both for the gender of authors of texts (Koppel et al. 2002, Argamon et al. 2003) and for that of characters in texts (Hota et al. 2006, 2007 on Shakespeare); Argamon et al. (in prep. a) study gender as a variable in Alexander Street Press's Black Drama corpus (gender of authors and characters) and (in prep. b) in French literature (gender of authors).

Surprisingly, some of the main findings of these studies show significant overlap: Female authors *and characters* use more personal pronouns and negations than males; male authors *and characters* use more determiners and quantifiers. Not only, then, do dramatic characters in Shakespeare and modern drama, like male and female authors, prove susceptible to automated analysis, but the feature sets that separate male from female characters show remarkable continuity with those that separate male from female authors.

For a scholar of Greek, this overlap between actual people and dramatic characters holds great promise. Since there are barely any extant Greek texts written by women, these results for English give us some hope that the corpus of Greek drama may serve as evidence for women's language in classical Greek.

After all, if the results for English-language literature showed significant differences between male and female characters, but no parallels with the differences found between male and female authors, we would be left with a study of gender characterization by dramatists with no known relation to the language of actual women. This is certainly of interest from a literary and stylistic point of view, but from a linguistic point of view, the English data hold out the promise that what we learn from Greek tragedy will tell us about gendered use of Greek language more generally, which is arguably a question of larger import, and one we have practically no other means of learning about.

There is some contemporary evidence to suggest that Greek males considered the portrayal of women on stage to be true to life. Plato's Socrates advises in the *Republic* against actors portraying women. Imitation must lead to some aspects of these inferior beings rubbing off on the (exclusively male) actors (*Republic* 3.394-395). Aristophanes, the comic playwright, has Euripides boast (*Frogs* 949f.) that he made tragedy democratic, allowing a voice to women and slaves alongside men.

Gender, of course, also continues to fascinate modern readers of these plays. For instance, Griffith (1999: 51) writes, on Antigone: "Gender lies at the root of the problems of *Antigone*. (...) Sophocles has created one of the most impressive female figures ever to walk the stage." Yet there are no full-scale studies of the linguistic characteristics of female speech on the Greek stage (*pace* McClure 1999).

In this paper, we report our results on data mining for gender in Greek drama. We started with the speakers in four plays of Sophocles (*Ajax, Antigone, Electra,* and *Trachiniae*), for a total of thirty characters, in order to test, first of all, whether a small, non-lemmatized Greek drama corpus would yield any results at all. We amalgamated the text of all characters by hand into individual files per speaker and analyzed the resulting corpus with PhiloMine, the data mining extension to PhiloLogic (http://philologic.uchicago.edu/philomine).

In this initial experiment, words were not lemmatized, and only occurrences of individual words, not bigrams or trigrams, were used. In spite of the modest size, results have been positive. The small corpus typically resulted in results of "100% correct" classification on different tasks, which is to be expected as a result of overfitting to the small amount of data. More significantly, results on cross-validation were in the 80% range, whereas results on random falsification stayed near 50%. We were aware of other work on small corpora (Yu 2007 on Dickinson), but were heartened by these positive results with PhiloMine, which had so far been used on much larger collections.

In our presentation, we will examine two questions in more depth, and on the basis of a larger corpus.

First, there is the overlap found between work on English and French. Argamon et al. (2007) laid down the gauntlet:

"The strong agreement between the analyses is all the more remarkable for the very different texts involved in these two studies. Argamon et al. (2003) analyzed 604 documents from the BNC spanning an array of fiction and non-fiction categories from a variety of types of works, all in Modern British English (post-1960), whereas the current study looks at longer, predominantly fictional French works from the 12th - 20th centuries. This cross-linguistic similarity could be supported with further research in additional languages."

So do we find the same tendencies in Greek, and if so, are we dealing with human, or 'Western' cultural, universals? Our initial results were mixed. When we ran a multinomial Bayes (MNB) analysis on a balanced sample, we did indeed see some negations show up as markers for female characters (3 negations in a top 50 of 'female' features; none in the male top

50), but pronouns and determiners show up in feature sets for both the female and male corpus. An emphatic form of the pronoun 'you' turned up as the most strongly male feature in this same analysis, and two more personal pronouns showed up in the male top ten, as against only one in the female top ten. Lexical items, on the other hand, were more intuitively distributed. Words such as 'army', 'man', 'corpse' and 'weapons' show up prominently on the male list; two past tense forms of 'die' show up in the female top ten. A larger corpus will allow us to report more fully on the distribution of function words and content words, and on how selections for frequency influence classifier results.

Secondly, after expanding our corpus, regardless of whether we find similar general results for Greek as for English and French, we will also be able to report on variation among the three tragedians, and give more fine-grained analysis. For instance, in our initial sample, we categorized gods and choruses as male or female along with the other characters (there are usually indications in the text as to the gender of the chorus in a given play, say 'sailors', 'male elders', 'women of Trachis'). Given the formal requirements of the genre, we expect that it will be trivial to classify characters as 'chorus' vs. 'non-chorus', but it will be interesting to see whether gender distinctions hold up within the group of choruses, and to what extent divinities conform to gender roles. The goddess Athena was the character most often mis-classified in our initial sample; perhaps this phenomenon will be more widespread in the full corpus. Such a finding would suggest (if not for the first time, of course) that authority and gender intersect in important ways, even as early as the ancient Greeks' conceptions of their gods.

In conclusion, we hope to demonstrate that data mining Greek drama brings new insights, despite the small size of the corpus and the intense scrutiny that it has already seen over the centuries. A quantitative study of this sort has value in its own right, but can also be a springboard for close readings of individual passages and form the foundation for a fuller linguistic and literary analysis.

References

Argamon, S., M. Koppel, J. Fine, A. Shimoni 2003. "Gender, Genre, and Writing Style in Formal Written Texts", *Text* 23(3).

Argamon, S., C. Cooney, R. Horton, M. Olsen, S. Stein (in prep. a). "Gender, Race and Nationality in Black Drama."

Argamon, S., J.-B. Goulain, R. Horton, M. Olsen (in prep. b). "Vive la Différence! Text Mining Gender Difference in French Literature."

Griffith, M. (ed.), 1999. *Sophocles: Antigone.*

Hota, S., S. Argamon, M. Koppel, and I. Zigdon, 2006. "Performing Gender: Automatic Stylistic Analysis of Shakespeare's Characters." Digital Humanities Abstracts 2006.

Hota, S., S. Argamon, R. Chung 2007. "Understanding the Linguistic Construction of Gender in Shakespeare via Text Mining." Digital Humanities Abstract 2007.

Koppel, M., S. Argamon, A. Shimoni 2002. "Automatically Categorizing Written Texts by Author Gender", *Literary and Linguistic Computing* 17:4 (2002): 401-12.

McClure, L., 1999. *Spoken Like a Woman: Speech and Gender in Athenian Drama.*

Yu, B. 2007. *An Evaluation of Text-Classification Methods for Literary Study.* Diss. UIUC.

Information visualization and text mining: application to a corpus on posthumanism

Ollivier Dyens
odyens@alcor.concordia.ca
Université Concordia, Canada

Dominic Forest
dominic.forest@umontreal.ca
Université de Montréal, Canada

Patric Mondou
patric.mondou@uqam.ca
Université du Quebec à Montréal, Canada

Valérie Cools
Université Concordia, Canada

David Johnston
Université Concordia, Canada

The world we live in generates so much data that the very structure of our societies has been transformed as a result. The way we produce, manage and treat information is changing; and the way we present this information and make it available raises numerous questions. Is it possible to give information an intuitive form, one that will enable understanding and memory? Can we give data a form better suited to human cognition? How can information display enhance the utility, interest and necessary features of data, while aiding memories to take hold of it? This article will explore the relevance of extracting and visualizing data within a corpus of text documents revolving around the theme of posthumanism. When a large quantity of data must be represented, visualizing the information implies a process of synthesis, which can be assisted by techniques of automated text document analysis. Analysis of the text documents is followed by a process of visualization and abstraction, which allows a global visual metaphoric representation of information. In this article, we will discuss techniques of data mining and the process involved in creating a prototype information-cartography software; and suggest how information-cartography allows a more intuitive exploration of the main themes of a textual corpus and contributes to information visualization enhancement.

Introduction

Human civilization is now confronted with the problem of information overload. How can we absorb an ever-increasing quantity of information? How can we both comprehend it and filter it according to our cognitive tools (which have adapted to understand general forms and structures but struggle with sequential and cumulative data)? How can we mine the riches that are buried in information? How can we extract the useful, necessary, and essential pieces of information from huge collections of data?

Most importantly, how can we guarantee the literal survival of memory? For *why* should we remember when an uncountable number of machines remember for us? And *how* can we remember when machines won't allow us to forget? Human memory has endured through time precisely because it is unaware of the greater part of the signals the brain receives (our senses gather over 11 million pieces of information per second, of which the brain is only aware of a maximum of 40 (Philipps, 2006, p.32)). Memory exists because it can forget.

Is an electronic archive a memory? We must rethink the way that we produce and handle information. Can we give it a more intuitive form that would better lend itself to retention and understanding? Can we better adapt it to human cognition? How can we extrapolate narratives from databases? How can we insert our memories into this machine memory? The answer is simple: by visualizing it.

For the past few years, the technical aspect of information representation and visualization has been the subject of active research which is gaining more and more attention (Card *et al.*, 1999; Chen, 1999; Don *et al.*, 2007; Geroimenko and Chen, 2005; Perer and Shneiderman, 2006; Spence, 2007). Despite the richness of the work that has been done, there is still a glaring lack of projects related to textual analyses, specifically of literary or theoretical texts, which have successfully integrated advances in the field of information visualization.

Objectives

Our project is part of this visual analytical effort. How can we visually represent posthumanism? How can we produce an image of its questions and challenges? How can we transform the enormous quantity of associated information the concept carries into something intuitive? We believe that the creation of a thematic map is part of the solution. Why a map? Because the information-driving suffocation we experience is also our inability to create cognitive maps (as Fredric Jameson suggested). And so we challenged ourselves to create a map of posthumanism, one that could be read and understood intuitively.

To do this, we have chosen to use Google Earth (GE) as the basic interface for the project. GE's interface is naturally intuitive. It corresponds to our collective imagination and it also allows users to choose the density of information they desire. GE allows the user to "dive" into the planet's geographical, political, and social layers, or to stay on higher, more general levels. GE "tells" the story of our way of seeing the world. GE shows us the world the way science describes it and political maps draw it.

This leads us to confront three primary challenges: the first was to find an area to represent posthumanism. The second was to give this area a visual form. The third was to integrate, in an intuitive way, the significant number of data and texts on the subject.

Methodology

In order to successfully create this thematic map, we first compiled a significant number of texts about posthumanism. The methodology used to treat the corpus was inspired by previous work in the field of text mining (Forest, 2006; Ibekwe-Sanjuan, 2007; Weiss *et al*., 2005). The goal of the analysis is to facilitate the extraction and organization of thematic groups. Non-supervised clustering technique is used in order to produce a synthetic view of the thematic area of the subject being treated. The data is processed in four main steps:

1. Segmentation of the documents, lexical extraction and filtering

2. Text transformation using vector space model

3. Segment classification

4. Information visualization

Having established and sorted a sizeable information-population, a new continent was needed. The continental outline of Antarctica was adopted. Antarctica was chosen because its contours are relatively unknown and generally unrecognizable; it tends to be thought of more as a white sliver at the bottom of the world map than a real place. Politically neutral Antarctica, whose shape is curiously similar to that of the human brain, is a large area surrounded by oceans. These qualities made it symbolically ideal for utilization in a map of posthumanism, a *New World* of the 21st century.

Antarctica also allowed us to avoid information overload typical of a populated area: it carries minimal associative memories and historic bias; few people have lived there. To differentiate our new continent, the contour of Antarctica was moved into the mid Pacific. The continent was then reskinned visually with terrain suggestive of neurology, cybernetics, and symbiosis. The end result was a new land mass free of memories ready for an abstract population.

Visualizing the results

After identifying the thematic structure of the corpus, themes are converted into regions. Themes are assigned portions of the continent proportional to their occurrence. Inside each major region are 2 added levels of analysis: sub-regions and sub-sub-regions, each represented by several keywords. Keywords correspond to clusters discovered during the automated text analysis. So it is possible to burrow physically downward into the subject with greater and greater accuracy; or to rest at an attitude above the subject and consider the overall 'terrain'. Each area is associated with a category of documents. From each area it is possible to consult the documents associated with each region.

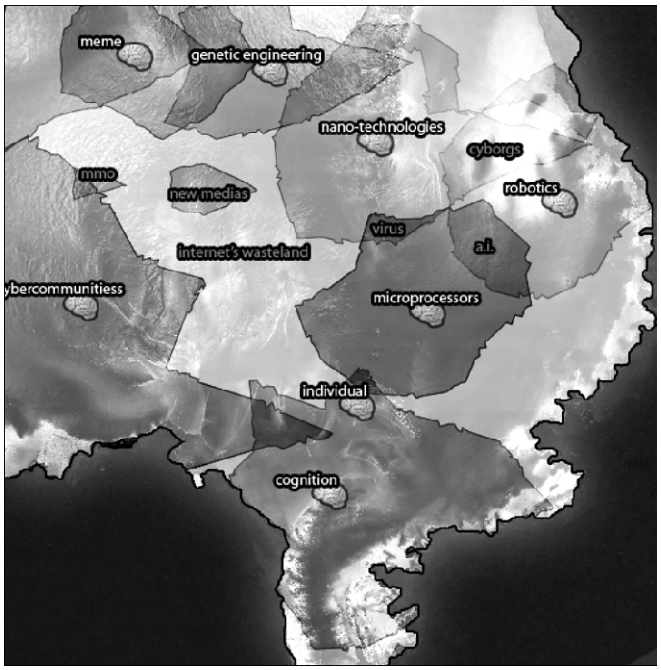


Figure 1. A general level of the visualized thematic map

The system offers users a certain number of areas, based on the algorithms used to process the data. These areas represent themes. Clicking on the brain icons allows the user to read an excerpt from one of the texts that is part of the cluster.

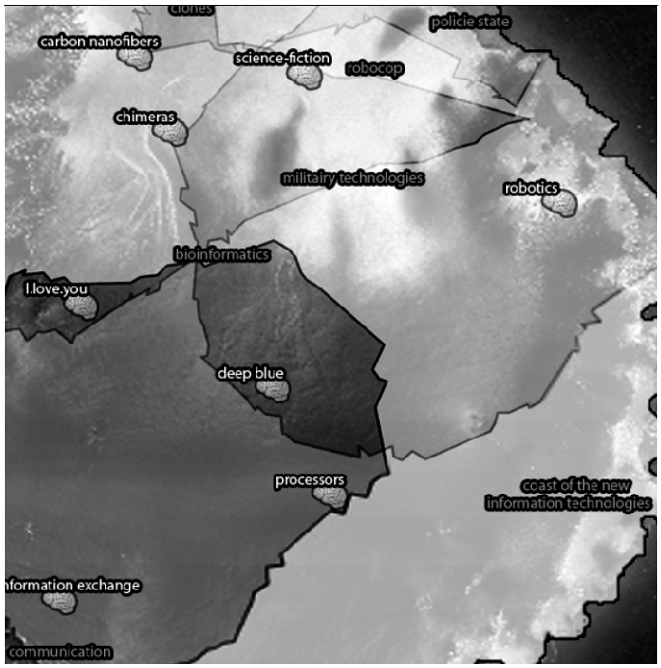


Figure2. A specific level of the thematic map

When the user zooms in on a region, the application shows the next level in the hierarchy of data visualization. Within one theme (as shown below) several sub-themes appear (in red). A greater number of excerpts is available for the same area.

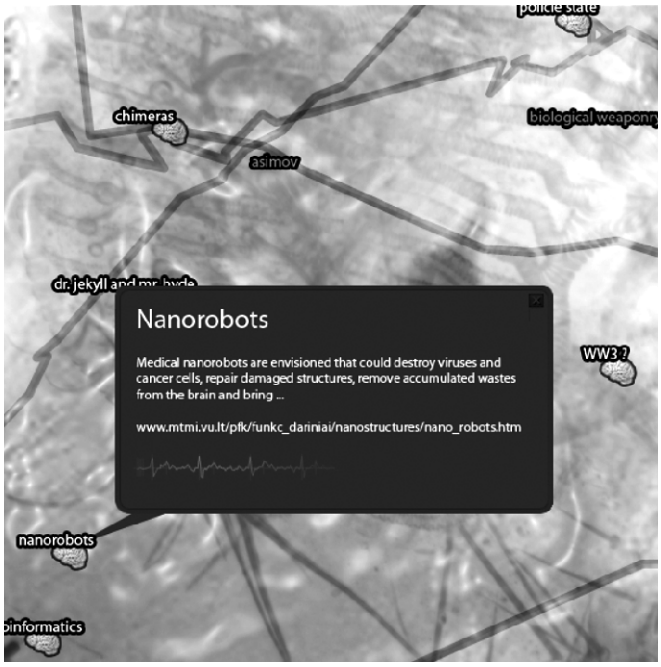


Figure 3. Looking at a document from the thematic map

The icons indicating the availability of texts may be clicked on at any time, allowing the user to read the excerpt in a small pop-up window, which includes a link to the whole article. This pop-up window can serve to show pertinent images or other hyperlinks.

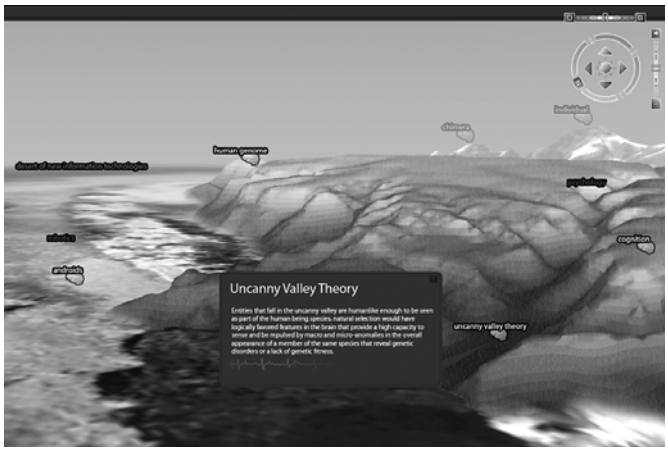


Figure 4. 3-dimensional rendering of the thematic map

At any time, the user can rotate the point of view or change its vertical orientation. This puts the camera closer to the ground and allows the user to see a three dimensional landscape.

Conclusion

We see the visualization of data, textual or otherwise, as part of a fundamental challenge: how to transform information into knowledge and understanding. It is apparent that the significant amount of data produced by research in both science and the humanities is often much too great for any one individual. This overload of information sometimes leads to social disinvestment as the data eventually cancel each other out. We think that giving these data an intuitive form will make their meaning more understandable and provide for their penetration into the collective consciousness. Posthumanism seems particularly well adapted to pioneer this system because it questions the very definition of what it is to be human.

References

Alpaydin, E. (2004). *Introduction to machine learning*. Cambridge (Mass.): MIT Press.

Aubert, N. (2004). "Que sommes-nous devenus?" *Sciences humaines, dossier l'individu hypermoderne,* no 154, novembre 2004, pp. 36-41.

Card, S. K., Mackinlay, J. and Shneiderman, B. (1999). *Readings in information visualization: using vision to think*. Morgan Kaufmann.

Chen, C. (1999). *Information visualisation and virtual environments*. New York: Springer-Verlag.

Clement, T., Auvil, L., Plaisant, C., Pape, G. et Goren, V. (2007). "Something that is interesting is interesting them: using text mining and visualizations to aid interpreting repetition in Gertrude Steins The Making of Americans." *Digital Humanities 2007*. University of Illinois, Urbana-Champaign, June 2007.

Don, A., Zheleva, E., Gregory, M., Tarkan, S., Auvil, L., Clement, T., Shneiderman, B. and Plaisant, C. (2007). *Discovering interesting usage patterns in text collections: integrating text mining with visualization*. Rapport technique HCIL-2007-08, Human-Computer Interaction Lab, University of Maryland.

Garreau, J. (2005). *Radical Evolution*. New York: Doubleday.

Geroimenko, V. and Chen, C. (2005). *Visualizing the semantic web: XML-based Internet and information visualization*. New York: Springer-Verlag.

Forest, D. (2006). *Application de techniques de forage de textes de nature prédictive et exploratoire à des fins de gestion et d'analyse thématique de documents textuels non structurés*. Thèse de doctorat, Montréal, Université du Québec à Montréal.

Frenay, R. (2006). *Pulse, The coming of age of systems, machines inspired by living things*. New York: Farrar, Strauss and Giroux.

Harraway, D. (1991). *Simians, cyborgs, and women. The reinvention of nature*. New York: Routledge.

Ibekwe-SanHuan, F. (2007). *Fouille de textes. Méthodes, outils et applications.* Paris: Hermès.

Jameson, F. (1991). *Postmodernism, or, the cultural logic of late capitalism*. Verso.

Joy, B. (2000). "Why the future doesn't need us." *Wired*. No 8.04, April 2000.

Kelly, K. (1994). *Out of control. The rise of neo-biological civilization*. Readings (Mass.): Addison-Wesley Publishing Company.

Kurzweil, R. (2005). *The singularity is near: when humans transcend biology*. Viking.

Levy, S. (1992). *Artificial life*. New York: Vintage Books, Random House.

Manning, C. D. and H. Schütze. (1999). *Foundations of statistical natural language processing.* Cambridge (Mass.): MIT Press.

Perer, A. and Shneiderman, B. (2006). "Balancing systematic and flexible exploration of social networks." *IEEE Transactions on Visualization and Computer Graphics*. Vol. 12, no 5, pp. 693-700.

Philips, H. (2006). "Everyday fairytales." *New Scientist*, 7-13 October 2006.

Plaisant, C., Rose, J., Auvil, L., Yu, B., Kirschenbaum, M., Nell Smith, M., Clement, T. and Lord, G. (2006). "Exploring erotics in Emily Dickinson's correspondence with text mining and visual interfaces." *Joint conference on digital libraries 2006*. Chapel Hill, NC, June 2006.

Ruecker, S. 2006. "Experimental interfaces involving visual grouping during browsing." *Partnership: the Canadian journal of library and information practice and research*. Vol. 1, no 1, pp. 1-14.

Salton, G. (1989). *Automatic Text Processing.* Reading (Mass.): Addison-Wesley.

Spence, R. (2007). *Information visualization: design for interaction*. Prentice Hall.

Tattersall, I. (2006). "How we came to be human." *Scientific American*, special edition: becoming human. Vol 16, no 2, pp. 66-73.

Unsworth, J. (2005). New methods for humanities research. The 2005 Lyman award lecture, National Humanities Center, 11 November 2005.

Unsworth, J. (2004). "Forms of attention: digital humanities beyond representation." *The face of text: computer-assisted text analysis in the humanities. The third conference of the Canadian symposium on text analysis (CaSTA)*. McMaster University, 19-21 November 2004.

Weiss, S. M., Indurkhya, N., Zhang, T. and Damereau, F. J. (2005). *Text mining. Predictive methods for analyzing unstructured information*. Berlin; New York: Springer-Verlag.

How Rhythmical is Hexameter: A Statistical Approach to Ancient Epic Poetry

Maciej Eder
maciejeder@poczta.ijp-pan.krakow.pl
Polish Academy of Sciences, Poland

In this paper, I argue that the later a given specimen of hexamater is, the less rhythmical it tends to be. A brief discussion of the background of ancient Greek and Latin metrics and its connections to orality is followed by an account of spectral density analysis as my chosen method. I then go on to comment on the experimental data obtained by representing several samples of ancient poetry as coded sequences of binary values. In the last section, I suggest how spectral density analysis may help to account for other features of ancient meter.

The ancient epic poems, especially the most archaic Greek poetry attributed to Homer, are usually referred as an extraordinary fact in the history of European literature. For present-day readers educated in a culture of writing, it seems unbelievable that such a large body of poetry should have been composed in a culture based on oral transmission. In fact, despite of genuine singers' and audience's memory, epic poems did not emerge at once as fixed texts, but they were re-composed in each performance (Lord 2000, Foley 1993, Nagy 1996). The surviving ancient epic poetry displays some features that reflect archaic techniques of oral composition, formulaic structure being probably the most characteristic (Parry 1971: 37-117).

Since formulaic diction prefers some fixed rhythmical patterns (Parry 1971: 8-21), we can ask some questions about the role of both versification and rhythm in oral composition. Why was all of ancient epic poetry, both Greek and Latin, composed in one particular type of meter called hexameter? Does the choice of meter influence the rhythmicity of a text? Why does hexameter, in spite of its relatively restricted possibilities of shaping rhythm, differ so much from one writer to another some (cf. Duckworth 1969: 37-87)? And, last but not least, what possible reasons are there for those wide differences between particular authors?

It is commonly known that poetry is in general easier to memorize than prose, because rhythm itself tends to facilitate memorization. In a culture without writing, memorization is crucial, and much depends on the quality of oral transmission. In epic poems from an oral culture rhythm is thus likely to be particularly important for both singers and hearers, even though they need not consciously perceive poetic texts as rhythmical to benefit from rhythm as an aid to memory.

It may then be expected on theoretical grounds that non-oral poems, such as the Latin epic poetry or the Greek hexameter of the Alexandrian age, will be largely non-rhythmical, or at least display weaker rhythm effects than the archaic poems of Homer and Hesiod. Although formulaic diction and other techniques of oral composition are noticeable mostly in Homer's epics (Parry 1971, Lord 2000, Foley 1993, etc.), the later hexameters, both Greek and Latin, also display some features of oral diction (Parry 1971: 24-36). The metrical structure of hexameter might be quite similar: strongly rhythmical in the oldest (or rather, the most archaic) epic poems, and less conspicuous in poems composed in written form a few centuries after Homer. The aim of the present study is to test the hypothesis that the later a given specimen of hexameter is, the less rhythmical it tends to be.

Because of its nature versification easily lends itself to statistical analysis. A great deal of work has already been done in this field, including studies of Greek and Latin hexameter (Jones & Gray 1972, Duckworth 1969, Foley 1993, etc.). However, the main disadvantage of the methods applied in existing research is that they describe a given meter as if it were a set of independent elements, which is actually not true. In each versification system, the specific sequence of elements plays a far more important role in establishing a particular type of rhythm than the relations between those elements regardless their linear order (language "in the mass" vs. language "in the line"; cf. Pawlowski 1999).

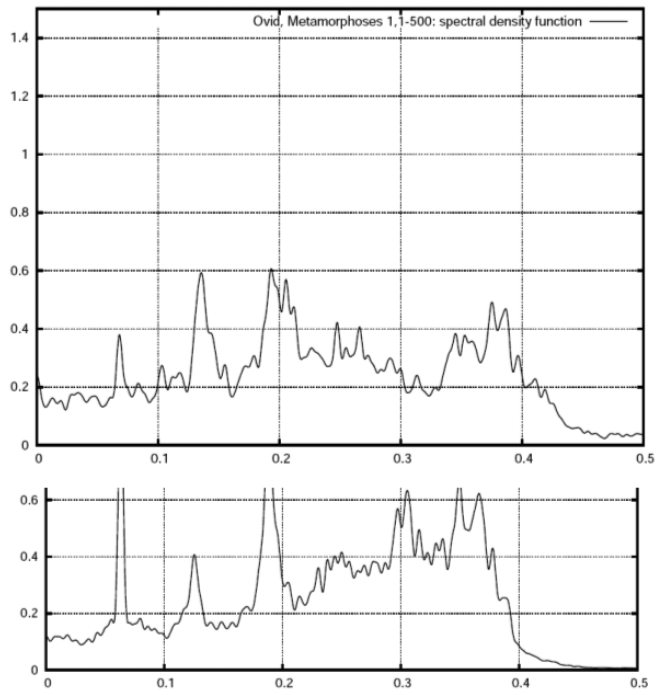
Fortunately, there are a few methods of statistical analysis (both numeric and probabilistic) that study verse by means of an ordered sequence of elements. These methods include, for example, time series modeling, Fourier analysis, the theory of Markov chains and Shannon's theory of information. In the present study, spectral density analysis was used (Gottman 1999, Priestley 1981, etc.). Spectral analysis seems to be a very suitable tool because it provides a cross-section of a given time series: it allows us to detect waves, regularities and cycles which are not otherwise manifest and open to inspection. In the case of a coded poetry sample, the spectrogram shows not only simple repetitions of metrical patterns, but also some subtle rhythmical relations, if any, between distant lines or stanzas. On a given spectrogram, a distinguishable peak indicates the existence of a rhythmical wave; numerous peaks suggest a quite complicated rhythm, while a pure noise (no peaks) on the spectrogram reflects a non-rhythmical data.

To verify the hypothesis of hexameter's decreasing rhythmicity, 7 samples of Greek and 3 samples of Latin epic poetry were chosen. The specific selection of sample material was as follows: 3 samples from Homeric hexameter (books 18 and 22 from the Iliad, book 3 from the Odyssey), 1 sample from Hesiod (Theogony), Apollonius (Argonautica, book 1), Aratos (Phainomena), Nonnos (Dionysiaca, book 1), Vergil (Aeneid, book 3), Horace (Ars poetica), and Ovid (Metamorphoses, book 1). In each sample, the first 500 lines were coded in such a way that each long syllable was assigned value 1, and each short syllable value 0. Though it is disputed whether ancient verse was purely quantitative or whether it also had some prosodic features (Pawlowski & Eder 2001), the quantity-based nature of Greek and Roman meter was never questioned. It is probable that rhythm was generated not only by quantity (especially in live performances), but it is certain that quantity itself played an essential role in ancient meter. Thus, in the coding procedure, all prosodic features were left out except the quantity of syllables (cf. Jones & Gray 1972, Duckworth 1969, Foley 1993, etc.). A binary-coded series was then obtained for each sample, e.g., book 22 of the Iliad begins as a series of values:

1110010010010011100100100100111001110010010011...

The coded samples were analyzed by means of the spectral density function. As might be expected, on each spectrogram there appeared a few peaks indicating the existence of several rhythmical waves in the data. However, while the peaks suggesting the existence of 2- and 3-syllable patterns it the text were very similar for all the spectrograms and quite obvious, the other peaks showed some large differences between the samples. Perhaps the most surprising was the peak echoing the wave with a 16-syllable period, which could be found in the samples of early Greek poems by Homer, Hesiod, Apollonius, and Aratos (cf. Fig. 1). The same peak was far less noticeable in the late Greek hexameter of Nonnos, and almost absent in the samples of Latin writers (cf. Fig. 2). Other differences between the spectrograms have corroborated the observation: the rhythmical effects of the late poems were, in general, weaker as compared with the rich rhythmical structure of the earliest, orally composed epic poems.

Although the main hypothesis has been verified, the results also showed some peculiarities. For example, the archaic poems by Homer and Hesiod did not differ significantly from the poems of the Alexandrian age (Apollonius, Aratos), which was rather unexpected. Again, the rhythm of the Latin hexameter turned out to have a different underlying structure than that of all the Greek samples. There are some possible explanations of those facts, such as that the weaker rhythm of the Latin samples may relate to inherent differences between Latin and Greek. More research, both in statistics and in philology, is needed, however, to make such explanations more nuanced and more persuasive.



Bibliography

Duckworth, George E. *Vergil and Classical Hexameter Poetry: A Study in Metrical Variety*. Ann Arbor: University of Michigan Press, 1969.

Foley, John Miles. *Traditional Oral Epic: "The Odyssey", "Beowulf" and the Serbo-Croatian Return Song*. Berkeley: University of California Press, 1993.

Gottman, John Mordechai, and Anup Kumar Roy. *Sequential Analysis: A Guide for Behavioral Researchers*. Cambridge: Cambridge University Press, 1990.

Jones, Frank Pierce, and Florence E. Gray. Hexameter Patterns, Statistical Inference, and the Homeric Question: An Analysis of the La Roche Data. *Transactions and Proceedings of the American Philological Association* 103 (1972): 187-209.

Lord, Albert B. Ed. Stephen Mitchell and Gregory Nagy. *The Singer of Tales*. Cambridge, MA: Harvard University Press, 2000.

Nagy, Gregory. *Poetry as Performance: Homer and Beyond*. Cambridge: Cambridge University Press, 1996.

Parry, Milman. Ed. Adam Parry. *The Making of Homeric Verse: The Collected Papers of Milman Parry*. Oxford: Clarendon Press, 1971.

Pawlowski, Adam. Language in the Line vs. Language in the Mass: On the Efficiency of Sequential Modeling in the Analysis of Rhythm. *Journal of Quantitative Linguistics* 6.1 (1999): 70-77.

Pawlowski, Adam, and Maciej Eder. Quantity or Stress? Sequential Analysis of Latin Prosody. *Journal of Quantitative Linguistics* 8.1 (2001): 81-97.

Priestley, M. B. *Spectral Analysis and Time Series*. London: Academic Press, 1981.

TEI and cultural heritage ontologies

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

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

Introduction

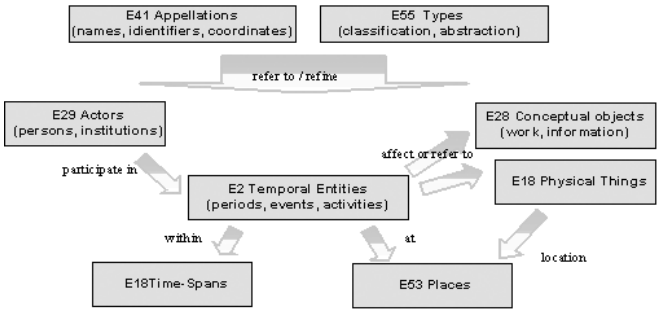
Since the mid 1990s there has been an increase in the interest for the design and use of conceptual models (ontologies) in humanities computing and library science, as well as in knowledge engineering in general. There is also a wish to use such models to enable information interchange. TEI has in its 20 years of history concentrated on the mark up of functional aspects of texts and their parts. That is, a person name is marked but linking to information about the real world person denoted by that name was not in the main scope. The scope of TEI has gradually broadened, however, to include more real world information external to the text in question. The Master project (Master 2001) is an early example of this change.

In TEI P5 a series of new elements for marking up real world information are introduced and several such elements from the P4 are adjusted. TEI P5 is meant to be a set of guidelines for the encoding of a large variety of texts in many cultural contexts. Thus the set of real world oriented elements in TEI P5 should not formally be bound to a single ontology. The ontological part of TEI P5 is, however, close connected to the authors implicit world view. Thus we believe it is important to study this part of TEI P5 with some well defined ontology as a yardstick. Our long experience with memory institution sector makes CIDOC CRM (Conceptual Reference Model) a natural choice. CIDOC CRM (Crofts 2005) has been proven useful as an intellectual aid in the formulation of the intended scope of the elements in a new mark up schemes and we believe the model can be useful to clarify the ontological part of TEI. This will also clarify what is needed in order to harmonize it with major standards like CIDOC CRM, FRBR, EAD and CDWA Lite.

CIDOC CRM

CIDOC CRM is a formal ontology intended to facilitate the integration, mediation and interchange of heterogeneous cultural heritage information. It was developed by interdisciplinary teams of experts, coming from fields such as computer science, archaeology, museum documentation, history of arts, natural history, library science, physics and philosophy, under the aegis of the International Committee for Documentation (CIDOC) of the International Council of Museums (ICOM). The harmonisation of CIDOC CRM and IFLA's FRBR (FRBR 1998) is in the process of being completed. The EAD has already been mapped to CIDOC CRM (Theodoridou 2001).

CIDOC CRM is defined in an object oriented formalism which allow for a compact definition with abstraction and generalisation. The model is event centric, that is, actors, places and objects are connected via events. CIDOC CRM is a core ontology in the sense that the model does not have classes for all particulars like for example the Art and Architecture Thesaurus with thousands of concepts. CIDOC CRM has little more than 80 classes and 130 properties. The most central classes and properties for data interchange are shown below.



Example: The issue of a medieval charter can be modelled as an activity connecting the issuer, witnesses, scribe and place and time of issue. The content of the charter is modelled as a conceptual object and the parchment as a physical thing. In cases where it is necessary for a scholarly analysis and when sufficient information has been preserved, an issuing of a charter can be broken down into a series of smaller events, e.g., the actual declaration in a court, the writing of the parchment and the attachment of the seals. This conceptual analysis can be can be used an intellectual aid in the formulation a data model and implementation.

In 2005 the CRM was reformulated as a simple XML DTD, called CRM-Core, to enable CRM compliant mark up of multimedia metadata (Sinclair 2006). A CRM-Core XML package may contain information about a single instance of any class in the CRM and how it may be connected to other objects via events and properties. The German Council of Museum has based its standard for XML based museum data interchange, MUSEUMDAT, on a combination of the Getty standard CDWA Lite and CRM Core. The CDWA Lite revision group currently considers these changes to CDWA Lite in order to make it compatible with CRM.

TEI P5 ontology elements in the light of CIDOC CRM

In TEI P5 the new ontologically oriented elements is introduced in the module NamesPlaces described in chapter *13 Names, Dates, People, and Places*. There are additional elements described in chapter *10 Manuscript Description*, in the TEI header and in connection with bibliographic descriptions as well. In this paper we concentrate on the elements in chapter 13.

The central elements in this module are: *person, personGrp, org, place* and *event*. *Person, personGrp* and *org* are "elements which provide information about people and their relationships". CIDOC CRM has the corresponding classes with a common superclass *E29 Actor*.

The element *event* is defined as "contains data relating to any kind of significant event associated with a person, place, or organization" and is similar to the CIDOC CRM class E5 Event and its subclasses. In the discussion of the marriage example in chapter 13, *event* element is presented as a "freestanding" element. In the formal definition it is limited to *person* and *org*. To make this coherent, the formal part will have to be extended or the example have to be changed.

Still event is problematic. The marriage example demonstrates that it is impossible to express the role a person has in an event. Without knowing the English marriage formalism one doesn't know if the "best man" participated. The very generic element *persRel* introduced in P5 does not solve this problem. A possible solution to this problem would be to introduce an *EventStateLike* model class with elements for roles and participants.

The model classes *orgStateLike, personStateLike, personTraitLik, placeStateLike, placeTraitLike* group elements used to mark up characteristics of persons, organisations and places. The elements in *...TraitLike* model classes contain information about permanent characteristics and the elements in *...StateLike* information about more temporal characteristics. The model classes contain the generic *Trait* and *State* elements in addition to specialised elements. The intention is to link all characteristics relating to a person, organisation or place. It is not possible to make a single mapping from these classes into CIDOC-CRM. It will depend partly on which type of trait or strait is used, and partly on the way in which it is used. Many characteristics will correspond to persistent items like *E55 Types, E3 String and E41 Appellation*, and are connected to actors and places through the properties *P1 is identified, P2 has type and P2 has note*. Other elements like *floruit*, which is used to describe a person's active period, are temporal states corresponding to the CIDOC-CRM temporal entity *E3 Condition State*. From an ontological point of view the two elements state and trait can be considered as generic mechanism for typed linking between the major classes.

All the elements in *...TraitLike* and *...StateLike* model classes can be supplied with the attributes *notAfter* and *notBefore* defining the temporal extension of their validity. This is a very powerful mechanism for expressing synoptically information based on hidden extensive scholarly investigation about real world events. As long as the justification for the values in these elements is not present, however, it is hard to map this information into an event oriented conceptual model like the CRM. Thus, it is important to include descriptions of methods for such justification in the guidelines, including examples.

TEI ontology – conclusion

The new elements in TEI P5 bring TEI a great step in the direction of an event oriented model. Our use of CRM Core as a yardstick has shown that small extensions to and adjustments of the P5 elements will enable the expression of CRM Core packages by TEI elements. This is a major change to our previous suggestions (Ore 2006) in which the ontological module was outside TEI.

To continue this research, an extended TEI tagset should be developed with element for abstracts corresponding to the ones in FRBR and CRM. This will not change the ontological structure of TEI significantly. But these adjustments will make the ontological information in a TEI document compliant with the other cultural heritage models like for example EAD, FRBR/FRBRoo, CIDOC CRM and CDWA-Lite. There is an ongoing harmonisation process between all these initiatives in which it is important that TEI is a part.

Bibliography

Crofts, N., Doerr, M., Gill, T., Stead, S. and Stiff M. (eds.) (2005): *Definition of the CIDOC Conceptual Reference Model.* (June 2005). URL: http://cidoc.ics.forth.gr/docs/cidoc_crm_version_4.2.doc (checked 2007-11-15)

CDWA Lite www.getty.edu/research/conducting_research/standards/cdwa/cdwalite.html (checked 2007-11-25)

FRBR (1998). Functional Requirement for Bibliographic Records. Final Report. International Federation of Library Associations. URL: http://www.ifla.org/VII/s13/frbr/frbr.pdf (checked 2007-11-24)

MASTER (2001). "Manuscript Access through Standards for Electronic Records (MASTER)." Cover Pages: Technology Reports. URL: http://xml.coverpages.org/master.html (checked 2007-11-25)

MUSEUMDAT (www.museumdat.org/, checked 2007-11-25)

Ore, Christian-Emil and Øyvind Eide (2006). "TEI, CIDOC-CRM and a Possible Interface between the Two." P. 62-65 in Digital Humanities 2006. Conference Abstracts. Paris, 2006.

Sinclair, Patrick & al. (2006). "The use of CRM Core in Multimedia Annotation." Proceedings of First International Workshop on Semantic Web Annotations for Multimedia (SWAMM 2006). URL: http://cidoc.ics.forth.gr/docs/paper16.pdf (checked 2007-11-25)

TEI P5 (2007). Guidelines for Electronic Text Encoding and Interchange. URL: http://www.tei-c.org/Guidelines/P5/ (checked 2007-11-15)

Theodoridou, Maria and Martin Doerr (2001). *Mapping of the Encoded Archival Description DTD Element Set to the CIDOC CRM*, Technical Report FORTH-ICS/TR-289. URL: http://cidoc.ics.forth.gr/docs/ead.pdf (checked 2007-11-25)

Homebodies and Gad-Abouts: A Chronological Stylistic Study of 19th Century French and English Novelists

Joel Goldfield
joel@cs.fairfield.edu
Fairfield University, USA

David L. Hoover
david.hoover@nyu.edu
New York University, USA

The assumption that authorial style changes over time is a reasonable one that is widely accepted in authorship attribution. Some important studies concentrate on individual authors, using a wide array of methods, and yielding varied results, however, so that is unwise to generalize (see the overview in Stamou 2007). Henry James's style, for example, changes in a remarkably consistent and extreme way (Hoover 2007), and Austen's stylistic development also seems consistent (Burrows 1987), but Beckett's texts are individually innovative without showing consistent trends (Stamou 2007: 3). For some authors, different investigators have obtained inconsistent results. We know of no general study that investigates how authorial style changes over a long career. We thus explore how working and otherwise living abroad for periods exceeding two years may affect an author's vocabulary and style.

Our presentation begins a broadly based study of the growth and decay of authorial vocabulary over time. Although we limit our study to nineteenth-century prose, to reduce the number of possible variables to a manageable level, we compare French and English authors, allowing us to test whether any discovered regularities are cross-linguistic. Because authors' vocabularies seem related to their life experiences, we compare authors who spent most of their lives in a single geographical area with others who traveled and lived extensively abroad. For the purposes of this study, we define extensive living and working abroad as at least two consecutive years, somewhat longer than the contemporary American student's or faculty member's stay of a semester or a year. This differentiation allows us to investigate in what significant ways, if any, extensive foreign travel affects an author's vocabulary.

Our investigation of these questions requires a careful selection of authors and texts. Although research is still ongoing, we have selected the following eight authors–two in each of the four possible categories–for preliminary testing and further biographical study:

Domestic authors:

English

• George Meredith (1828-1909). German schooling at age fifteen, but apparently no significant travel later; fifteen novels published between 1855 and 1895 available in digital form.

• Wilkie Collins (1824-1889). No foreign travel mentioned in brief biographies; 20 novels available over 38 years.

French

• Honoré de Balzac (1799-1850). He traveled little outside of France until 1843. Subsequent excursions abroad in the last seven years of his life, mainly for romance, were relatively infrequent and never exceeded five consecutive months. Over two dozen novels in digitized format are available.

• Jules Barbey d'Aurevilly (1808-1889). Raised in France and schooled in the law, he was devoted to his native Normandy and seldom ventured abroad. Initially cultivating the image of the dandy, his novels and novellas create a curious paradox in his later writing between sexually suggestive themes and nostalgia for earlier aesthetics and a defense of Catholicism. His literary productivity can be divided into the three periods of 1831-1844 (first published novel and novella), 1845-1873 (return to Catholicism; work as both literary critic and novelist), and 1874-1889. At least fourteen novels or novellas are available in digitized format.

Traveling authors:

English

• Dickens (1812-1870). Seven novels available 1836-43, foreign travel for 3 yrs 1844-47, and more travel in 1853-55; four novels after 1855 available.

• Trollope (1815-1883). Travel to Brussels in 1834, but briefly; six novels available before 1859, Postal missions to Egypt, Scotland, and the West Indies, 1858-59; 1871 trip to Australia, New Zealand, and US; travel to Ceylon and Australia, 1875, South Africa 1877, Iceland 1878; five novels available after 1878.

French

• Arthur de Gobineau (1816-1882). Raised partly in France, partly in Germany and Switzerland, he learned German and began the study of Persian in school while his mother, accused of fraud and estranged from her military husband, kept the family a step ahead of the French police. Three periods encompassing his publication of fiction and often related to career travel can be divided as follows: 1843-1852 (at least 4 novellas and 1 novel); 1853-1863 (1 novella); 1864-1879 (10 novellas, 2 novels). Living mainly in France from 1831-1849, he was a protégé of Alexis de Tocqueville, who brought him into the French diplomatic service in 1849. Gobineau was then stationed in Switzerland and Germany (1849-1854), Persia (1855-1858, 1861-1863), Newfoundland (1859), Greece (1864-1868), Brazil (1868-1870) and Sweden (1872-1873). Following travel through Europe in 1875-77, he left the diplomatic service. His first short fiction was published and serialized in 1846. At least a dozen novellas written throughout his career (mostly written and published as collections) and two novels (1852 and 1874) are available in digitized format.

• Victor Hugo (1802-85). Raised in France except for a six-month period in a religious boarding school in Madrid (1811-12), Hugo began writing his first novel, Bug-Jargal (1826) in 1820. This initial literary period includes 1820-1851. Aside from a few short trips lasting less than three months, Hugo lived and wrote mainly in his homeland until his exile on the Island of Guernsey during the reign of Louis-Napoléon from 1851-1870. The third period encompasses his triumphant return to Paris in 1871 until his death, celebrated by a state funeral, in 1885.

Research on the French authors is being facilitated by use of PhiloLogic and the ARTFL database, complemented by local digitized works and tools that are also used to investigate the English authors.

Preliminary testing must be done on all the authors to discover overall trends or principles of vocabulary development before investigating any possible effects of foreign travel. The importance of this can be seen in figures 1 and 2 below, two cluster analyses of English traveling authors, based on the 800 mfw (most frequent words). Dickens's novels form two distinct groups, 1836-43 and 1848-70, a division coinciding with his 1844-47 travels. For Trollope, the match is not quite so neat, but it is suggestive. Ignoring *Nina Balatka*, a short romance set in Prague, and *La Vendee*, Trollope's only historical novel, set in France in the 1790's, only two remaining novels in the early group were published after his travel to Egypt, Scotland, and the West Indies in 1858-59.

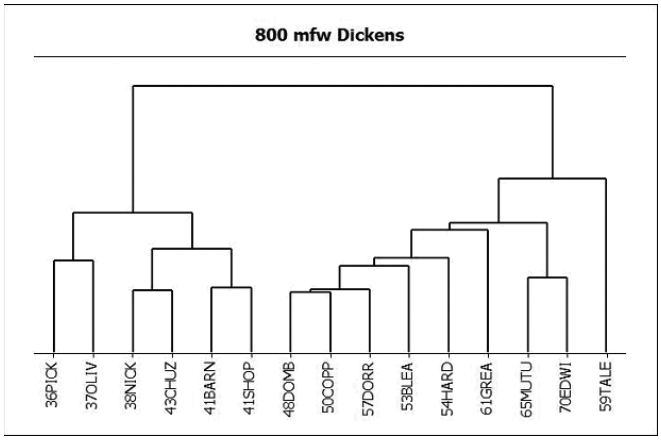


Fig. 1 – Fifteen Dickens Novels, 1836-1870

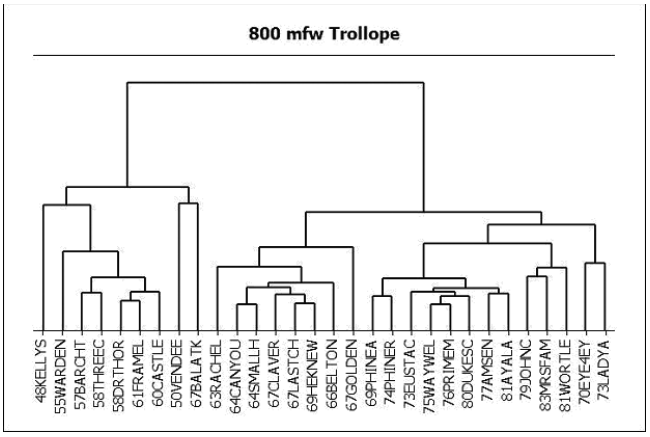


Fig. 2 – Thirty Trollope Novels, 1848-1883

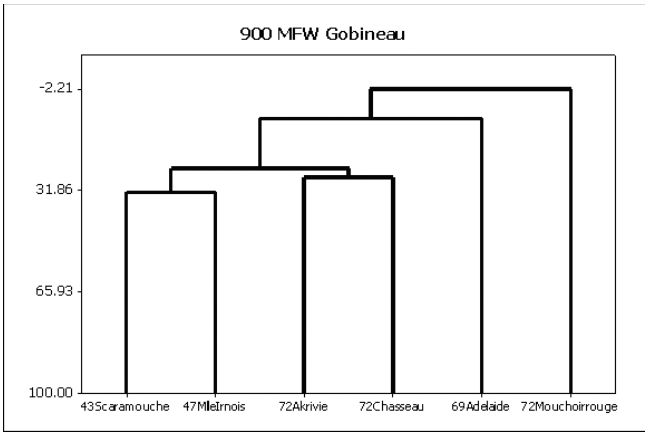


Fig. 5 – Six Texts by Gobineau

Unfortunately, comparing cluster analyses of the English domestic authors casts doubt on any simple correlation between travel and major stylistic change. Meredith, like Dickens and Trollope, shows a sharp break between his two early novels 1855-57 and the rest of his novels (see Fig. 3). And, ignoring *Antonina* (an historical novel about the fall of Rome), Collins's novels also form early and late groups (see Fig. 4).

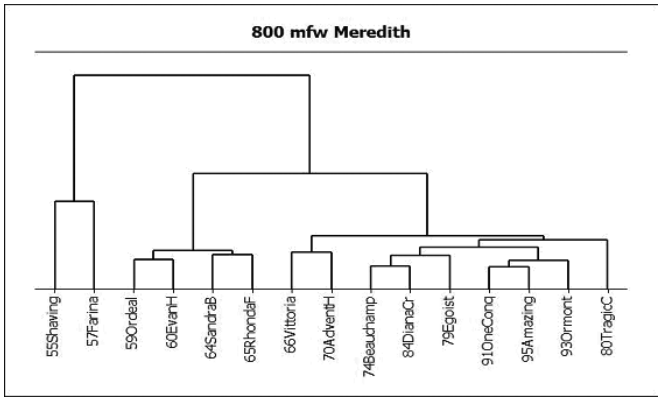


Fig. 3 – Fifteen Meredith Novels, 1855-895

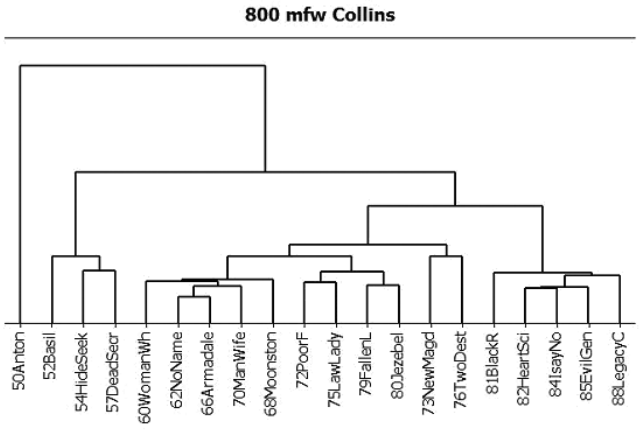


Fig. 4 – Twenty Collins Novels, 1850-1888

Furthermore, although our results for French authors are more preliminary, a test of six works by Gobineau shows that two of his earliest texts, written before any extensive travel, are quite similar to some of his latest texts (see Fig. 5).

We are still developing the techniques for the study of the effects of travel, but preliminary testing based on a division of each author's career into early, middle, and late periods allows us to check for consistent trends rather than simple differences between early and late texts, and to begin comparing the four categories of authors. Choosing novelists with long careers allows us to separate the three periods, selecting natural gaps in publication where possible, but creating such gaps where necessary by omitting some texts from the study. For traveling authors, these divisions also take into account the timing of their travel, creating gaps that include the travel.

Three stylistic periods create six patterns of highest, middle, and lowest frequency for each word in the an author's texts. Depending on the number and size of the novels, we include approximately the 8,000 to 10,000 most frequent words, all those frequent enough to show a clear increase or decrease in frequency; we delete words that appear in only one period. Thus, as shown in Fig. 6, each of the six patterns would be expected to occur about one-sixth (17%) of the time by chance.

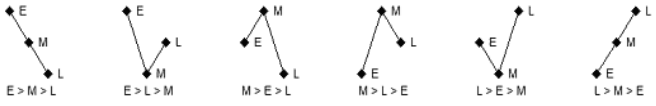


Fig. 6 –Six Patterns of Change ("E" = early; "M" = middle; "L" = late)

Results for the English authors, shown in Fig. 7, are both surprising and suggestive (note that the axis crosses at the "expected" 16.7% level, so that it is easy to see which patterns are more or less frequent than expected for each author. Gradual decrease in frequency, E > M > L, is the only pattern more frequent than expected for all four authors (Meredith's figure is only very slightly more frequent than expected), and both M > E > L and L > E > M are less frequent than expected for all four authors. Although the patterns for these authors suggest some regularity in the growth and decay of vocabulary, no simple relationship emerges. Consider also the surprising fact that vocabulary richness tends to decrease chronologically for Dickens, Trollope, and possibly Collins, while only Meredith shows increasing vocabulary richness. (These comments are

based on a relatively simple measure of vocabulary richness, the number of different words per 10,000 running words; for more discussion of vocabulary richness, see Tweedie and Baayen, 1998 and Hoover, 2003.) These facts contradict the intuitive assumption that the main trend in a writer's total vocabulary should be the learning of new words.

Similar comparisons are being developed for the French authors in question. The conference presentation will include not only a comparison of style within each language group, but between language groups. Such comparisons also build on strengths of corpus stylistics important to translation (Goldfield 2006) and a possible related future for comparative literature (Apter 2005).

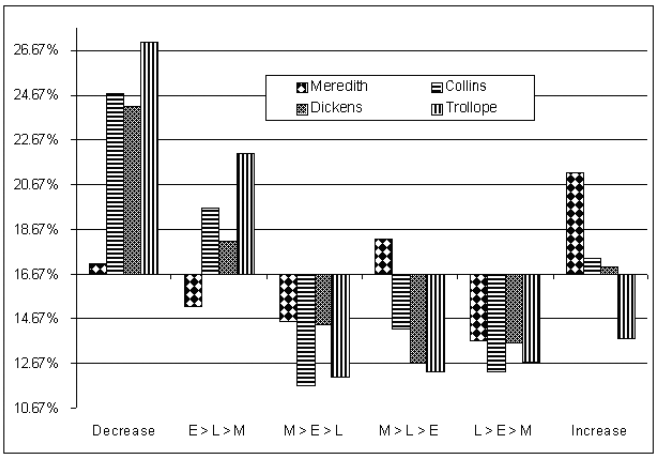


Fig. 7 – Patterns of Frequency Change in the Words of Six Authors.

References

Apter, Emily (2006) *The Translation Zone: A New Comparative Literature,* Princeton.

Burrows, J. F. (1992a). "Computers and the study of literature." In Butler, C. S., ed. *Computers and Written Texts*. Oxford: Blackwell, 167-204.

Goldfield, J. (2006) "French-English Literary Translation Aided by Frequency Comparisons from ARTFL and Other Corpora," *DH2006: Conference Abstracts*: 76-78.

Hoover, D. L. (2003) "Another Perspective on Vocabulary Richness." Computers and the Humanities, 37(2), 151-78.

Hoover, D. L. (2007) "Corpus Stylistics, Stylometry, and the Styles of Henry James," *Style* 41(2) 2007: 174-203.

Stamou, Constantina. (2007) "Stylochronometry: Stylistic Development, Sequence of Composition, and Relative Dating." *LLC* Advance Access published on October 1, 2007.

Tweedie, F. J. and R. H. Baayen. 1998. "How Variable May a Constant Be? Measures of Lexical Richness in Perspective". *Computers and the Humanities* 32 (1998), 323-352.

The Heinrich-Heine-Portal. A digital edition and research platform on the web

(www.heine-portal.de)

Nathalie Groß
hhp@uni-trier.de
University Trier, Germany

Christian Liedtke
liedtke@math.uni-duesseldorf.de
Heinrich-Heine-Institut Düsseldorf, Germany

In this paper we are presenting the Heinrich Heine Portal, one of the most sophisticated web resources dedicated to a German classical author. It virtually combines the two most definitive critical editions (DHA=Düsseldorfer Heine-Ausgabe and HSA=Heine Säkularausgabe) in print together with digital images of the underlying textual originals within an elaborated electronic platform. The project, which has been established in 2002, is organized as a co-operation between the Heinrich-Heine-Institut (Düsseldorf) and the Competence Centre for Electronic Publishing and Information Retrieval in the Humanities (University of Trier). It has been funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Organisation) as well as the Kunststiftung Nordrhein-Westfalen (Foundation for the Arts of North Rhine Westphalia). The work within the project consists of two major parts. On the one hand it has to transfer the printed texts into a digital representation which serves as the basis for its electronic publication. On the other hand it aims at a complete revision of important parts of the printed critical edition and provides new results of the Heinrich-Heine research community.

The first part of the workflow is organized as a typical retro-digitization project. Starting with the printed editions, consisting of nearly 26,500 pages with an amount of about 72 millions of characters, the text was sent to a service partner in China, where it was typed in by hand. During this process two digital versions of the text were produced and then sent back to Germany, where they were automatically collated. After that the listed differences were manually corrected by comparing them with the printed original. The result of this step is a digital text that corresponds with the printed version providing a quality of nearly 99.995%. The next task was to transform this basic digital encoding into a platform independent representation, which then can be used as the main data format for all following project phases. In order to achieve this goal a pool of program routines was developed which uses the typographical information and contextual conditions to generate XML-markup according to the TEI guidelines. Because of the heterogeneous structures of the text (prose, lyrics, critical apparatus, tables of contents etc.) this was a very time consuming step. At the end of this process a reusable version of the data that can be put

into the archives on a long time basis exists. Using the XML encoding the digital documents were imported into an online database, where different views onto the data are stored separately, e.g. metadata for the main information about the letter corpus (list of senders, addressees, dates and places etc.) or bibliographical information on the works. In order to get access to the information from the database the project has been supplied with a sophisticated graphical user interface, which has been built with the help of a content management framework (ZOPE).

Besides providing online access to the basic editions DHA and HSA the Heinrich-Heine-Portal offers a completely revised and updated version of the letter corpus and displays newly-discovered letters and corrigenda, which are being added constantly. Additionally, the portal is an electronic archive which contains and presents more than 12,000 digital images of original manuscripts and first editions, linked to the text and the apparatus of the edition. Most of those images were made available by the Heinrich-Heine-Institut, Düsseldorf, which holds nearly 60% of the Heine-manuscripts known today. The collection of the Heine-Institut was completely digitized for this project. In addition, 32 other museums, libraries and literary archives in Europe and the United States are cooperating with the Heine-Portal and have given us consent to present manuscripts from their collections. Among them are the British Library and the Rothschild Archive in London, the Russian Archives of Social and Political History in Moscow, the Foundation of Weimar Classics and many others. One of the long-term goals of the Heine-Portal is a "virtual unification" of all of Heine's manuscripts. Beyond that the Heine-Portal offers extensive bibliographies of primary sources and secondary literature, from Heine's own first publications in a small journal (1817) up to the most recent secondary literature. Other research tools of the Heine-Portal are a powerful search engine and a complex hyperlink structure which connects the texts, commentaries and the different sections of the Portal with each other and a database of Heine's letters with detailed information on their edition, their availability and the institutions which hold them.

Our paper will describe the technical and philological aspects of the work process that was necessary for the completion of the Heine-Portal, it will give an account of its main functions and demonstrate their use for students and scholars alike, and it will discuss possible pedagogical applications for schools as well as university teaching.

Bibliography

Bernd Füllner and Johannes Fournier: Das Heinrich-Heine-Portal. Ein integriertes Informationssystem. In *Standards und Methoden der Volltextdigitalisierung. Beiträge des Internationalen Kolloquiums an der Universität Trier, 8./9. Oktober 2001*. Hrsg. von Thomas Burch, Johannes Fournier, Kurt Gärtner u. Andrea Rapp. Trier 2002 (Abhandlungen der Akademie der Wissenschaften und der Literatur, Mainz; Geistes- und Sozialwissenschaftliche Klasse), pp. 239-263.

Bernd Füllner and Christian Liedtke: Volltext, Web und Hyperlinks. Das Heinrich-Heine-Portal und die digitale Heine-Edition. In *Heine-Jahrbuch* 42, 2003, pp. 178-187.

Bernd Füllner and Christian Liedtke: Die Datenbanken des Heinrich-Heine-Portals. Mit fünf unbekannten Briefen von und an Heine. In *Heine-Jahrbuch* 43, 2004, pp. 268-276.

Nathalie Groß: Das Heinrich-Heine-Portal: Ein integriertes Informationssystem. In *Uni-Journal* 3/2004, pp. 25-26.

Nathalie Groß: Der Digitale Heine – Ein Internetportal als integriertes Informationssystem. In *Jahrbuch für Computerphilologie - online*. Hg. v. Georg Braungart, Karl Eibl und Fotis Jannidis. München 2005, pp. 59-73. Available online: http://computerphilologie.uni-muenchen.de/jg04/gross/gross.html

Christian Liedtke: Die digitale Edition im Heinrich-Heine-Portal – Probleme, Prinzipien, Perspektiven. In *editio. Internationales Jahrbuch für Editionswissenschaft*. Tübingen 2006.

Bernd Füllner and Nathalie Groß: Das Heinrich-Heine-Portal und digitale Editionen. Bericht über die Tagung im Heinrich-Heine-Institut in Düsseldorf am 6. Oktober 2005. In *Heine-Jahrbuch*. 45. 2006, pp. 240-248.

The New Middle High German Dictionary and its Predecessors as an Interlinked Compound of Lexicographical Resources

Kurt Gärtner
gaertnek@staff.uni-marburg.de
Universität Trier, Germany

The paper gives, firstly, a brief overview over the history of Middle High German (MHG) lexicography leading to the new MHG dictionary. Secondly, it describes the digitisation of the old MHG dictionaries which form a subnet in a wider net of German dictionaries. Thirdly, this network, together with specific tools and digital texts, is seen as an essential source for compiling the new MHG dictionary, of which, fourthly, the essential features, esp. its online version, are mentioned briefly, but will be demonstrated in detail in the full paper.

1. Looking back at the beginnings of Humanities Computing in the sixties, one of the major subjects was the support the computer would give to lexicographers. Many e-texts have been created then for mere lexicographical purposes (Wisbey 1988). The production of concordances and indices was at its heights between 1970 and 1980 (cf. Gärtner 1980). However, the road to real dictionary making came only gradually in sight when lemmatization was required and semantics began to dominate the discussion of computer application in lexicography. The work on the dictionaries to the medieval Germanic languages benefited a great deal from this development. This has certainly been the case regarding MHG, the period of German from ca. 1050 to ca. 1350 resp. – in the earlier periodisations – to 1500. During the 20th century, the situation of MHG lexicography had become more and more deplorable. Since Lexer's dictionary appeared in 1878, all plans for a new dictionary had been unsuccessful. Therefore, many editors of MHG texts published after 1878 provided glossaries as a means of presenting such material that was not included in the earlier dictionaries edited by Benecke/Müller/Zarncke (BMZ) and Lexer with its *Nachträge*. It was not until 1985 that a group of researchers at the University of Trier started compiling the *Findebuch* which was published in 1992. The *Findebuch* was a compilation of glossaries to editions in order to make up for the lack of a new MHG dictionary. Also in the late eighties, on the basis of the work on the *Findebuch*, planning began to work out a new MHG dictionary (cf. Gärtner/Grubmüller 2000). This time the plans were successful, and work on the new dictionary started in 1994. The scientific use of the computer had meanwhile become a *conditio sine qua non*, also for the new MHG dictionary (cf. Gärtner/Plate/Recker 1999).

2. The work on a new dictionary relies to a great deal on its predecessors. As the *Findebuch* had been compiled with extensive use of the computer and existed in machine readable form, the need for digitizing the old MHG dictionaries was strongly felt from the beginning of the new dictionary's planning. The old dictionaries (BMZ, Lexer with *Nachträge*) and the *Findebuch* are closely interconnected and can only be used simultaneously. They were ideal candidates for the composition of an electronic dictionary compound. Therefore, as a supporting research project to the new dictionary, the old ones were digitized and interlinked thus forming a digital network (MWV). The work on this project has been described to the LLC readers in detail (Fournier 2001; see also Burch/Fournier 2001). Moreover, this network formed the starting point for a wider net that interlinks many more dictionaries to the German language and its dialects (cf. Woerterbuchnetz). The most prominent among them is the 33vols. *Deutsches Wörterbuch* (DWB) by the brothers Grimm.

3. The work on the new MHG dictionary (MWB) was considerably supported by two more projects. First and most essential for lexicographers working in different places was the implementation of a web based workbench for the composition of dictionaries. The development of a toolkit for a collaborative editing and publishing of the MWB has also been described to LLC readers in detail (Queens/Recker 2005). The second project supporting the work on the new dictionary was the extension of the digital text archive. At the beginning of the work on the MWB, the lexicographical workbench had to rely only on a small corpus of digital texts. This number was effectively increased by a collaborative project with the *Electronic Text Center* of the University of Virginia at Charlottesville (cf. Recker 2002 and MHDTA/MHGTA 2004). By now nearly all the texts of the *Findebuch* corpus have been digitized and added to the corpus of the new dictionary.

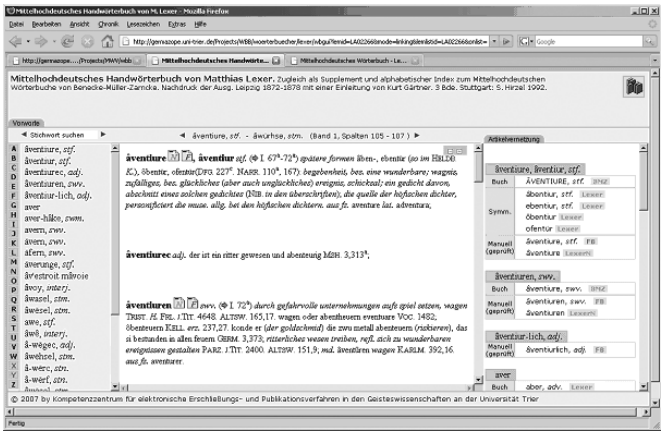
4. In 2006 the first instalment of the new MWB was published together with a CD-ROM containing a PDF of the printed form. The online version was to follow in 2007.

All digital resources created for the new dictionary are firstly to the advantage of the lexicographer, and secondly and still more important to the advantage of the user as well. The lexicographer's workbench provides an alphabetical list of headwords (*Lemmaliste*), it offers an easy and quick access to the old dictionaries (BMZ, Lexer, *Findebuch*), to the lemmatized archive of instances (*Belegearchiv*) in form of a KWIC-concordance, and finally to the whole corpus of e-texts which can be searched for more evidences. Not only the lexicographer, but also the user of the online version has access to all information on the usage of a MHG word, as I will demonstrate in detail.

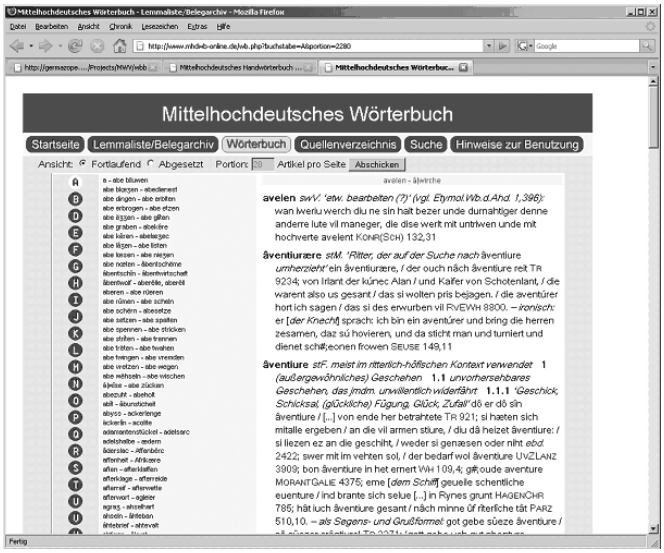
As the new MWB is a long term project, the online version not only offers the already published instalments, but also the complete material the forthcoming fascicles are based upon. Furthermore, for the entries still to be worked out the old dictionaries are always present for consultation together with the new material of the archive and the corpus of e-texts. Thus, linking old and new lexicographical resources

proves immensely valuable not only for the workbench of the lexicographer, but also for the user of his work.

Images



Lexer's MHG Dictionary within the Dictionary Net



The Online Version of the New MHG Dictionary

Bibliography

(Apart from the dictionaries, preferably research papers in English are listed)

BMZ = Georg Friedrich Benecke, Wilhelm Müller, Friedrich Zarncke: *Mittelhochdeutsches Wörterbuch*. Leipzig 1854–1866. Nachdruck mit einem Vorwort und einem zusammengefaßten Quellenverzeichnis von Eberhard Nellmann sowie einem alphabetischen Index von Erwin Koller, Werner Wegstein und Norbert Richard Wolf. Stuttgart 1990.

Burch, Thomas / Fournier, Johannes / Gärtner, Kurt / Rapp, Andrea (Eds.): *Standards und Methoden der Volltextdigitalisierung. Beiträge des Internationalen Kolloquiums an der Universität Trier, 8./9. Oktober 2001* (Abhandlungen der Akademie der Wissenschaften und der Literatur, Mainz; Geistes- und sozialwissenschaftliche Klasse; Einzelveröffentlichung Nr. 9). Stuttgart 2002.

Burch, Thomas / Fournier, Johannes: Middle High German Dictionaries Interlinked. In: Burch/Fournier/Gärtner/Rapp 2002, p. 297-301. Online-version: http://germazope.uni-trier.de/Projects/MWV/publikationen

DWB = *Deutsches Wörterbuch* der Brüder Grimm. Erstbearbeitung. 33 vols. Leipzig 1864-1971. Online-Version (OA): http://www.DWB.uni-trier.de Offline-Version: *Der Digitale Grimm*. Deutsches Wörterbuch der Brüder Grimm. Elektronische Ausgabe der Erstbearbeituung. Bearbeitet von Hans-Werner Bartz, Thomas Burch, Ruth Christmann, Kurt Gärtner, Vera Hildenbrandt, Thomas Schares und Klaudia Wegge. Hrsg. vom Kompetenzzentrum für elektronische Erschließungs- und Publikationsverfahren in den Geisteswissenschaften an der Universität Trier in Verbindung mit der Berlin-Brandenburgischen Akademie der Wissenschaften. 2 CD-ROMs, Benutzerhandbuch, Begleitbuch. 1. Auflage. Frankfurt am Main: Zweitausendeins 2004, 5. Auflage, 2007.

Findebuch = Kurt Gärtner, Christoph Gerhardt, Jürgen Jaehrling, Ralf Plate, Walter Röll, Erika Timm (Datenverarbeitung: Gerhard Hanrieder): *Findebuch zum mittelhochdeutschen Wortschatz. Mit einem rückläufigen Index*. Stuttgart 1992.

Fournier, Johannes: New Directions in Middle High German Lexicography: Dictionaries Interlinked Electronically. In: *Literary and Linguistic Computing* 16 (2001), p. 99-111.

Gärtner, Kurt: Concordances and Indices to Middle High German. In: *Computers and the Humanities* 14 (1980), p. 39-45.

Gärtner, Kurt: Comprehensive Digital Text Archives: A Digital Middle High German Text Archive and its Perspectives, in: First EU/NSF Digital Libraries All Projects Meeting, Rome March 25-26, 2002. http://delos-noe.iei.pi.cnr.it/activities/internationalforum/All-Projects/us.html

Gärtner, Kurt / Wisbey, Roy: Die Bedeutung des Computers für die Edition altdeutscher Texte. In: *Kritische Bewahrung. Beiträge zur deutschen Philologie*. Festschrift für Werner Schröder zum 60. Geburtstag. Hrsg. von E.-J. Schmidt. Berlin 1974, p. 344-356.

Gärtner, Kurt / Grubmüller, Klaus (Eds.): *Ein neues Mittelhochdeutsches Wörterbuch: Prinzipien, Probeartikel, Diskussion* (Nachrichten der Akademie der Wissenschaften in Göttingen. I. Philologisch-historische Klasse, Jg. 2000, Nr. 8). Göttingen 2000.

Gärtner, Kurt / Plate, Ralf / Recker, Ute: Textvorbereitung und Beleggewinnung für das Mittelhochdeutsche Wörterbuch. In: *Literary and Linguistic Computing* 14, No. 3 (1999), p. 417-423.

Lexer = Lexer, Matthias: *Mittelhochdeutsches Handwörterbuch.* Nachdruck der Ausgabe Leipzig 1872-1878 mit einer Einleitung von Kurt Gärtner. 3 Bde. Stuttgart 1992.

Lexer Nachträge = Matthias Lexer: *Nachträge zum mittelhochdeutschen Handwörterbuch.* Leipzig 1878.

MHDTA/MHGTA = Final report on the Project „Digital Middle High German Text Archive" to the DFG and NSF. Charlottesville / Trier 2004. http://www.mhgta.uni-trier.de/MHGTA_Final_Report.pdf

MWB = Mittelhochdeutsches Wörterbuch Online-Version (OA): http://mhdwb-online.de Print-Version with CD: *Mittelhochdeutsches Wörterbuch.* Im Auftrag der Akademie der Wissenschaften und der Literatur Mainz und der Akademie der Wissenschaften zu Göttingen hrsg. von Kurt Gärtner, Klaus Grubmüller und Karl Stackmann. Erster Band, Lieferung 1ff. Stuttgart 2006ff. Project Homepage: http://www.mhdwb.uni-trier.de/

MWV = Mittelhochdeutsche Wörterbücher im Verbund. [Middle High German Dictionaries Interlinked] Online-Version (OA): http://www.MWV.uni-trier.de Offline-Version: *Mittelhochdeutsche Wörterbücher im Verbund.* Hrsg. von Thomas Burch, Johannes Fournier, Kurt Gärtner. CD-ROM und Begleitbuch. Stuttgart 2002.

Queens, Frank; Recker-Hamm, Ute: A Net-based Toolkit for Collaborative Editing and Publishing of Dictionaries. In: *Literary and Linguistic Computing* 20 (2005), p. 165-175.

Recker, Ute: Digital Middle High German Text Archive. In: Burch / Fournier / Gärtner / Rapp 2002, p. 307-309.

Wisbey, Roy: Computer und Philologie in Vergangenheit, Gegenwart und Zukunft. In: *Maschinelle Verarbeitung altdeutscher Texte IV.* Beiträge zum Vierten Internationalen Symposion, Trier 28. Februar bis 2. März 1988. Hrsg. von Kurt Gärtner, Paul Sappler und Michael Trauth. Tübingen 1991, p. 346-361.

Wörterbuch-Netz: http://www.woerterbuchnetz.de

Domestic Strife in Early Modern Europe: Images and Texts in a virtual anthology

Martin Holmes
mholmes@uvic.ca
University of Victoria, Canada

Claire Carlin
ccarlin@uvic.ca
University of Victoria, Canada

The seventeenth-century engravings and texts collected for the project Le mariage sous l'Ancien Régime: une anthologie virtuelle all belong to the polemical genre. The institution of marriage was undergoing intense scrutiny and criticism in light of Reformation questioning of Catholic practices, and popular discourse and images reflected this malaise. The cuckold attacked both verbally and physically by a nagging wife, or, conversely, a sassy wife receiving her "correction" are themes of the Middle Ages reproduced frequently during the early modern period, but with new sophistication as engravings became the primary site for the representation of problems with marriage, as DeJean has shown. Whereas polemical writings flourished in the first half of the century according to Banderier and Carlin 2002, images gained more complexity in design and incorporated increasing amounts of commentary during the reign of Louis XIV (1661-1715). New variations on medieval topics occur, as bourgeois women in salons comment on marriage or wrestle with their husbands over the pants in the family. A novel twist on spousal conflict appears in engravings such as "L'invention des femmes" (Lagniet) and "Operateur céphalique" (Anon.): inspired by the Renaissance interest in dissection, the notion that behaviour could be modified through radical brain surgery introduces a new kind of violence into marriage satire.

From the beginning of the project, our approach to construction of the corpus has been based on the ideal that texts and images must both be full and equivalent members of the collection. Images have been treated as texts, and our goal has been to make explicit, in textual annotations, as much of the significant information they encode as possible. In this respect, the Mariage project is very different from other markup projects which integrate images and texts, such as those described in Porter 2007; while projects such as Pembroke 25 (Szarmach & Hall n.d.) and the Electronic Aelfric integrate images and text in sophisticated ways, the images are of manuscript pages, and the at the heart of these digital editions is the process of transcription. By contrast, while the engravings in the Mariage collection may include fairly extensive textual components — see, for instance, "Le Fardeau du Menage" (Guérard 1712), which incorporates over 60 lines of verse — the text is ancillary to the scenes depicted. Thus, the engravings in the Mariage project are somewhere on a continuum between the page-images of a digital edition of a manuscript, and the images

of artworks in a project such as the On-Line Picasso Project (Mallen n.d.).

At the simplest level, marking up the images means transcribing and clarifying any text which actually appears on the engravings. This makes the text on the images as searchable as that of the poetry and prose. Following this, we have begun to identify and discuss the significance of individual figures in the engravings, as well as linking these figures to similar elements in other engravings, or to texts in the collection. Figure 1 shows this kind of linking in as it appears to the reader in a Web browser. The wife's lover is a recurrent figure in images and texts of the period, and the annotation links to textual references, and also to a segment of another image.

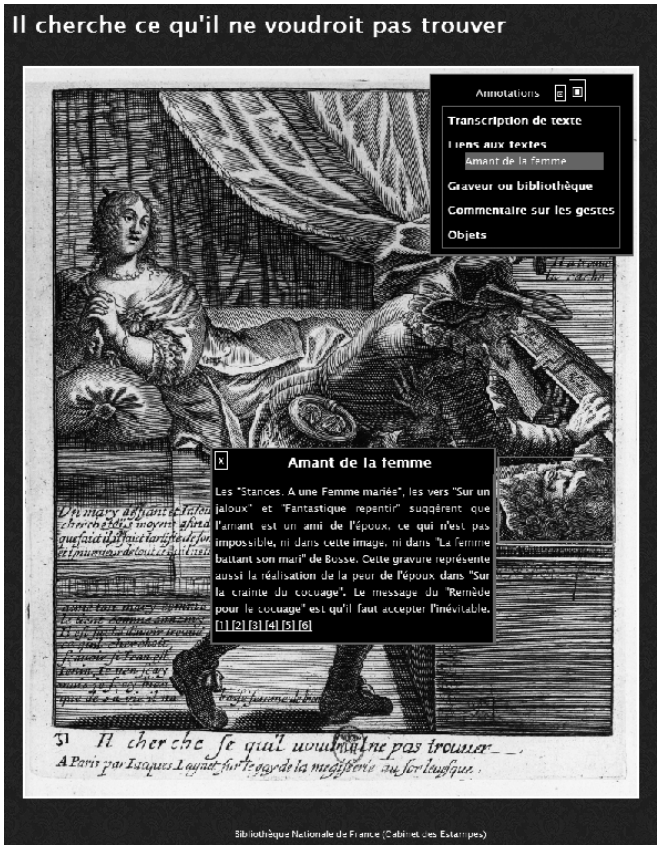


Figure 1: Linking between image areas and other texts.

In addition, the search system on the site has been designed to retrieve both blocks of text and segments of images, based on annotations to the images. Figure 2 shows part of the result set from a search for "frapp*", including sections of images and a line from a poem. In the case of an image hit, clicking on the link will lead to the annotated image with the "hit annotation" selected.

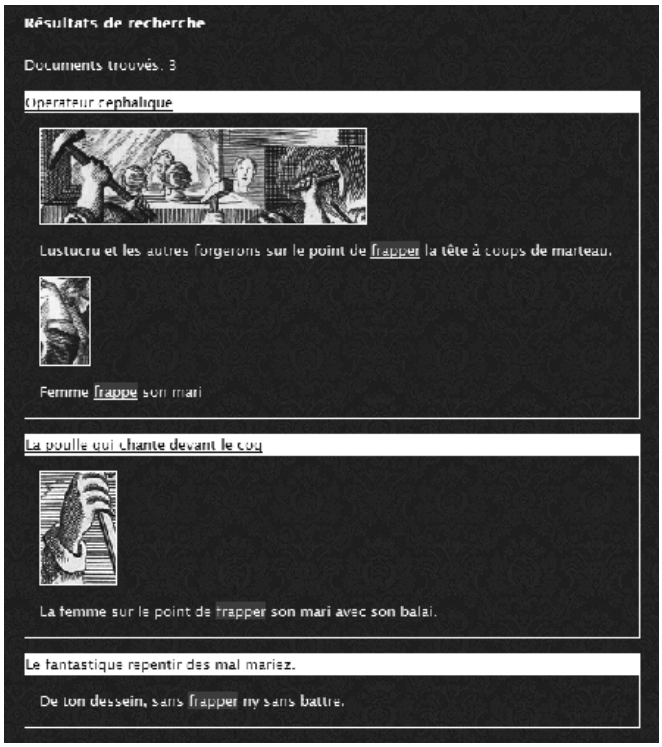


Figure 2: Results of a search, showing sections of images as well as a poem retrieved from the database.

We are now marking up symbols and devices, and have begun to contemplate the possibility that we might mark up virtually every object which appears in the engravings. Although this would be rather laborious, it would enable the discovery of more correspondences and perhaps more symbolism than is currently apparent. To be able to search for "poule", and retrieve all the depictions of hens which appear in the collection (as well as instances of the word itself), and view them together on the same page, would provide a powerful tool for researchers. This level of annotation (the identification and labelling of everyday objects) does not require any significant expertise, but it will contribute a great deal to the value of the collection.

Using the combination of textual and image markup, our team has been able to uncover links across the decades as well as slowly developing differences among diverse types of polemical images. Research questions we have begun to explore include the following:

- How does the use of key vocabulary evolve over the sixteenth to eighteenth centuries?

- How does the use of objects, especially the instruments of violence, change over this period?

- How does the use of stock verses and verse forms develop over the course of this period?

In addition to discussing the research tools and insights emerging out of the *Mariage* project, this presentation will look at the development of the *Image Markup Tool,* an open-

source Windows application which was created initially for use in the Mariage project, but which is now used in a range of other projects as well. The *IMT* has acquired a number of new features in response to the requirements of the Mariage project, and is also developing in reaction to changes in TEI P5. The original approach of the *IMT* was to blend SVG with TEI to produce a dual-namespace document (Carlin, Haswell and Holmes 2006), but the incorporation of the new <facsimile>, <surface> and <zone> elements in the TEI transcr module (TEI Consortium 2007, 11.1) now provide a native TEI framework which appears suited to image markup projects such as Mariage, and in December 2007, a new version of the *Image Markup Tool* was released, which relinquishes the SVG-based approach in favour of a pure TEI schema. This simplification of the file format will, we hope, make it much easier for end-users to integrate markup produced using the IMT into Web applications and other projects.

Further developments in the *IMT* are projected over the next few months. The current version of the application is really focused on marking up the single, standalone images which are central to the Mariage project. However, <facsimile> is described in the TEI Guidelines as containing "a representation of some written source in the form of a set of images rather than as transcribed or encoded text" (TEI Consortium 2007, 11.1). The *IMT* really ought to be capable of producing documents which contain multiple images, and we plan to extend it to make this possible.

References

Anon. n.d. "Operateur Cephalique". Bibliothèque Nationale de France Cote RES TF 7 4e (Cliché P123-IFFNo190). [http://mariage.uvic.ca/xhtml.xq?id=operateur_cephalique] Accessed 2007-11-05.

Banderier, G. "*Le mariage au miroir des poètes satiriques français (1600-1650)." In Le mariage dans l'Europe des XVIe et XVIIe siècles: réalités et représentations,* vol. 2. Ed. R. Crescenzo et al. Presses de l'Université Nancy II, 200. 243-60.

Carlin, C. "Misogamie et misogynie dans les complaintes des mal mariés au XVIIe siècle." In *La Femme au XVIIe siècle*. Ed. Richard G. Hodgson. Biblio 17, vol. 138. Tübingen: Gunter Narr Verlag, 2002. 365-78.

Carlin, C., E. Haswell and M. Holmes. 2006. "Problems with Marriage: Annotating Seventeenthcentury French Engravings with TEI and SVG." *Digital Humanities 2006 Conference Abstracts*. 39-42. [http://www.allc-ach2006.colloques.paris-sorbonne.fr/DHs.pdf]. July 2006. Accessed 2007-11-05.

Carlin, C. et al. 2005-2007. *Le mariage sous L'Ancien Régime: une anthologie virtuelle.* [http://mariage.uvic.ca/] Accessed 2007-11-05.

DeJean, J. 2003. "Violent Women and Violence against Women: Representing the 'Strong' Woman in Early Modern France." Signs, 29.1 (2003): 117-47.

Guérard, N. 1712. "Le Fardeau du Menage". Bibliothèque Nationale de France Cote EE 3A PET FOL, P. 7 (Cliché 06019648). [http://mariage.uvic.ca/xhtml.xq?id=fardeau] Accessed 2007-11-05.

Holmes, M. 2007. Image Markup Tool v. 1.7. [http://www.tapor.uvic.ca/~mholmes/image_markup/] Accessed 2007-11-05.

Lagniet, J. ca. 1660. "L'Invention des femmes." Bibliothèque Nationale de France, Cote TE 90 (1) FOL (Cliché 06019644). [http://mariage.uvic.ca/xhtml.xq?id=invention_femmes_1] Accessed 2007-11-05.

Mallen, E., ed. n.d. *The Picasso Project.* Texas A&M University. [http://picasso.tamu.edu/] Accessed 2007-11-05.

Porter, D. C. 2007. "Examples of Images in Text Editing." *Digital Humanities 2007 Conference Abstracts.* 159-160.

Szarmach, Paul, and Thomas H. Hall. n.d. *Digital Edition of Cambridge, Pembroke College MS 25.* (Pembroke 25 project.) Cited in Porter 2007.

TEI Consortium, eds. 2007. "Digital Facsimiles." *Guidelines for Electronic Text Encoding and Interchange.* [Last modified date: 2007-10-28]. [http://www.tei-c.org/release/doc/tei-p5- doc/en/html/PH.html] Accessed 2007-11-07.

Rescuing old data: Case studies, tools and techniques

Martin Holmes
mholmes@uvic.ca
University of Victoria, Canada

Greg Newton
gregster@uvic.ca
University of Victoria, Canada

Much of the work done by digital humanists could arguably be described as "rescuing old data". Digitizing texts is a way of rescuing them from obsolescence, from obscurity, or from physical degradation. However, digitizing in itself does not amount to permanent rescue; it is merely transformation, and digital texts have turned out to have lifetimes vastly shorter than printed texts. As Besser pointed out in 1999,

> "Though most people tend to think that (unlike analog information) digital information will last forever, we fail to realize the fragility of digital works. Many large bodies of digital information (such as significant parts of the Viking Mars mission) have been lost due to deterioration of the magnetic tapes that they reside on. But the problem of storage media deterioration pales in comparison with the problems of rapidly changing storage devices and changing file formats. It is almost impossible today to read files off of the 8-inch floppy disks that were popular just 20 years ago, and trying to decode Wordstar files from just a dozen years ago can be a nightmare. Vast amounts of digital information from just 20 years ago is, for all practical purposes, lost."

As Kirschenbaum (2007) says, "The wholesale migration of literature to a born-digital state places our collective literary and cultural heritage at real risk." To illustrate this point, consider that searches for the original news source of Besser's claim about the Viking Mars mission reveal links to articles in Yahoo, Reuters, and Excite, but all of these documents are now unavailable. Only the Internet Archive has a copy of the story (Krolicki 2001).

This paper will examine two cases in which we have recently had to rescue digitized content from neardeath, and discuss some tools and techniques which we developed in the process, some of which are available for public use. Our experiences have taught us a great deal, not just about how to go about retrieving data from obsolete formats, but also about how better to protect the data we are generating today from obsolescence. Our paper is not intended to address the issue of digital preservation (how best to prevent good digital data from falling into obsolescence); an extensive literature already exists addressing this problem (Vogt-O'Connor 1999, Besser 1999, and others). The fact is that, despite our best efforts, bit-rot is inexorable, and software and data formats always tend towards obsolescence. We are concerned, here, with discussing the techniques we have found effective in rescuing data which has already become difficult to retrieve.

Case Study #1: The Nxaʔamxcín (Moses) Dictionary Database

During the 1960s and 70s, the late M. Dale Kincaid did fieldwork with native speakers of Nxaʔamxcín (Moses), an aboriginal Salish language. His fieldnotes were in the form of a huge set of index cards detailing vocabulary items, morphology and sample sentences. In the early 1990s, the data from the index cards was digitized using a system based on a combination of Lexware and WordPerfect, running on DOS, with the goal of compiling a print dictionary. The project stalled after a few years, although most of the data had been digitized; the data itself was left stored on an old desktop computer. When this computer finally refused to boot, rescuing the data became urgent, and the WordPerfect files were retrieved from its hard drive. We then had to decide what could usefully be done with it. Luckily, we had printouts of the dictionary entries as processed by the Lexware/WordPerfect system, so we knew what the original output was supposed to look like. The data itself was in WordPerfect files.

The natural approach to converting this data would be to open the files in an more recent version of WordPerfect, or failing that, a version contemporary with the files themselves. However, this was not effective, because the files themselves were unusual. In addition to the standard charset, at least two other charsets were used, along with an obsolete set of printer fonts, which depended on a particular brand of printer, and a specific Hercules graphics card. In addition, the original fonts were customized to add extra glyphs, and a range of macros were used. In fact, even the original authors were not able to see the correct characters on screen as they worked; they had to proof their work from printouts. When the files were opened in WordPerfect, unusual characters were visible, but they were not the "correct" characters, and even worse, some instances of distinct characters in the original were collapsed into identical representations in WordPerfect. Attempts to open the files in other word processors failed similarly.

Another obvious approach would be to use libwpd, a C++ library designed for processing WordPerfect documents, which is used by OpenOffice.org and other word-processing programs. This would involve writing handlers for the events triggered during the document read process, analysing the context and producing the correct Unicode characters. Even given the fact that libwpd has only "Support for a substantial portion of the WordPerfect extended character set", this technique might well have succeeded, but the tool created as a result would have been specific only to WordPerfect files, and to this project in particular. We decided that with a similar investment of time, we would be able to develop a more generally-useful tool; in particular, we wanted to create a tool which could be used by a non-programmer to do a similar task in the future.

Comparing the contents of the files, viewed in a hex editor, to the printouts, we determined that the files consisted of:

-blocks of binary information we didn't need (WordPerfect file headers)

-blocks of recognizable text

-blocks of "encoded text", delimited by non-printing characters

The control characters signal switches between various WordPerfect character sets, enabling the printing of non-ascii characters using the special printer fonts. Our task was to convert this data into Unicode. This was essentially an enormous search-and-replace project. Here is a sample section from the source data:

```
.rt?   ÀEOTBELÀÀ1SOHÀkÀSOEOTÀ

1infl   transitive
11tr   ÀEOTBELÀÀ1SOHÀÀUSSOHÀkÀSOEOTÀ©À9SOHÀn
11g   *pull
q   should this be under a separate root?
```

(where [EOT] = end of transmission, [BEL] = bell, [SOH] = start of header, [SO] = shift out, and [US] = unit separator). This image shows the original print output from this data, including the rather cryptic labels such as "11tr" which were to be used by Lexware to generate the dictionary structure:

```
.rt?   √čkʷ

1infl   transitive
11tr   √čákʷ-ən
11g   *pull
q   should this be under a separate root?
```

Working with the binary data, especially in the context of creating a search-and-replace tool, was problematic, so we transformed this into a pure text representation which we could work with in a Unicode text editor. This was done using the Linux "cat" command. The command "cat -v input_file > output_file" takes "input_file" and prints all characters, including non-printing characters, to "output_file", with what the cat manual refers to as "nonprinting characters" encoded in "hat notation". This took us from this:

```
À[EOT][BEL]ÀÀ1[SOH]ÀkÀ[SO][EOT]À
```

to this:

```
M-@^D^GM-@M-@1^AM-@kM-@^N^DM-@
```

From here, our task was in two stages: to go from the hat-notation data to a Unicode representation, like this:

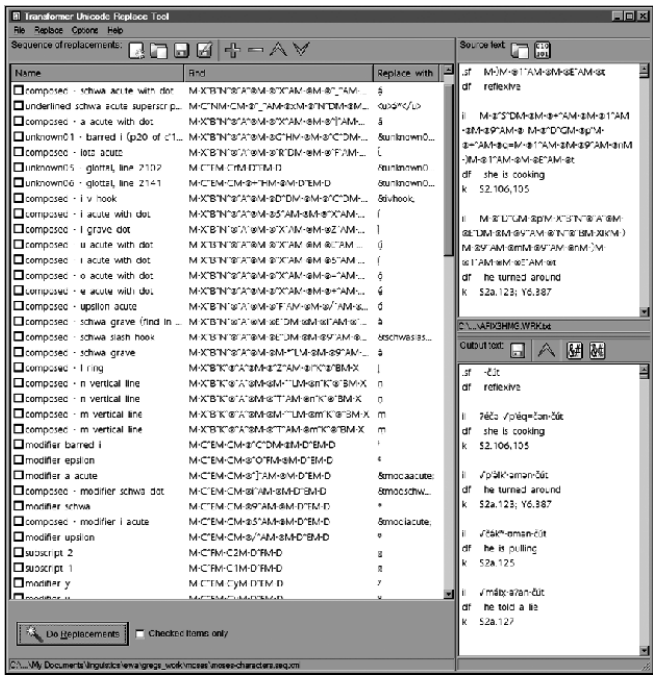
```
<ENTRY level="001"  id="√čkʷ">
<rt>√čkʷ</rt>
<infl mode="1">transitive</infl>
<tr mode="11">√čákʷ-ən</tr>
<g mode="11">*pull</g>
<q>should this be under a separate root?</q>
</ENTRY>
```

and thence to a TEI XML representation:

```
<entry xml:id="čkW">
   <form>
      <pron>
         <seg>čkʷ</seg>
      </pron>
      <hyph>√<m sameAs="">čkʷ</m>
      </hyph>
   </form>
      [...]
</entry>
```

In the first stage, we established a table of mappings between control character sequences and Unicode sequences. However, dozens of such sequences in our data contained overlaps; one sequence mapping to one character might appear as a component of a larger sequence mapping to a different character. In other words, if we were to reconstruct the data using search-and-replace operations, the order of those operations would be crucial; and in order to fix upon the optimal progression for the many operations involved, some kind of debugging environment would be needed.

This gave rise to the Windows application Transformer, an open-source program designed for Unicodebased search-and-replace operations. It provides an environment for specifying, sequencing and testing multiple search-and-replace operations on a text file, and then allows the resulting sequence to be run against a batch of many files. This screenshot shows Transformer at work on one of the Moses data files.



The ability to test and re-test replacement sequences proved crucial, as we discovered that the original data was inconsistent. Data-entry practices had changed over the lifetime of the project. By testing against a range of files from different periods and different data-entry operators, we were able to devise a sequence which produced reliable results across the whole set, and transform all the data in one operation.

Having successfully created Unicode representations of the data, we were now able to consider how to convert the results to TEI XML. The dataset was in fact hierarchically structured, using a notation system which was designed to be processed by Lexware. First, we were able to transform it into XML using the Lexware Band2XML converter (http://www.ling.unt.edu/~montler/convert/Band2xml.htm); then an XSLT transformation took us to TEI P5.

The XML is now being proofed and updated, and an XML database application is being developed.

Case Study #2: The Colonial Despatches project

During the 1980s and 1990s, a team of researchers at the University of Victoria, led by James Hendrickson, transcribed virtually the entire correspondence between the colonies of British Columbia and Vancouver Island and the Colonial Office in London, from the birth of the colonies until their incorporation into the Dominion of Canada. These documents include not only the despatches (the spelling with "e" was normal in the period) between colonial governors and the bureacracy in London; each despatch received in London went through a process of successive annotation, in the form of bureaucratic "minutes", through which its significance was discussed, and appropriate responses or actions were mooted, normally leading to a decision or response by a government minister. These documents are held in archives in BC, Toronto and in the UK, and were transcribed both from originals and from microfilm. It would be difficult to overestimate the historical significance of this digital archive, and also its contemporary relevance to the treaty negotiations which are still going on between First Nations and the BC and Canadian governments.

The transcriptions take the form of about 9,000 text files in Waterloo SCRIPT, a markup language used primarily for preparing print documents. 28 volumes of printed text were generated from the original SCRIPT files, and several copies still exist. The scale of the archive, along with the multithreaded and intermittent nature of the correspondence, delayed as it was by the lengthy transmission times and the range of different government offices and agents who might be involved in any given issue, make this material very well-suited to digital publication (and rather unwieldy and difficult to navigate in print form). Our challenge is converting the original script files to P5 XML.

This is an example of the Waterloo Script used in this project:

```
.sr $$cdorigin = Douglas ;.sr $$cdaddres = Lytton
..cm ================================================
...DES;.des No. 3
.'ref 12721, CO 60/1, p. 213; received 14 December
.adr Victoria Vancouver's Island
.dat 12 October 1858
.br;Sir
.par;1. I take the liberty of submitting for the information of Her
Majesty's Government, a report on my observations on the state of public
.ix 'Gold fields' 'conditions at'
.ix 'Douglas, James' 'inspection of gold fields'
affairs, during a late visit to Fraser's River, necessarily brief, as my
time is engrossed not only with the Executive duties of Government, but
also in attending to all the details of inferior departments, which must
hereafter devolve on other officers.
```

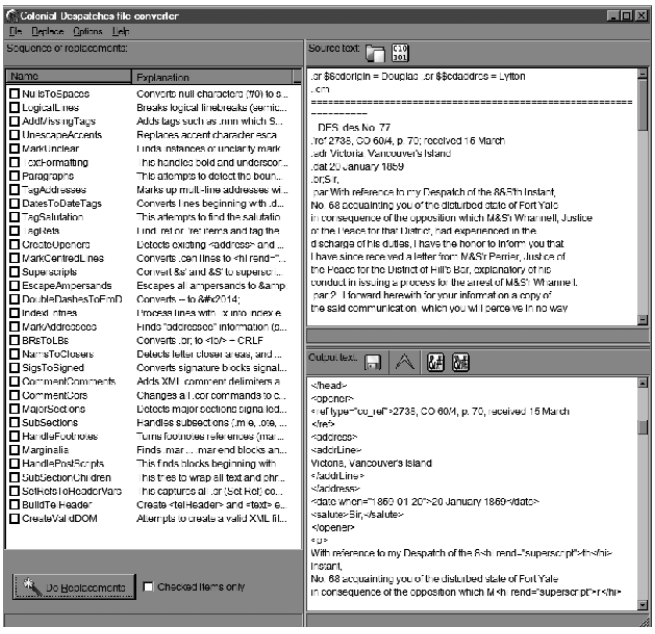
We can see here the transcribed text, interspersed with structural milestones (.par;), editorial annotations such as index callouts and footnote links, and other commands such as .adr, which are actually user-defined macros that invoke sets of formatting instructions, but which are useful to us because they help identify and categorize information (addresses, dates, etc.). Although this is structured data, it is far from from ideal. It is procedural rather than hierarchical (we can see where a paragraph begins, but we have to infer where it ends); it is only partially descriptive; and it mixes editorial content with transcription.

This is rather more of a challenge than the Moses data; it is much more varied and more loosely structured. Even if a generic converter for Waterloo SCRIPT files were available (we were unable to find any working converter which produced useful output such as XML), it would essentially do no more than produce styling/printing instructions in a different format; it would not be able to infer the document structure, or convert

(say) a very varied range of handwritten date formats into formal date representations. To create useful TEI XML files, we need processing which is able to make complex inferences from the data, in order to determine for instance where for example an <opener> begins and ends; decide what constitutes a <salute></salute>; or parse a human-readable text string such as "12719, CO 60/1, p. 207; received 14 December" into a structured reference in XML.

The most effective approach here was to write routines specific to these texts and the macros and commands used in them. As in the Moses project, we used the technique of stringing together a set of discrete operations, in an environment where they could be sequenced, and individual operations could be turned on and off, while viewing the results on individual texts. We gutted the original Transformer tool to create a shell within which routines programmed directly into the application were used in place of search-and-replace operations. The resulting interface provides the same options for sequencing and suppressing operations, and for running batch conversions on large numbers of files, but the underlying code for each operation is project-specific, and compiled into the application.

This screenshot shows the application at work on a despatch file:



The bulk of the original SCRIPT files were converted in this way, at a "success rate" of around 98% --meaning that at least 98% of the files were converted to well-formed XML, and proved valid against the generic P5 schema used for the project. The remaining files (around 150) were partially converted, and then edited manually to bring them into compliance. Some of these files contained basic errors in their original encoding which precluded successful conversion. Others were too idiosyncratic and complex to be worth handling in code. For instance, 27 of the files contained "tables" (i.e. data laid out in tabular format), each with distinct formatting settings, tab width, and so on, designed to lay them out effectively on a page printed by a monospace printer. The specific formatting information was not representative of anything in the original documents (the original documents are handwritten, and very roughly laid out); rather, it was aimed specifically at the print process. In cases like this, it was more efficient simply to encode the tables manually.

The project collection also contains some unencoded transcription in the form of Word 5 documents. To retrieve this data and other similar relics, we have built a virtual machine running Windows 3.11, and populated it with DOS versions of Word, WordStar and WordPerfect. This machine also provides an environment in which we can run the still-famous TACT suite of DOS applications for text-analysis. The virtual machine can be run under the free VMWare Server application, and we hope to make it available on CD-ROM at the presentation.

Conclusions

The data files in both the projects we have discussed above were all in standard, documented formats. Nevertheless, no generic conversion tool or process could have been used to transform this data into XML, while preserving all of the information inherent in it. We were faced with three core problems:

- **Idiosyncracy**. Waterloo SCRIPT may be well documented, but any given corpus is likely to use macros written specifically for it. WordPerfect files may be documented, but issues with character sets and obsolete fonts can render them unreadable. Every project is unique.

- **Inconsistency**. Encoding practices evolve over time, and no project of any size will be absolutely consistent, or free of human error. The conversion process must be flexible enough to accommodate this; and we must also recognize that there is a point at which it becomes more efficient to give up, and fix the last few files or bugs manually.

- **Non-explicit information.** Much of the information we need to recover, and encode explicitly, is not present in a mechanical form in the original document. For example, only context can tell us that the word "Sir" constitutes a <salute>; this is evident to a human reader of the original encoding or its printed output, but not obvious to a conversion processor, unless highly specific instructions are written for it.

In Transformer, we have attempted to create a tool which can be used for many types of textual data (and we have since used it on many other projects). Rather than create a customized conversion tool for a single project, we have tried to create an environment for creating, testing and applying conversion scenarios. We are currently planning to add scripting support to the application. For the Colonial Despatches project, we resorted to customizing the application by adding specific application code to accomplish the conversion. A scripting

feature in Tranformer would have enabled us to write that code in script form, and combine it with conventional search-and-replace operations in a single process, without modifying the application itself.

Although this paper does not focus on digital preservation, it is worth noting that once data has been rescued, every effort should be made to encode and store it in such a way that it does not require rescuing again in the future. Good practices include use of standard file formats, accompanying documentation, regular migration as described in Besser (1999), and secure, redundant storage. To these we would add a recommendation to print out all your data; this may seem excessive, but if all else fails, the printouts will be essential, and they last a long time. Neither of the rescue projects described above would have been practical without access to the printed data. Dynamic rendering systems (such as Web sites that produce PDFs or HTML pages on demand, from database back-ends) should be able to output all the data in the form of static files which can be saved. The dynamic nature of such repositories is a great boon during development, and especially if they continue to grow and to be edited, but one day there may be no PHP or XSLT processor that can generate the output, and someone may be very glad to have those static files. We would also recommend creating virtual machines for such complex systems; if your project depends on Tomcat, Cocoon and eXist, it will be difficult to run when there are no contemporary Java Virtual Machines.

References

Besser, Howard. 1999. "Digital longevity." In Maxine Sitts (ed.) *Handbook for Digital Projects: A Management Tool for Preservation and Access*, Andover MA: Northeast Document Conservation Center, 2000, pages 155-166. Accessed at <http://www.gseis.ucla.edu/~howard/Papers/sfs-longevity.html>, 2007-11-09.

Holmes, Martin. 2007. Transformer. <http://www.tapor.uvic.ca/~mholmes/transformer/>

Hsu, Bob. n.d. Lexware.

Kirschenbaum, Matthew. 2007. "Hamlet.doc? Literature in a Digital Age." *The Chronicle of Higher Education Review*, August 17, 2007. Accessed at <http://chronicle.com/free/v53/i50/50b00801.htm>, 2007-11-09.

Krolicki, Kevin. 2001. "NASA Data Point to Mars 'Bugs,' Scientist Says." *Yahoo News* (from Reuters). Accessed at <http://web.archive.org/web/20010730040905/http://dailynews.yahoo.com/h/nm/20010727/ sc/space_mars_life_dc_1.html>, 2007-11-09.

Hendrickson, James E. (editor). 1988. *Colonial Despatches of British Columbia and Vancouver Island.* University of Victoria.

libwpd - a library for importing WordPerfect (tm) documents. <http://libwpd.sourceforge.net/>

The Nxaʔamxcín (Moses) Dictionary Database. <http://lettuce.tapor.uvic.ca/cocoon/projects/moses/>

Vogt-O'Connor, Diane. 1999. "Is the Record of the 20th Century at Risk?" *CRM: Cultural Resource Management* 22, 2: 21–24.

Digital Editions for Corpus Linguistics: A new approach to creating editions of historical manuscripts

Alpo Honkapohja
ahonkapo@welho.com
University of Helsinki, Finland

Samuli Kaislaniemi
samuli.kaislaniemi@helsinki.fi
University of Helsinki, Finland

Ville Marttila
Ville.Marttila@helsinki.fi
University of Helsinki, Finland

Introduction

One relatively unexplored area in the expanding field of digital humanities is the interface between textual and contextual studies, namely linguistics and history. Up to now, few digital editions of historical texts have been designed with linguistic study in mind. Equally few linguistic corpora have been designed with the needs of historians in mind. This paper introduces a new interdisciplinary project, Digital Editions for Corpus Linguistics (DECL). The aim of the project is to create a model for new, linguistically oriented online digital editions of historical manuscripts.

Digital editions, while on a theoretical level clearly more useful to scholars than print editions, have not yet become the norm for publishing historical texts. To some extent this can be argued to result from the lack of a proper publishing infrastructure and user-friendly tools (Robinson 2005), which limit the possibilities of individual scholars and small-scale projects to produce digital editions.

The practical approach taken by the DECL project is to develop a detailed yet flexible framework for producing electronic editions of historical manuscript texts. Initially, the development of this framework will be based on and take place concurrently with the work on three digital editions of Late Middle and Early Modern English manuscript material. Each of these editions—a Late Medieval bilingual medical handbook, a family of 15th-century culinary recipe collections, and a collection of early 17th-century intelligence letters—will serve both as a template for the encoding guidelines for that particular text type and as a development platform for the common toolset. Together, the toolset and the encoding guidelines are designed to enable editors of manuscript material to create digital editions of their material with reasonable ease.

Theoretical background

The theoretical basis of the project is dually grounded in the fields of manuscript studies and corpus linguistics. The aim of the project is to create a solid pipeline from the former to the latter: to facilitate the representation of manuscript reality in a form that is amenable to corpus linguistic study. Since context and different kinds of metatextual features are important sources of information for such fields as historical pragmatics, sociolinguistics and discourse analysis, the focus must be equally on *document, text* and *context*. By document we refer to the actual manuscript, by text to the linguistic contents of the document, and by context to both the historical and linguistic circumstances relating to text and the document. In practice this division of focus means that all of these levels are considered equally important facets of the manuscript reality and are thus to be represented in the edition.

On the level of the text, DECL adopts the opinion of Lass (2004), that a digital edition should preserve the text as accurately and faithfully as possible, convey it in as flexible a form as possible, and ensure that any editorial intervention remains visible and reversible. We also adopt a similar approach with respect to the document and its context: the editorial framework should enable and facilitate the accurate encoding and presentation of both the diplomatic and bibliographical features of the document, and the cultural, situational and textual contexts of both the document and the text. In keeping with the aforementioned aims, the development of both the editions, and the tools and guidelines for producing them, will be guided by the following three principles.

Flexibility

The editions seek to offer a flexible user interface that will be easy to use and enable working with various levels of the texts, as well as selecting the features of the text, document and context that are to be included in the presentation or analysis of the text. All editions produced within the framework will build on similar logic and general principles, which will be flexible enough to accommodate the specific needs of any text type.

Transparency

The user interfaces of the editions will include all the features that have become expected in digital editions. But in addition to the edited text and facsimile images of the manuscripts, the user will also be able to access the raw transcripts and all layers of annotation. This makes all editorial intervention transparent and reversible, and enables the user to evaluate any editorial decisions.

Expandability

The editions will be built with future expansion and updating in mind. This expandability will be three-dimensional in the sense that new editions can be added and linked to existing ones, and both new documents and new layers of annotation or

information can be added to existing editions. Furthermore, the editions will not be hardwired to a particular software solution, and their texts can be freely downloaded and processed for analysis with external software tools. The editions will be maintained on a web server and will be compatible with all standards-compliant web browsers.

Technical methods

Following the aforementioned principles, the electronic editions produced by the project will reproduce the features of the manuscript text as a faithful diplomatic transcription, into which linguistic, palaeographic and codicological features will be encoded, together with associated explanatory notes elucidating the contents and various contextual aspects of the text. The encoding standard used for the editions will be based on and compliant with the latest incarnation of the TEI XML standard (P5, published 1.11.2007), with any text-type specific features incorporated as additional modules to the TEI schema. The XML-based encoding will enable the editions to be used with any XML-aware tools and easily converted to other document or database standards. In addition to the annotation describing the properties of the document, text and context, further layers of annotation—e.g. linguistic analysis—can be added to the text later on utilising the provisions made in the TEI P5 standard for standoff XML markup.

The editorial strategies and annotation practices of the three initial editions will be carefully coordinated and documented to produce detailed guidelines, enabling the production of further compatible electronic editions. The tools developed concurrently with and tested on the editions themselves will make use of existing open source models and software projects—such as GATE or Heart of Gold, teiPublisher and Xaira—to make up a sophisticated yet customisable annotation and delivery system. The TEI-based encoding standard will also be compatible with the ISO/TC 37/SC 4 standard, facilitating the linguistic annotation of the text.

Expected results

One premise of this project is that creating digital editions based on diplomatic principles will help raise the usefulness of digitised historical texts by broadening their scope and therefore also interest in them. Faithful reproduction of the source text is a requisite for historical corpus linguistics, but editions based on diplomatic transcripts of manuscript sources are equally amenable to historical or literary enquiry. Combining the approaches of different disciplines—historical linguistics, corpus linguistics, history—to creating electronic text databases should lead to better tools for all disciplines involved and increase interdisciplinary communication and cooperation. If they prove to be successful, the tools and guidelines developed by DECL will also be readily applicable to the editing and publication of other types of material, providing a model for integrating the requirements and desires of different disciplines into a single solution.

The first DECL editions are being compiled at the Research Unit for Variation, Contacts and Change in English (VARIENG) at the University of Helsinki, and will form the bases for three doctoral dissertations. These editions, along with a working toolset and guidelines, are scheduled to be available within the next five years.

Since the aim of the DECL project is to produce an open access model for multipurpose and multidisciplinary digital editions, both the editions created by the DECL project and the tools and guidelines used in their production will be published online under an open access license. While the project strongly advocates open access publication of scholarly work, it also acknowledges that this may not be possible due to ongoing issues with copyright, for example in the case of facsimile images.

The DECL project is also intended to be open in the sense that participation or collaboration by scholars or projects working on historical manuscript materials is gladly welcomed.

References

DECL (Digital Editions for Corpus Linguistics). <http://www.helsinki.fi/varieng/domains/DECL.html>.

GATE (A General Architecture for Text Engineering). <http://gate.ac.uk>. Accessed 15 November 2007.

Heart of Gold. <http://www.delph-in.net/heartofgold/>. Accessed 23 November 2007.

ISO/TC 37/SC 4 (Language resource management). <http://www.iso.org/iso/iso_technical_committee.html?commid=297592>. Accessed 15 November 2007.

Lass, Roger. 2004. "Ut custodiant litteras: Editions, Corpora and Witnesshood". *Methods and Data in English Historical Dialectology*, ed. Marina Dossena and Roger Lass. Bern: Peter Lang, 21–48. [Linguistic Insights 16].

Robinson, Peter. 2005. "Current issues in making digital editions of medieval texts—or, do electronic scholarly editions have a future?". *Digital Medievalist* 1:1 (Spring 2005). <http://www.digitalmedievalist.org>. Accessed 6 September 2006.

TEI (Text Encoding Initiative). <http:/www.tei-c.org>. Accessed 15 November 2007.

teiPublisher. <http://teipublisher.sourceforge.net/docs/index.php>. Accessed 15 November 2007.

VARIENG (Research Unit for Variation, Contacts and Change in English). <http://www.helsinki.fi/varieng/>.

Xaira (XML Aware Indexing and Retrieval Architecture). <http://www.oucs.ox.ac.uk/rts/xaira/>. Accessed 15 November 2007.

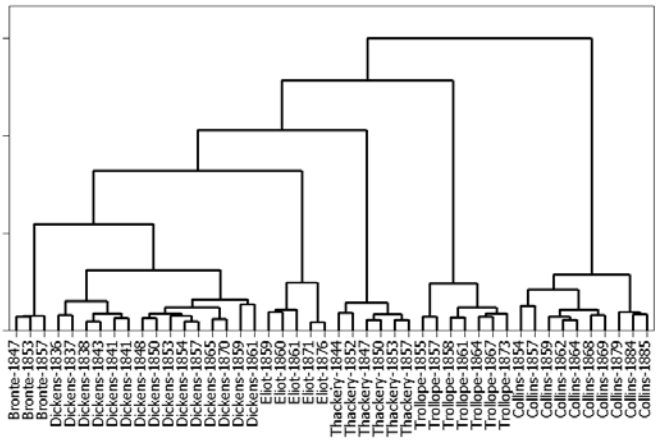
The Moonstone and The Coquette: Narrative and Epistolary Styles

David L. Hoover
david.hoover@nyu.edu
New York University, USA

In his seminal work on Austen, John F. Burrows demonstrates that characters can be distinguished from each another on the basis of the frequencies of the most frequent words of their dialogue treating the characters as if they were authors (1987). Computational stylistics has also been used to study the distinctive interior monologues of Joyce's characters in *Ulysses* (McKenna and Antonia 1996), the styles of Charles Brockden Brown's narrators (Stewart 2003), style variation within a novel (Hoover 2003), and the interaction of character separation and translation (Rybicki 2006). Here I address two kinds of local style variation: the multiple narrators of Wilkie Collins's *The Moonstone* (1868) and the multiple letter writers in Hannah Webster Foster's sentimental American epistolary novel *The Coquette* (1797).

The Moonstone has several narrators whose styles seem intuitively distinct, though all the narrations share plot elements, characters, and physical and cultural settings. *The Coquette*, based on an infamous true story, and very popular when it was published, is read today more for its cultural significance and its proto-feminist tendencies than for its literary merit. Nevertheless, one would expect the coquette, the evil seducer, the virtuous friend, and the disappointed suitor to write distinctively. Treating the narrators and letter writers of these two novels as different authors will test how successfully Collins and Foster distinguish their voices and shed light on some practical and theoretical issues of authorship and style.

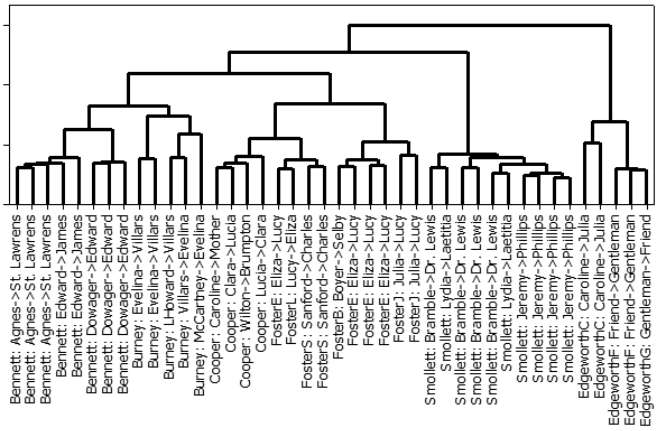
Fig. 1--Six Victorians



400MFW, no pron., culled 80%, Ward Linkage, Sq. Euclid. Dist.

Computational stylistics cannot be applied to narrators and letter writers of these novels, however, unless they can distinguish Collins and Foster from their contemporaries. Experiments on a 10-million word corpus of forty-six Victorian novels confirm that Collins is easily distinguished from five of his contemporaries, as shown in Fig. 1. For *The Coquette*, I have made the task more difficult by comparing Foster's letter writers to those of five other late 18th-century epistolary novels by five authors. I separated all the letters by writer and addressee and retaining only the 22 writers with the most letters, then combined the letters between each single writer and addressee and cut the combined texts into 42 sections of about 3,500 words. Both Delta and cluster analysis do a good job on this difficult task, and many Delta analyses give completely correct results. Cluster analysis is slightly less accurate, but several analyses are correct for five of the six authors; they also show that individual letter writers strongly tend to group together within the cluster for each novel (see Fig. 2).

Fig. 2-- *The Coquette* and Five Others



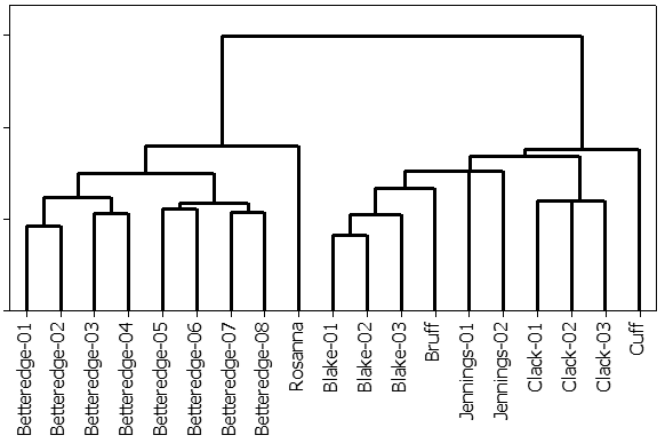
550MFW, No Pron., culled 70%, Ward Linkage, Sq. Euclid. Dist.

Examining the individual novels reveals a sharp contrast: Collins's narrators are internally consistent and easy to distinguish, while Foster's letter writers are much less internally consistent and much more difficult to distinguish. For Collins, cluster analysis consistently groups all narrative sections of 6 of the 7 narrators. When a modified technique developed especially for investigating intra-textual style variation is used (Hoover 2003), the results are even better. As Fig. 3 shows, all sections by all narrators sometimes cluster correctly (the sections range from 4,300 to 6,900 words).

"Tuning" the analysis to produce better clustering may seem circular in an analysis that tests whether Collins's narrators have consistent idiolects. But this objection can be answered by noting the consistency of the groupings. The stable clustering across large ranges of analyses with different numbers of MFW that is found here is obviously more significant than frequently-changing clustering.

Note also that the sections strongly tend to occur in narrative order in Fig. 3: every two-section cluster consists of contiguous sections. This "echo" of narrative structure provides further evidence that the analysis is accurately characterizing the narrators' styles.

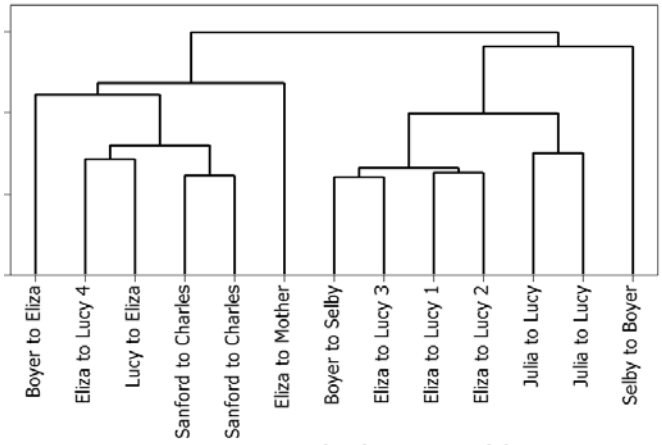
Fig. 3--Seven Narrators in *The Moonstone*



800MFW, Style Variation, Ward Linkage, Sq. Euclid. Dist.

A fine-grained investigation of *The Moonstone* involving smaller sections and including more narrators puts Collins's ability to differentiate his narrators to a sterner test, but some cluster analyses are again completely correct. The distinctiveness of the narrators of *The Moonstone* is thus confirmed by computational stylistics: the narrators behave very much as if they were literally different authors. Given the length of the novel, the generic constraints of narrative, and the inherent similarities of plot and setting, this is a remarkable achievement.

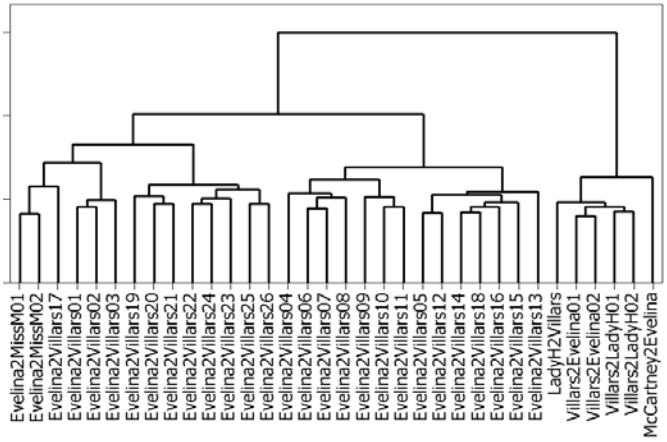
Fig. 4--13 Sections of *The Coquette*



500MFW No Pron. Ward Linkage, Sq. Euclid. Dist.

For my analysis of *The Coquette*, I have added 3 more sections of letters, so that there are 13 sections of approximately 3,000 words by 6 writers, with separate sections by a single writer to two different addressees. Although these sections should be considerably easier to group than those in the fine-grained analysis of *The Moonstone*, the results are not encouraging. Cluster analyses based on the 300-700 MFW produce very similar but not very accurate results, in all of which letters by Boyer and Eliza appear in both main clusters (see Fig. 4). Lucy's letters to Eliza also cluster with the last section of Eliza's letters to Lucy in all of these analyses a misidentification as strong as any correct one. In a real authorship attribution problem, such results would not support the conclusion that all of Boyer's or Eliza's letters were written by a single author. Perhaps one could argue that Boyer's letters to Selby should be different from his letters to Eliza, and perhaps it is appropriate that Eliza's last section includes only despairing letters written after Boyer has rejected her. Yet such special pleading is almost always possible after the fact. Further, any suggestion that Boyer's letters to Eliza should be distinct from those to Selby is impossible to reconcile with the fact that they cluster so consistently with Eliza's early letters to Lucy. And if Eliza's early and late letters should be distinct, it is difficult understand the clustering of the early letters with those of Boyer to Selby and the consistent clustering of the late letters with Lucy's letters to Eliza. It is difficult to avoid the conclusion that Foster has simply failed to create distinct and consistent characters in *The Coquette*.

Fig. 5--34 Sections of Evelina



750MFW No Pron. Ward Linkage, Sq. Euclid. Dist.

In contrast, Fanny Burney, whose *Evelina* is included in the novels compared with *The Coquette*, above, creates very distinct voices for the letter writers. Although only four of the writers have parts large enough for reasonable analysis, Evelina writes to both her adoptive father Mr. Villars and her friend Miss Mirvan, and Villars writes both to Evelina and to Lady Howard. One might expect significant differences between letters to these very different addressees. Evelina's style might also be expected to change over the course of this bildungsroman. However, analyses of all 34 sections of letters from *Evelina* (approximately 2,500 words long), show that Burney's characters are much more distinct and consistent than Foster's, as a representative analysis shows (see Fig. 5). This dendogram also strongly reflects the narrative structure of the novel. Burney, like Collins, is very successful in creating distinctive voices for her characters.

Criticism of Foster's novel has paid little attention to the different voices of the characters, but what commentary there is does not suggest that the patterns shown above should have been predictable. Smith-Rosenberg, for example, suggests a contrast between Eliza and "the feminized Greek chorus of Richman, Freeman, and Eliza's widowed mother, who, at the end, can only mouth hollow platitudes" (2003: 35). Although these women and Julia are often considered a monolithic group urging conventional morality, the distinctness of Julia's

sections, especially from Lucy's (see Fig. 5) might suggest a reexamination of this notion, and might reveal how the styles of Julia and the other women are related to more significant differences of opinion, character, or principle. Various suggestions about changes in Eliza over the course of the novel might also benefit from a closer investigation of the language of the letters. Because Foster's novel is of interest chiefly on cultural, historical, and political grounds rather than literary ones, however, such an investigation is more likely to advance the theory and practice of computational stylistics than the criticism of *The Coquette*. It is clear, at any rate, that computational stylistics is adequate to the task of distinguishing narrators and letter writers, so long as the author is adequate to the same task.

References

Burrows, J. (1987) *Computation into Criticism*. Oxford: Clarendon Press.

Hoover, D. (2003) Multivariate Analysis and the Study of Style Variation', *LLC* 18: 341-60.

Smith-Rosenberg, C. (2003) Domesticating Virtue: Coquettes and Revolutionaries in Young America', In M. Elliott and C. Stokes (eds),*American Literary Studies:A Methodological Reader,* New York: NewYork Univ. Press.

Rybicki, J. (2006). "Burrowing into Translation: Character Idiolects in Henryk Sienkiewicz's Trilogy and its Two English Translations," *LLC* 21:91-103.

Stewart, L."Charles Brockden Brown: Quantitative Analysis and Literary Interpretation," *LLC* 2003 18: 129-38.

Term Discovery in an Early Modern Latin Scientific Corpus

Malcolm D. Hyman
hyman@mpiwg-berlin.mpg.de
Max Planck Institut für Wissenschaftsgeschichte, Germany

This paper presents the results of a pilot project aimed at the development of automatic techniques for the discovery of salient technical terminology in a corpus of Latin texts. These texts belong to the domain of mechanics and date from the last quarter of the sixteenth and the first quarter of the seventeenth century, a period of intense intellectual activity in which engineers and scientists explored the limits of the Aristotelian and Archimedean paradigms in mechanics. The tensions that arose ultimately were resolved by the new "classical mechanics" inaugurated by Newton's *Principia* in 1687 (cf. Damerow et al. 2004).

The work presented here forms part of a larger research project aimed at developing new computational techniques to assist historians in studying fine-grained developments in long-term intellectual traditions, such as the tradition of Western mechanics that begins with the pseudo-Aristotelian *Problemata Mechanica* (ca. 330 B.C.E.). This research is integrated with two larger institutional projects: the working group "Mental Models in the History of Mechanics" at the Max Planck Institute for the History of Science in Berlin, and the German DFG-funded Collaborative Research Center (CRC) 644 "Transformations of Antiquity."

The purpose of this paper is to present initial results regarding the development of efficient methods for technical term discovery in Early Modern scientific Latin. The focus is on the identification of term variants and on term enrichment. The methodology employed is inspired by Jacquemin (2001), whose approach allows for the use of natural language processing techniques without the need for full syntactic parsing, which is currently not technically feasible for Latin.

The present paper extends prior work in term discovery along two vectors. First, work in term discovery has primarily addressed the languages of Western Europe (especially English and French), with some work also in Chinese and Japanese. Latin presents some typological features that require modifications to established techniques. Chief among these is the rich inflectional morphology (both nominal and verbal) of Latin, which is a language of the synthetic type. Latin also exhibits non-projectivity, i.e. syntactic constituents may be represented non-continuously (with the intrusion of elements from foreign constituents). Although the non-projectivity of Renaissance Latin is considerably less than what is found in the artistic prose (and *a fortiori* poetry) of the Classical language (Bamman and Crane 2006), term detection must proceed within a framework that allows for both non-projectivity and (relatively) free word order within constituents.

Second, researchers in the field of term discovery have focused almost exclusively on contemporary scientific corpora in domains such as biomedicine. In contemporary scientific literature, technical terms are characterized by a particularly high degree of denotative monosemicity, exhibit considerable stability, and follow quite rigid morphological, syntactic, and semantic templates. Although these characteristics are also applicable to the terminology of Latin scientific texts, they are applicable to a lesser degree. In other words, the distinction between technical terminology and ordinary language vocabulary is less clear cut than in the case of contemporary scientific and technical language. The lesser degree of monosemicity, stability, and structural rigidity of terminology holds implications for automatic term discovery in corpora earlier than the twentieth (or at least nineteenth) century.

The corpus of Early Modern mechanics texts in Latin is well-designed for carrying out experiments in adapting established techniques of term discovery to historical corpora. Mechanics is by this time a scientific discipline that possesses an extensive repertoire of characteristic concepts and terminology. Thus it is broadly comparable to contemporary scientific corpora, while still presenting unique features that merit special investigation. Several thousand pages of text are available in XML format, which have been digitized by the Archimedes Project, an international German/American digital library venture jointly funded by the DFG and NSF. It will be possible to extend future work to a multilingual context, by examining in addition closely-related vernacular works (in Italian, Spanish, and German) that are contemporary with the Latin corpus. (Some of these are translations and commentaries.)

The set of technical terminology discovered by the methods presented in this paper is intended to further the computationally-assisted framework for exploring conceptual change and knowledge transfer in the history of science that has been described by Hyman (2007). This framework employs latent semantic analysis (LSA) and techniques for the visualization of semantic networks, allowing change in the semantic associations of terms to be studied within a historical corpus. The concluding section of the present paper will survey the applications of technical term discovery within historical corpora for the study of the conflict, competition, evolution, and replacement of concepts within a scientific discipline and will suggest potential applications for other scholars who are concerned with related problems.

References

Bamman, D. and G. Crane. 2006. The design and use of a Latin dependency treebank. In *Proceedings of the TLT 2006*, edd. J. Hajič and J. Nivre, Prague, pp. 67–78.

Damerow, P., G. Freudenthal, P. McLaughlin, J. Renn, eds. 2004. *Exploring the Limits of Preclassical Mechanics: A Study of Conceptual Development in Early Modern Science: Free Fall and Compounded Motion in the Work of Descartes, Galileo, and Beeckman*. 2d ed. New York.

Hyman, M.D. 2007. Semantic networks: a tool for investigating conceptual change and knowledge transfer in the history of science. In *Übersetzung und Transformation*, edd. H. Böhme, C. Rapp, and W. Rösler, Berlin, pp. 355–367.

Jacquemin, C. 2001. *Spotting and Discovering Terms through Natural Language Processing*. Cambridge, MA.

Markup in Textgrid

Fotis Jannidis
jannidis@linglit.tu-darmstadt.de
Technische Universität Darmstadt, Germany

Thorsten Vitt
vitt@linglit.tu-darmstadt.de
Technische Universität Darmstadt, Germany

The paper will discuss the decisions in relation to markup which have been made in Textgrid. The first part of the paper will describe the functionality and principal architecture of Textgrid, the second part will discuss Textgrid's baseline encoding. Textgrid is a modular platform for collaborative textual editing and a first building block for a community grid for the humanities. Textgrid consists of a toolkit for creating and working with digital editions and a repository offering storage, archiving and retrieval.

Textgrid's architecture follows a layered design, built for openness on all levels. At its base there is a middleware layer providing generic utilities to encapsulate and provide access to the data grid's storage facilities as well as external archives. Additionally, indexing and retrieval facilities and generic services like authorisation and authentication are provided here.

A service layer built on the middleware provides automated text processing facilities and access to semantic resources. Here, Textgrid offers domain-specific services like a configurable streaming editor or a lemmatizer which uses the dictionaries stored in Textgrid. All services can be orchestrated in workflows, which may also include external services.

Every service deploys standard web service technologies. As well, tools in the service layer can work with both data managed by the middleware and data streamed in and out of these services by the caller, so they can be integrated with environments outside of Textgrid.

The full tool suite of Textgrid is accessible via TextGridLab, a user interface based on Eclipse which, besides user interfaces to the services and search and management facilities for Textgrid's content, also includes some primarily interactive tools. The user interface provides integrated access to the various tools: For example, an XML Editor, a tool to mark up parts of an image and link it to the text, and a dictionary service. From the perspective of the user, all these tools are part of one application.

This software framework is completely based on plug-ins and thus reflects the other layers' extensibility: it can be easily extended by plug-ins provided by third parties, and although there is a standalone executable tailored for the philologist users, TextGridLab's plugins can be integrated with existing Eclipse installations, as well.

Additionally, the general public may read and search publicized material by means of a web interface, without installing any specialized software.

Designing this infrastructure it would have been a possibility to define one data format which can be used in all services including search and retrieval and publishing. Instead the designers chose a different approach: each service or software component defines its own minimal level of format restriction. The XML editor, which is part of the front end, is designed to process all files which are xml conform; the streaming editor service can handle any kind of file etc. The main reason for this decision was the experience of those people involved and the model of the TEI guidelines to allow users as much individual freedom to choose and use their markup as possible even if the success of TEI lite and the many project specific TEI subsets seem to point to the need for defining strict standards.

But at some points of the project more restrictive format decisions had to be made. One of them was the result of the project's ambition to make all texts searchable in a way which is more useful than a mere full text search. On the other hand it isn't realistic to propose a full format which will allow all future editors, lexicographers and corpus designers to encode all features they are interested in. So Textgrid allows all projects to use whatever XML markup seems necessary but burdens the project with designing its own interface to these complex data structures. But in this form the project data are an island and there is no common retrieval possible. To allow a retrieval across all data in Textgrid which goes beyond the possibilities of a full text research, the Textgrid designers discussed several possibilities but finally settled down on a concept which relies very much on text types like drama, prose, verse, letter etc. and we differentiate between basic text types like verse and container text types like corpora or critical editions.

Interproject search is enabled by transforming all texts into a rudimentary format which contains the most important information of the specific text type. This baseline encoding is not meant to be a core encoding which covers all important information of a text type but it is strictly functional. We defined three demands which should be met by the baseline encoding, which is meant to be a subset of the TEI:

1) Intelligent search. Including often used aspects of text types into the search we try to make text retrieval more effective. A typical example would be the 'knowledge' that a word is the lemma of a dictionary entry, so a search for this word would mark this subtree as a better hit than another where it is just part of a paragraph.

2) Representation of search results. The results of an inter-project search have to be displayed in some manner which preserves some important aspects of the source texts.

3) Automatic reuse and further processing of text. A typical example for this would be the integration of a dictionary in a network of dictionaries. This aspect is notoriously

underrepresented in most design decisions of modern online editions which usually see the publication as the natural goal of their project, a publication which usually only allows for reading and searching as the typical forms of text usage.

Our paper will describe the baseline encoding format for some of the text types supported by Textgrid at the moment including the metadata format and discuss in what ways the three requirements are met by them.

One of the aims of our paper is to put our arguments and design decisions up for discussion in order to test their validity. Another aim is to reflect on the consequences of this approach for others like the TEI, especially the idea to define important text types for the humanities and provide specific markup for them.

Breaking down barriers: the integration of research data, notes and referencing in a Web 2.0 academic framework

Ian R. Johnson
johnson@acl.arts.usyd.edu.au
University of Sydney, Australia

In this paper I argue that the arbitrary distinction between bibliographic data and research data – which we see in the existence of specialised library catalogues and bibliographic systems on the one hand, and a multitude of *ad hoc* notes, digitized sources, research databases and repositories on the other – is a hangover from a simpler past, in which publication and bibliographic referencing was a well-defined and separate part of the research cycle.

Today published material takes many different forms, from books to multimedia, digital artworks and performances. Research data results from collection and encoding of information in museums, archives, libraries and fieldwork, as well as output from analysis and interpretation. As new forms of digital publication appear, the boundary between published material and research data blurs. Given the right enabling structures (eg. peer review) and tools (eg. collaborative editing), a simple digitized dataset can become as valuable as any formal publication through the accretion of scholarship. When someone publishes their academic musings in a personal research blog, is it analogous to written notes on their desk (or desktop) or is it grey literature or vanity publishing?

The drawing of a distinction between bibliographic references and other forms of research data, and the storing of these data in distinct systems, hinders construction of the linkages between information which lie at the core of Humanities research. Why on earth would we want to keep our bibliographic references separate from our notes and developing ideas, or from data we might collect from published or unpublished sources? Yet standalone desktop silos (such as EndNote for references, Word for notes and MSAccess for data) actively discourage the linking of these forms of information.

Bespoke or ad hoc databases admirably (or less than admirably) fulfill the particular needs of researchers, but fail to connect with the wider world. These databases are often desktop-based and inaccessible to anyone but the user of their host computer, other than through sharing of copies (with all the attendant problems of redundancy, maintenance of currency and merging of changes). When accessible they often lack multi-user capabilities and/or are locked down to modification by a small group of users because of the difficulties of monitoring and rolling back erroneous or hostile changes. Even when

accessible to the public, they are generally accessible through a web interface which allows human access but not machine access, and cannot therefore be linked programmatically with other data to create an integrated system for analyzing larger problems.

For eResearch in the Humanities to advance, all the digital information we use – bibliographic references, personal notes, digitized sources, databases of research objects etc. – need to exist in a single, integrated environment rather than in separate incompatible systems. This does not of course mean that the system need be monolithic – mashups, portals and Virtual Research Environments all offer distributed alternatives, dependant on exposure of resources through feed and web services. The 'silo' approach to data is also breaking down with the stunning success of web-based social software such as the Wikipedia encyclopaedia or Del.icio.us social bookmarking systems. These systems demonstrate that – with the right level of control and peer review – it is possible to build substantial and highly usable databases without the costs normally associated with such resources, by harnessing the collaborative enthusiasm of large numbers of people for data collection and through data mining of collective behaviour.

To illustrate the potential of an integrated Web 2.0 approach to heterogeneous information, I will discuss Heurist (HeuristScholar.org) – an academic social bookmarking application which we have developed, which provides rich information handling in a single integrated web application – and demonstrate the way in which it has provided a new approach to building significant repositories of historical data.

Heurist handles more than 60 types of digital entity (easily extensible), ranging from bibliographic references and internet bookmarks, through encyclopaedia entries, seminars and grant programs, to C14 dates, archaeological sites and spatial databases. It allows users to attach multimedia resources and annotations to each entity in the database, using private, public, and group-restricted wiki entries. Some entries can be locked off as authoritative content, others can be left open to all comers.

Effective geographic and temporal contextualisation and linking between entities provides new opportunities for Humanities research, particularly in History and Archaeology. Heurist allows the user to digitize and attach geographic data to any entity type, to attach photographs and other media to entities, and to store annotated, date-stamped relationships between entities. These are the key to linking bibliographic entries to other types of entity and building, browsing and visualizing networks of related entities.

Heurist represents a first step towards building a single point of entry Virtual Research Environment for the Humanities. It already provides 'instant' web services, such as mapping, timelines, styled output through XSLT and various XML feeds (XML, KML, RSS) allowing it to serve as one component in a decentralized system. The next version will operate in a peer-to-peer network of instances which can share data with one another and with other applications.

The service at HeuristScholar.org is freely available for academic use and has been used to construct projects as varied as the University of Sydney Archaeology department website, content management for the Dictionary of Sydney project (a major project to develop an online historical account of the history of Sydney) and an historical event browser for the Rethinking Timelines project.

Constructing Social Networks in Modern Ireland (C.1750-c.1940) Using ACQ

Jennifer Kelly
jennifer.kelly@nuim.ie
National University of Ireland, Maynooth, Ireland

John G. Keating
john.keating@nuim.ie
National University of Ireland, Maynooth, Ireland

Introduction

The Associational Culture in Ireland (ACI) project at NUI Maynooth explores the culture of Irish associational life from 1750 to 1940, not merely from the point of view of who, what, where and when, but also to examine the 'hidden culture' of social networking that operated behind many clubs and societies throughout the period. Recently commissioned government research on civic engagement and active citizenship in Ireland has highlighted the paucity of data available for establishing 'trends in volunteering, civic participation, voting and social contact in Ireland' (Taskforce on Active Citizenship, Background Working Paper, 2007, p. 2). The same research has also confirmed the importance in Ireland of informal social networking compared to many other economically developed countries (Report of the Taskforce on Active Citizenship, 2007). The objective of the ACI project is to provide a resource to enable scholars of Irish social and political life to reconstruct and highlight the role that the wider informal community information field played in the public sphere in Ireland from the mid-eighteenth century. The project will also provide long-term quantitative digital data on associational culture in Ireland which is compatible with sophisticated statistical analysis, thereby enabling researchers to overcome one of the hindrances of modern-day purpose based social surveys: the short timeframe of data currently available.

Associational Culture and Social Networking

All historians are aware of the importance of social networks that underpin the foundations of social and political life in the modern world (Clark, 2000; Putnam, 2000). However, given the often-transient nature of much of these networks, they can be quite difficult to reconstruct.

One way to examine social networks is to trace them through the structures of a developing associational culture where the cultivation of social exclusivity and overlapping membership patterns provide insights into the wider organisation of civil society at local, regional and national levels. To this end, the ACI project mines a wide range of historical sources to piece together as comprehensive a view as possible of the various voluntary formal associations that existed in Ireland during the period c.1750-c.1940.

The first phase of the project concentrated on collecting data on Irish associational culture and social networking in the period 1750-1820; the current phase centres on the period 1820-1880 and the third phase will focus on the years between 1880 and 1940. The research results so far testify to the vibrancy of associational activity in Ireland and already patterns in social networking are becoming apparent in different parts of the country. The results so far indicate that particular forms of associational culture were popular in different parts of the country from the mid-eighteenth century, which in turn produced their own particular social networking systems. In many respects these patterns were maintained into the nineteenth century with similar continuities in patterns of sociability even though these continuities were sometimes expressed through different organisations, e.g. the ubiquitous Volunteering movement of the later eighteenth century gave way to local Yeomanry units in the beginning of the nineteenth century. Associational culture among the urban middling strata also appeared to increase somewhat moving into the nineteenth century with the increasing appearance of charitable and temperance societies in different parts of the country.

Software Development

Given the vast range of data available in the sources and the multiplicity of question types posed by the findings, the organisation and dissemination of the research was one of the main priorities for the project. The desire to present a quantitative as well as qualitative profile of Irish associational culture, which can be expanded upon by future research, presented particular difficulties in terms of the production of the project's findings. A multi-skilled, partnership approach was adopted, fusing historical and computer science expertise to digitally represent the data recovered from the sources and expose the resultant social networking patterns through the construction of an appropriate online database for the project. Encompassing over seventy individual data fields for each association, the online ACI database is fully searchable by organisation and by individual – the later feature in particular allowing social patterns behind Irish associational culture to be traced over the course of two centuries.

Bradley (2005) points out XML is well suited to document oriented project materials (such as written text), whereas relational database implementations are better suited to data oriented project materials. Furthermore, he indicates that projects may often use both, given that textual materials often contain data materials that are more suited to relational models. He argues that even if one begins with a text, the materials may be data-orientated and better served by a relational model where certain characteristics, for example, linkages between data objects and powerful querying are more easily expressed using SQL (Structured Query Language) than XML

searching facilities, for example, XPATH or XQUERY. It was necessary, for our project to use a relational model to encode data orientated materials, even though data were derived from newspaper articles which are typically more suited to encoding in XML. Our relational objects are densely linked, and it is often necessary to build SQL queries incorporating multiple joins across many database tables.

The database contains over 100 individual database tables related to associations, members, sources, relationships, associate members, locations, etc. Authenticated data entry and access is available using online forms. Test, rollback and moderation facilities are available for different classes of user. Research queries may be formulated using a specifically designed English language-type formal grammar called Associational Culture Query Language (ACQL), which is parsed by the software system to extract and present information from the database – parsing involves evaluating ACQL and converting it into appropriately constructed SQL suitable for querying the database. Users may construct ACQL manually or they may use an online query builder (see Figure 1).

The software engineering process began with all partners involved in a collaborative process concentrating on the construction of a Specification of Requirements (SR) document which essentially formed the contract between the ACI historians and software engineers – this key document was used to derive all phases of the software development process, for example, System Analysis, System Design Specifications, Software Specifications, and Testing and Maintenance. These phases were implemented using a rapid prototyping model, where successive passes through a design-development-testing cycle provided new prototypes which could be evaluated against the SR document. As is typical of such development projects, there was some requirement drift, but the rapid prototyping approach ensured insignificant divergence between prototypes and expectations outlined in the SR document.

In order to produce the SR document, the ACI historians composed a series of research questions, which were used to extract the information categories necessary for the construction of social networks. During the collaborative process, these research questions and sources were used to identify and model the relationships between the associations and individuals in the sources. A key requirement was that ACI historians wanted to be able to answer specific research questions related to social networking, for example, "Are there any illegal women's associations/clubs active in Ireland after 1832?" It was necessary to revise these research questions during the Analysis phase, however, as it became apparent that the historians did not expect a yes/no answer to this question, but rather a list of the associations and clubs, if they existed. Most research questions went through a revision process where the cardinality or result context was explicit, i.e. "List the illegal women's associations active in Ireland after 1832". This research question could be reformulated in ACQL as follows:

```
LIST THE BODIES (NAME) (
WHERE THE GENDER IS "Female" AND
WHERE THE STATUS IS "Illegal" AND
WHERE THE FOUNDING DATE
IS GREATER THAN 1st OF January 1832
)
```

The parsing software then converts this ACQL query into the following SQL:

```
SELECT BodyName.bodyName
FROM  Body, BodyGender,
BodyName, BodyStatus
WHERE  (Body.idBody = BodyName.
Body_idBody) AND
  (Body.idBody = BodyGender.
Body_idBody AND
   BodyGender.gender ='Female'  AND
 (Body.idBody = BodyStatus.
Body_idBody AND
  BodyStatus.typeOfStatus ='Illegal'  AND
  (Body.foundingDate > '1832-01-1')
 )
);
```

This query is then executed by the Relational Database Management System (RDBMS) and the results are returned to the environment for presentation to the user.

We examined over one hundred questions of this type ensuring that our relational model provided appropriate answers. We developed ACQL to map research questions into SQL queries, thereby removing the requirement for the user to have knowledge of the database, table joins, or even SQL. Researchers need not have an intimate knowledge of the relationship between the database tables and their associated fields to perform searches. Another advantage of this approach is that the underlying database structure could change and the users would not have to change the format of their queries implemented in ACQL.

Ongoing and Future Developments

The success of this project depends on the research community using and contributing to the information contained in the online database. We are heartened, however, by the reports of Warwick et al (2007) who have shown that when information aggregation sites have user-friendly interfaces, contain quality peer-reviewed information, and fit research needs, the research community scholars are more likely adopt the digital resource. We believe that ACQL is a crucial component in making this system usable by professional historians interested in social networking. In particular, as the volume of data within the database expands in digital format, the potential for developing further social analysis tools such as sociograms will be initiated. All in all, the ACI database, by providing quantitative and qualitative data on specific

associations, regions, and groups of individuals, comparable with international data sources, will greatly aid historical and social science researchers to establish Irish trends in civic participation, social inclusion, marginalisation and grassroots organising in modern Ireland.

References

Bradley, J. (2005) Documents and Data: Modelling Materials for Humanities Research in XML and Relational Databases. *Literary and Linguistic Computing*, Vol. 20, No. 1.

Clark, P. (2000) *British clubs and societies, 1500-1800: the origins of an associational world.* Oxford: Oxford University Press.

Putnam, R. (2000) *Bowling alone: the collapse and revival of American community*. New York: Simon & Schuster.

Taskforce on Active Citizenship (2007) *Statistical Evidence on Active Citizenship in Ireland, Background Working Paper*. Retrieved March 25, 2008 from the World Wide Web: http://www.activecitizen.ie/UPLOADEDFILES/Mar07/Statistical%20Report%20(Mar%2007).pdf.

Taskforce on Active Citizenship (2007) *Report of the Taskforce on Active Citizenship*. Dublin: Secretariat of the Taskforce on Active Citizenship.

Warwick, C., Terras, M., Huntington, P. and Pappa, N. (2007). 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*, Vol. 22, No. 1. pp. 1-18.

Unnatural Language Processing: Neural Networks and the Linguistics of Speech

William Kretzschmar
kretzsch@uga.edu
University of Georgia, USA

The foundations of the linguistics of speech (i.e., language in use, what people actually say and write to and for each other), as distinguished from ?the linguistics of linguistic structure? that characterizes many modern academic ideas about language, are 1) the continuum of linguistic behavior, 2) extensive (really massive) variation in all features at all times, 3) importance of regional/social proximity to "shared" linguistic production, and 4) differential frequency as a key factor in linguistic production both in regional/social groups and in collocations in text corpora (all points easily and regularly established with empirical study using surveys and corpora, as shown in Kretzschmar Forthcoming a). Taken together, the basic elements of speech correspond to what has been called a ?complex system? in sciences ranging from physics to ecology to economics. Order emerges from such systems by means of self-organization, but the order that arises from speech is not the same as what linguists study under the rubric of linguistic structure. This paper will explore the relationship between the results of computational analysis of language data with neural network algorithms, traditionally accepted dialect areas and groupings, and order as it emerges from speech interactions.

In both texts and regional/social groups, the frequency distribution of features (language variants per se or in proximal combinations such as collocations, colligations) occurs as the same curve: a ?power law? or asymptotic hyperbolic curve (in my publications, aka the ?A-curve?). Speakers perceive what is "normal" or "different" for regional/social groups and for text types according to the A-curve: the most frequent variants are perceived as "normal," less frequent variants are perceived as "different," and since particular variants are more or less frequent among different groups of people or types of discourse, the variants come to mark identity of the groups or types by means of these perceptions. Particular variants also become more or less frequent in historical terms, which accounts for what we call "linguistic change," although of course any such "changes" are dependent on the populations or text types observed over time (Kretzschmar and Tamasi 2003). In both synchronic and diachronic study the notion of "scale" (how big are the groups we observe, from local to regional/social to national) is necessary to manage our observations of frequency distributions. Finally, our perceptions of the whole range of "normal" variants (at any level of scale) create "observational artifacts." That is, the notion of the existence of any language or dialect is actually an "observational artifact" that comes from our perceptions of the available variants (plus other information and attitudes), at one point in time and for a particular group of speakers, as mediated by the A-curve.

The notion "Standard," as distinct from "normal," represents institutional agreement about which variants to prefer, some less frequent than the "normal" variants for many groups of speakers, and this creates the appearance of parallel systems for "normal" and "Standard."

The best contemporary model that accommodates such processing is connectionism, parallel processing according to what anthropologists call "schemas" (i.e., George Mandler's notion of schemas as a processing mechanism, D'Andrade 1995: 122-126, 144-145). Schemas are not composed of a particular set of characteristics to be recognized (an object), but instead of an array of slots for characteristics out of which a pattern is generated, and so schemas must include a process for deciding what to construct. One description of such a process is the serial symbolic processing model (D'Andrade 1995: 136-138), in which a set of logical rules is applied in sequence to information available from the outside world in order to select a pattern. A refinement of this model is the parallel distributed processing network, also called the connectionist network, or neural net (D'Andrade 1995: 138-141), which allows parallel operation by a larger set of logical rules. The logical rules are Boolean operators, whose operations can be observed, for example, in simulations that Kauffman (1996) has built based on networks of lightbulbs. Given a very large network of neurons that either fire or not, depending upon external stimuli of different kinds, binary Boolean logic is appropriate to model "decisions" in the brain which arise from the on/off firing patterns. Kauffman's simulations were created to model chemical and biological reactions which are similarly binary, either happening or not happening given their state (or pattern) of activation, as the system cycles through its possibilities. The comparison yields similar results: as D'Andrade reports (1995: 139-140), serial processing can be '"brittle'--if the input is altered very slightly or the task is changed somewhat, the whole program is likely to crash" (or as Kauffman might say, likely to enter a chaotic state cycle), while parallel processing appears to be much more flexible given mixed or incomplete input or a disturbance to the system (or as Kauffman might say, it can achieve homeostatic order).

Computational modeling of neural networks appears, then, to be an excellent match for analysis of language data. Unfortunately, results to date have often been disappointing when applied to geographic language variation (Nerbonne and Heeringa 2001, Kretzschmar 2006). Neural network analysis cannot be shown reliably to replicate traditional dialect patterns. Instead, self-organizational patterns yielded by neural net algorithms appear to respond only in a general way to assumed dialect areas, and often appear to be derived not from the data but from conditions of its acquisition such as "field worker" effects (Kretzschmar Forthcoming b). However, this paper will show, using results from experiments with an implementation of a Self-Organizing Map (SOM) algorithm (Thill, Kretzschmar, Casas, and Yao Forthcoming), that application of the model from the linguistics of speech to computer neural network analysis of geographical language data can explain such anomalies. It is not the implementation of neural nets that is the problem, but instead lack of control over the scale of analysis, and of the non-linear distribution of the variants included in the analysis, that tends to cause the problems we observe. In the end, we still cannot validate traditional dialect areas from the data (because these areas were also derived without sufficient control over the dynamics of the speech model), but we can begin to understand more clearly how the results of neural network analysis do reveal important information about the distribution of the data submitted to them.

References

D'Andrade, Roy. 1995. *The Development of Cognitive Anthropology.* Cambridge: Cambridge University Press.

Kauffman, Stuart. 1996. *At Home in the Universe: The Search for the Laws of Self-Organization and Complexity.* New York: Oxford University Press.

Kretzschmar, William A., Jr. 2006. Art and Science in Computational Dialectology. *Literary and Linguistic Computing* 21: 399-410.

Kretzschmar, William A., Jr. Forthcoming a. *The Linguistics of Speech.* Cambridge: Cambridge University Press.

Kretzschmar, William A., Jr. Forthcoming b. The Beholder?s Eye: Using Self-Organizing Maps to Understand American Dialects. In Anne Curzan and Michael Adams, eds., *Contours of English* (Ann Arbor: University of Michigan Press).

Kretzschmar, William A., Jr., and Susan Tamasi. 2003. Distributional Foundations for a Theory of Language Change. *World Englishes* 22: 377-401.

Nerbonne, John, and Wilbert Heeringa. 2001. Computational Comparison and Classification of Dialects. *Dialectologia et Geolinguistica* 9: 69-83.

Thill, J., W. Kretzschmar, Jr, I. Casas, and X. Yao. Forthcoming. Detecting Geographic Associations in English Dialect Features in North America with Self-Organising Maps. In *Self-Organising Maps: Applications in GI Science*, edited by P. Agarwal and A. Skupin (London: Wiley).

Digital Humanities 'Readership' and the Public Knowledge Project

Caroline Leitch
cmleitch@uvic.ca
University of Victoria, Canada

Ray Siemens
siemens@uvic.ca
University of Victoria, Canada

Analisa Blake
University of Victoria, Canada

Karin Armstrong
karindar@uvic.ca
University of Victoria, Canada

John Willinsky
john.willinsky@ubc.ca
University of British Columbia, Canada,

As the amount of scholarly material published in digital form increases, there is growing pressure on content producers to identify the needs of expert readers and to create online tools that satisfy their requirements. Based on the results of a study conducted by the Public Knowledge Project and introduced at Digital Humanities 2006 (Siemens, Willinsky and Blake), continued and augmented since, this paper discusses the reactions of Humanities Computing scholars and graduate students to using a set of online reading tools.

Expert readers were asked about the value of using online tools that allowed them to check, readily, related studies that were not cited in the article; to examine work that has followed on the study reported in the article; to consider additional work that has been done by the same author; and to consult additional sources on the topic outside of the academic literature. In the course of this study, these domain-expert readers made it clear that reading tools could make a definite, if limited, contribution to their critical engagement with journal articles (especially if certain improvements were made). Their reactions also point to an existing set of sophisticated reading strategies common to most expert readers.

Our findings indicate that online tools are of most value to expert readers when they complement and augment readers' existing strategies. We have organized the results of the study around a number of themes that emerged during our interviews with domain expert readers as these themes speak to both readers' existing reading processes and the potential value of the online reading tools. By entering user responses into a matrix, we have been able to measure user responses and track both negative and positive reactions to different aspects of the online reading tools.

In addition to these findings, we also discovered that users' experiences with the online reading tools was influenced by their existing research methods, their familiarity with online research, and their expectations of online publishing. While many respondents felt that the "information environment" created by the online tools was beneficial to their evaluation and understanding of the material, they also expressed some dissatisfaction with their experience. Some users questioned the relevance and usefulness of the contextual material retrieved by the online tools. Users were also concerned with the perceived credibility of research published online and the limited amount of freely available online material.

The results of our study reveal both the potential strengths and perceived weaknesses of online reading environments. Understanding how users read and evaluate research materials, anticipating users' expectations of the reading tools and resources, and addressing user concerns about the availability of online material will lead to improvements in the design and features of online publishing.

Bibliography

Afflerbach Peter P. "The Influence of Prior Knowledge on Expert Readers' Main Idea Construction Strategies." *Reading Research Quarterly* 25.1 (1990): 31-46.

Alexander, P. A. *Expertise and Academic Development: A New Perspective on a Classic Theme.* Invited keynote address to the Biennial meeting of the European Association for the Research on Learning and Instruction (EARLI). Padova, Italy. August 2003.

Chen, S., F. Jing-Ping, and R. Macredie. "Navigation in Hypermedia Learning Systems: Experts vs. Novices." *Computers in Human Behavior* 22.2 (2006): 251-266.

Pressley, M., and P. Afflerbach. *Verbal Protocols of Reading: The Nature of Constructively Responsive Reading*. Hillsdale NJ: Erlbaum, 1995.

Schreibman, Susan, Raymond George Siemens, and John M. Unsworth, eds. *A Companion to Digital Humanities*. Oxford, UK: Blackwell, 2004

Siemens, Ray, Analisa Blake, and John Willinsky. "Giving Them a Reason to Read Online: Reading Tools for Humanities Scholars." *Digital Humanities* 2006. Paris.

Stanitsek, Geog. "Texts and Paratexts in Media." Trans. Ellen Klien. *Critical Inquiry* 32.1 (2005): 27-42.

Willinsky, J. "Policymakers' Use of Online Academic Research." *Education Policy Analysis Archives* 11.2, 2003, <http://epaa.asu.edu/epaa/v11n2/>.

Willinsky, J., and M. Quint-Rapoport. *When Complementary and Alternative Medicine Practitioners Use PubMed.* Unpublished paper. University of British Columbia. 2007.

Windschuttle, Keith. "Edward Gibbon and the Enlightenment." *The New Criterion* 15.10. June 1997. <http://www.newcriterion.com/archive/15/jun97/gibbon.htm>.

Wineburg, S. "Historical Problem Solving: A Study of the Cognitive Processes Used in the Evaluation of Documentary and Pictorial Evidence." *Journal of Educational Psychology* 83.1 (1991): 73-87.

Wineburg, S. "Reading Abraham Lincoln: An Expert/Expert Study in the Interpretation of Historical Texts." *Cognitive Science* 22.3 (1998): 319-346.

Wyatt, D., M. Pressley, P.B. El-Dinary, S. Stein, and R. Brown. "Comprehension Strategies, Worth and Credibility Monitoring, and Evaluations: Cold and Hot Cognition When Experts Read Professional Articles That Are Important to Them." *Learning and Individual Differences* 5 (1993): 49–72.

Using syntactic features to predict author personality from text

Kim Luyckx
kim.luyckx@ua.ac.be
University of Antwerp, Belgium

Walter Daelemans
walter.daelemans@ua.ac.be
University of Antwerp, Belgium

Introduction

The style in which a text is written reflects an array of meta-information concerning the text (e.g., topic, register, genre) and its author (e.g., gender, region, age, personality). The field of stylometry addresses these aspects of style. A successful methodology, borrowed from text categorisation research, takes a two-stage approach which (i) achieves automatic selection of features with high predictive value for the categories to be learned, and (ii) uses machine learning algorithms to learn to categorize new documents by using the selected features (Sebastiani, 2002). To allow the selection of linguistic features rather than (*n*-grams of) terms, robust and accurate text analysis tools are necessary. Recently, language technology has progressed to a state of the art in which the systematic study of the variation of these linguistic properties in texts by different authors, time periods, regiolects, genres, registers, or even genders has become feasible.

This paper addresses a not yet very well researched aspect of style, the author's personality. Our aim is to test whether personality traits are reflected in writing style. Descriptive statistics studies in language psychology show a direct correlation: personality is projected linguistically and can be perceived through language (e.g., Gill, 2003; Gill & Oberlander, 2002; Campbell & Pennebaker, 2003). The focus is on extraversion and neuroticism, two of "the most salient and visible personality traits" (Gill, 2003, p. 13). Research in personality prediction (e.g., Argamon et al., 2005; Nowson & Oberlander, 2007; Mairesse et al., 2007) focuses on openness, conscientiousness, extraversion, agreeableness, and neuroticism.

We want to test whether we can automatically predict personality in text by studying the four components of the Myers-Briggs Type Indicator: Introverted-Extraverted, Intuitive-Sensing, Thinking-Feeling, and Judging-Perceiving. We introduce a new corpus, the *Personae* corpus, which consists of Dutch written language, while other studies focus on English. Nevertheless, we believe our techniques to be transferable to other languages.

Related Research in Personality Prediction

Most of the research in personality prediction involves the Five-Factor Model of Personality: openness, conscientiousness, extraversion, agreeableness, and neuroticism. The so-called *Big Five* have been criticized for their limited scope, methodology and the absence of an underlying theory. Argamon et al. (2005) predict personality in student essays using functional lexical features. These features represent lexical and structural choices made in the text. Nowson & Oberlander (2007) perform feature selection and training on a small and clean weblog corpus, and test on a large, automatically selected corpus. Features include *n*-grams of words with predictive strength for the binary classification tasks. Openness is excluded from the experiments because of the skewed class distribution. While the two studies mentioned above took a bottom-up approach, Mairesse et al. (2007) approach personality prediction from a top-down perspective. On a written text corpus, they test the predictive strength of linguistic features that have been proposed in descriptive statistics studies.

Corpus Construction

Our 200,000-word *Personae* corpus consists of 145 BA student essays of about 1,400 words about a documentary on Artificial Life in order to keep genre, register, topic and age relatively constant. These essays contain a factual description of the documentary and the students' opinion about it. The task was voluntary and students producing an essay were rewarded with two cinema tickets. They took an online MBTI test and submitted their profile, the text and some user information. All students released the copyright of their text to the University of Antwerp and explicitly allowed the use of their text and personality profile for research, which makes it possible to distribute the corpus.

The Myers-Briggs Type Indicator (Myers & Myers, 1980) is a forced-choice test based on Jung's personality typology which categorizes a person on four preferences:

- **I**ntroversion and **E**xtraversion (attitudes): I's tend to reflect before they act, while E's act before they reflect.
- i**N**tuition and **S**ensing (information-gathering): N's rely on abstract or theoretical information, while S's trust information that is concrete.
- **F**eeling and **T**hinking (decision-making): While F's decide based on emotions, T's involve logic and reason in their decisions.
- **J**udging and **P**erceiving (lifestyle): J's prefer structure in their lives, while P's like change.

MBTI correlates with the Big Five personality traits of extraversion and openness, to a lesser extent with agreeableness and consciousness, but not with neuroticism (McCrae & Costa, 1989).

The participants' characteristics are too homogeneous for experiments concerning gender, mother tongue or region, but we find interesting distributions in at least two of the four MBTI preferences: .45 I vs. .55 E, .54 N vs. .46 S, .72 F vs. .28 F, and .81 J and .19 P.

Personality measurement in general, and the MBTI is no exception, is a controversial domain. However, especially for scores on IE and NS dimensions, consensus is that they are correlated with personality traits. In the remainder of this paper, we will provide results on the prediction of personality types from features extracted from the linguistically analyzed essays.

Feature Extraction

While most stylometric studies are based on token-level features (e.g., word length), word forms and their frequencies of occurrence, syntactic features have been proposed as more reliable style markers since they are not under the conscious control of the author (Stamatatos et al., 2001).

We use Memory-Based Shallow Parsing (MBSP) (Daelemans et al., 1999), which gives an incomplete parse of the input text, to extract reliable syntactic features. MBSP tokenizes, performs a part-of-speech analysis, looks for chunks (e.g., noun phrase) and detects subject and object of the sentence and some other grammatical relations.

Features occurring more often than expected (based on the chi-square metric) in either of the two classes are extracted automatically for every document. Lexical features (*lex*) are represented binary or numerically, in *n*-grams. *N*-grams of both fine-grained (*pos*) and coarse-grained parts-of-speech (*cgp*) are integrated in the feature vectors. These features have been proven useful in stylometry (cf. Stamatatos et al., 2001) and are now tested for personality prediction.

Experiments in Personality Prediction and Discussion

We report on experiments on eight binary classification tasks (e.g., I vs. not-I) (cf. Table 1) and four tasks in which the goal is to distinguish between the two poles in the preferences (e.g., I vs. E) (cf. Table 2). Results are based on ten-fold cross-validation experiments with TiMBL (Daelemans & van den Bosch, 2005), an implementation of memory-based learning (MBL). MBL stores feature representations of training instances in memory without abstraction and classifies new instances by matching their feature representation to all instances in memory. We also report random and majority baseline results. Per training

document, a feature vector is constructed, containing comma-separated binary or numeric features and a class label. During training, TiMBL builds a model based on the training data by means of which the unseen test instances can be classified.

Task	Feature set	Precision	Recall	F-score	Accuracy
Introverted	lex 3-grams *random*	56.70% *44.1%*	84.62% *46.2%*	67.90%	64.14%
Extraverted	cgp 3-grams random	58.09% 54.6.%	98.75% *52.5%*	73.15%	60.00%
iNtuitive	cgp 3-grams *random*	56.92% *48.7.%*	94.87% *48.7%*	71.15%	58.62%
Sensing	pos 3-grams *random*	50.81% *40.3.%*	94.03% *40.3.%*	65.97%	55.17%
Feeling	lex 3-grams *random*	73.76% *72.6%*	99.05% *73.3.%*	84.55%	73.79%
Thinking	lex 1-grams *random*	40.00% *28.2.%*	50.00% *27.5%*	44.44%	65.52%
Judging	lex 3-grams *random*	81.82% *77.6%*	100.00% *76.9%*	90.00%	82.07%
Perceiving	lex 2-grams *random*	26.76% *6.9%*	67.86% *7.1%*	38.38%	57.93%

Table 1:TiMBL results for eight binary classification tasks

Table 1 suggests that tasks for which the class distributions are not skewed (I, E, N and S) achieve F-scores between 64.1% and 73.2%. As expected, results for Feeling and Judging are high, but the features and methodology still allow for a score around 40% for tasks with little training data.

Task	Feature set	F-score [INFJ]	F-score [ESTP]	Average F-score	Accuracy
I vs. E	lex 3-grams *random* *majority*	67.53%	63.24%	65.38%	65.52% *49.7%* *55.2%*
N vs. S	pos 3-grams *random* *majority*	58.65%	64.97%	61.81%	62.07% *44.8%* *53.8%*
F vs.T	lex 3-grams *random* *majority*	84.55%	13.64%	49.09%	73.79% *60.7%* *72.4%*
J vs. P	lex 3-grams *random* *majority*	90.00%	13.33%	51.67%	82.07% *63.5%* *80.7%*

Table 2:TiMBL results for four discrimination tasks

Table 2 shows results on the four discrimination tasks, which allows us to compare with results from other studies in personality prediction. Argamon et al. (2005) find appraisal adjectives and modifiers to be reliable markers (58% accuracy) of neuroticism, while extraversion can be predicted by function words with 57% accuracy. Nowson & Oberlander (2007) predict high/low extraversion with a 50.6% accuracy, while the system achieves 55.8% accuracy on neuroticism, 52.9% on agreeableness, and 56.6% on conscientiousness. Openness is excluded because of the skewed class distribution. Taking a top-down approach, Mairesse et al. (2007) report accuracies of 55.0% for extraversion, 55.3% for conscientiousness, 55.8% agreeableness, 57.4% for neuroticism, and 62.1% for openness.

For the *I-E* task - correlated to extraversion in the *Big Five* - we achieve an accuracy of 65.5%, which is better than Argamon et al. (2005) (57%), Nowson & Oberlander (2007) (51%), and Mairesse et al. (2007) (55%). For the *N-S* task - correlated to openness - we achieve the same result as Mairesse et al. (2007) (62%). For the *F-T* and *J-P* tasks, the results hardly achieve higher than majority baseline, but nevertheless something is learned for the minority class, which indicates that the features selected work for personality prediction, even with heavily skewed class distributions.

Conclusions and Future Work

Experiments with TiMBL suggest that the first two personality dimensions (Introverted-Extraverted and iNtuitive-Sensing) can be predicted fairly accurately. We also achieve good results in six of the eight binary classification tasks. Thanks to improvements in shallow text analysis, we can use syntactic features for the prediction of personality type and author.

Further research using the *Personae* corpus will involve a study of stylistic variation between the 145 authors. A lot of the research in author recognition is performed on a closed-class task, which is an artificial situation. Hardly any corpora – except for some based on blogs (Koppel et al., 2006) – have more than ten candidate authors. The corpus allows the computation of the degree of variability encountered in text on a single topic of different (types) of features when taking into account a relatively large set of authors. This will be a useful complementary resource in a field dominated by studies potentially overestimating the importance of these features in experiments discriminating between only two or a small number of authors.

Acknowledgements

This study has been carried out in the framework of the Stylometry project at the University of Antwerp. The "Computational Techniques for Stylometry for Dutch" project is funded by the National Fund for Scientific Research (FWO) in Belgium.

References

Argamon, S., Dhawle, S., Koppel, M. and Pennebaker, J. (2005), Lexical predictors of personality type, *Proceedings of the Joint Annual Meeting of the Interface and the Classification Society of North America.*

Campbell, R. and Pennebaker, J. (2003), The secret life of pronouns: Flexibility in writing style and physical health, *Psychological Science* 14, 60-65.

Daelemans, W. and van den Bosch, A. (2005), *Memory-Based Language Processing,* Studies in Natural Language Processing, Cambridge, UK: Cambridge University Press.

Daelemans, W., Bucholz, S. and Veenstra, J. (1999), Memory-Based Shallow Parsing, *Proceedings of the 3rd Conference on Computational Natural Language Learning CoNLL*, pp. 53-60.

Gill, A. (2003), Personality and language: The projection and perception of personality in computer-mediated communication, PhD thesis, University of Edinburgh.

Gill, A. & Oberlander J. (2002), Taking care of the linguistic features of extraversion, *Proceedings of the 24th Annual Conference of the Cognitive Science Society,* pp. 363-368.

Koppel, M., Schler, J., Argamon, S. and Messeri, E. (2006), Authorship attribution with thousands of candidate authors, *Proceedings of the 29th ACM SIGIR Conference on Research and Development on Information Retrieval*, pp. 659-660.

Mairesse, F., Walker, M., Mehl, M. and Moore, R. (2007), Using linguistic cues for the automatic recognition of personality in conversation and text, *Journal of Artificial Intelligence Research*.

McCrae, R. and Costa, P. (1989), Reinterpreting the Myers-Briggs Type Indicator from the perspective of the Five-Factor Model of Personality, *Journal of Personality* 57(1), 17-40.

Myers, I. and Myers, P. (1980), *Gifts differing: Understanding personality type*, Mountain View, CA: Davies-Black Publishing.

Nowson, S. and Oberlander, J. (2007), Identifying more bloggers. Towards large scale personality classification of personal weblogs, *Proceedings of International Conference on Weblogs and Social Media ICWSM*.

Sebastiani, F. (2002), Machine learning in automated text categorization, *ACM Computing Surveys* 34(1), 1-47.

Stamatatos, E., Fakotakis, N. and Kokkinakis, G. (2001), Computer-based authorship attribution without lexical measures, *Computers and the Humanities* 35(2), 193-214.

An Interdisciplinary Perspective on Building Learning Communities Within the Digital Humanities

Simon Mahony
simon.mahony@kcl.ac.uk
King's College London

Introduction

Recent research at the Centre for Computing in the Humanities at King's College London has focussed on the role and place of the digital humanities in the academic curriculum of Higher Education (see Jessop:2005, Jessop:forthcoming). This work is based on the experience of both our undergraduate and postgraduate programmes focusing particularly on the way in which students are encouraged to integrate the content of a variety of digital humanities courses and apply it to their own research project. In the case of the undergraduates this is developed in conjunction with their home department. These courses are designed to train not just the new generation of young scholars in our discipline but also the majority who will gain employment in a variety of professions in industry and commerce.

Our students come from a range of disciplines and backgrounds within the humanities and what is highlighted in each case is the necessity to ensure that their projects meet the scholarly criteria of their home disciplines and the interdisciplinary aspects of humanities computing. This emphasises the need for training the students in collaborative method and reflective practice; the need to build a community of learning which will lead to a community of practice. This paper discusses recent research and initiatives within distance learning, focussing on how these can be repurposed for campus-based courses, and is illustrated by the findings of their use in a digital humanities course.

Context

There have been a number of initiatives that are pertinent to this topic. The published report on the accomplishments of the Summit on Digital Tools for the Humanities convened in 2005 at the University of Virginia (http://www.iath.virginia.edu/dtsummit/) identified areas where innovative change was taking place that could lead to what they referred to as "a new stage in humanistic scholarship". The style of collaboration enabled by digital learning community tools is identified as one such area. This has been further reinforced at the National Endowment of the Humanities hosted Summit Meeting of Digital Humanities Centers and Funders held in April 2007 at the University of Maryland.

(https://apps.lis.uiuc.edu/wiki/display/DHC/Digital+Humanities+Centers+Summit)

On the summit wiki among the areas of research priorities and funder priorities John Unsworth lists:

- Collaborative work
- Teaching and learning
- Collaboration among scholars

(https://apps.lis.uiuc.edu/wiki/display/DHC/Areas+of+research+priorities%2C+funder+priorities)

Building communities of learning and developing strategies for collaborative working has been the subject of much study within the distance learning community (Anderson:2004, Brown:2006, Perry and Edwards:2005, Swan:2002, *et al*) and here it is argued that this also needs to be a consideration for campus-based academic programmes. The growing trend among undergraduate programmes of a movement away from set courses to the introduction of modular and credit systems means that students no longer follow a single programme of study. Their time is fragmented between their chosen course options and they often only come together with their peers for 'core courses'. Thus in the last two decades study has become more of an individual rather than community-based activity. This trend needs to be compensated for by teaching collaborative skills, the very same skills that are at the heart of the majority of digital humanities research projects.

The 'Community of Inquiry' model (developed by Garrison, Anderson and Archer:2000 and 2004) draws out the basic elements which overlap to form the educational experience of the distance learner: social, cognitive, and teaching presence. This model is used as a framework to analyse the effectiveness of asynchronous discussion methodologies (which encourage reflective practice), with particular regard for cognitive presence (where students construct meaning through communication with their peers, which is particularly important in the development of critical thinking), when used in campus-based courses.

Building Networked Communities for Collaborative and Reflective Teaching and Learning

The highly collaborative nature of modern research practice makes it clear that future humanities scholars need to be trained in the collaborative process and to understand the importance of critical reflection (Jessop:2005, Jessop:forthcoming). This emphasis on collaborative practice represents a shift in the academic culture of humanities away from the popular funding model of a single researcher towards one of team working where no single person has complete control or ownership. This is closer to models in operation in the sciences where progress is often based on team efforts and reports frequently have many authors; we may need to develop protocols that borrow some aspects of science research practice. The results of the author's limited research into the effectiveness of the collaborative process in medical research practice are also to be included in this study.

To develop an environment that fosters collaboration and reflection students should be actively encouraged to engage with each other both inside and outside of the classroom. With social software (MySpace, Facebook, LiveJournal, *inter alia*) students are already building networked communities, and the blog and wiki have provided educators with simple, readily available tools to build learning communities. The wiki can be deployed as an experiential and formative learning environment outside of the classroom with students able to create their own content, comment on each others, and share resources using tools like del.icio.us and MyIntute. The blog supports this with a less formal reflective space which belongs to the students themselves rather than their course. The asynchronous nature of these media gives students the opportunity to reflect on their classmates' contribution in the processes of creating their own (Swan:2000) and so instil the practice of critical reflection. Further, with simple applications such as MyYahoo and iGoogle students can draw all their varied networks together along with course and departmental webpages thus giving a single interface or 'Personal Learning Portal' (PLP) through which to access and manage their online resources.

In their PLP students create a web interface for their own digital environment that includes:

- Content management where they integrate both personal and academic interests
- A networking system for connection with others
- Collaborative and individual workspace
- Communications setup
- A series of syndicated and distributed feeds

This model is based on a presentation given by Terry Anderson at the Centre for Distance Education, University of London in March 2007:

http://www.cde.london.ac.uk/support/news/generic3307.htm). In this Anderson discusses how the Personal Learning Environment (PLE), such as that used at Athabasca, is an improvement on the popular Virtual Learning Environment (VLE). That argument is developed here with stress upon the further advantages of the PLP introduced earlier.

What is notable is that this model represents an approach rather than a specific application and is portable and not

dependant on a single department or even institution. This ensures sustainability as it allows and encourages students to take the tools and skills from one area and apply them in others (arguably the basis of humanities computing, see McCarty and Short:2002). At the same time it puts the emphasis for the responsibility for managing their own learning and web resources on the students.

In this approach learners are encouraged to interact and collaborate in a way that does not occur when static webpages are viewed with a traditional browser. The pages on a wiki and the student's PLP are dynamic and mutable as they can be edited by the user through their web browser. Learners gain the ability to enrich the material and, unlike a print publication where those annotations are only for personal use, make these available for others. Such exchanges of ideas are central to the processes of building communities of learning and it is in this way that knowledge grows as we are able to push the boundaries of scholarship. The model that is developing here is one in which the student moves from being a reader of other peoples' material to active engagement with that material; a transition from being a 'reader' to being an 'interpreter'.

Conclusion

Education is an academic, individual, and a social experience that requires a sustainable community of learning. The tools and experiences developed in the distance learning field can be re-purposed for the 'analogue' students. The suggested model based on the student's PLP is grounded in collaborative practice, uses asynchronous discussion to develop reflective practice and ensure cognitive presence; it is sustainable and portable. Putting this in the wider context, it is by building a community of learners that we will instil the cooperative, collaborative, and reflective skills needed for a community of humanities scholars; skills that are equally in demand outside of the academy. The tools have already been subject to limited trials in the digital humanities programmes at King's College London but the current academic year will see a more extensive application of them across our teaching. The final version this paper will report on the results of the experiences of both teachers and learners of this model applied to a humanities computing course. The work contributes to the pedagogy of the digital humanities in the academic curricula both within the teaching of humanities computing and the development of tools for collaborative research and on-line learning communities.

Bibliography

Anderson T (2004) 'Theory and Practice of Online Learning', *Athabasca University online books*: http://www.cde.athabascau.ca/online_book/ (last accessed 01/11/07)

Anderson T (2007) 'Are PLE's ready for prime time?' http://terrya.edublogs.org/2006/01/09/ples-versus-lms-are-ples-ready-for-prime-time/ (last accessed 03/11/07)

Biggs J (1999) *Teaching for quality and learning at university,* Open University Press.

Brown J S (2006) 'New Learning Environments for the 21st Century' http://www.johnseelybrown.com/newlearning.pdf (last accessed 30/10/07)

Garrison D, Anderson T, and Archer W (2000) 'Critical inquiry in a text-based environment: computer conferencing in higher education'. *The Internet and Higher Education* Volume 2, Issue 3 pp.87-105

Garrison D, Anderson T, and Archer W (2001) 'Critical thinking, cognitive presence, and computer conferencing in distance education'. *The American Journal of Distance Education* Volume 15, Issue 1 pp.7-23.

Garrison D, Anderson T, and Archer W (2004) 'Critical thinking and computer conferencing: A model and tool for assessing cognitive presence'. http://communitiesofinquiry.com/documents/CogPresPaper_June30_.pdf (last accessed 26/10/07)

Jessop M (2005) 'Teaching, Learning and Research in Final year Humanities Computing Student Projects', *Literary and Linguistic Computing*. Volume 20, Issue 3.

Jessop M (Forthcoming) 'Humanities Computing Research as a Medium for Teaching and Learning in the Digital Humanities', *Digital Humanities Quarterly.*

Mahony S (2007) 'Using digital resources in building and sustaining learning communities', *Body, Space & Technology,* Volume 07/02.

McCarty W and Short H (2002) 'A Roadmap for Humanities Computing' http://www.allc.org/reports/map/ (last accessed 04/11/07)

Moon J (2004) *A Handbook of Reflective and Experiential Learning: Theory and Practice,* RoutledgeFalmer

Perry B and Edwards M (2005) 'Exemplary Online Educators: Creating a Community of Inquiry' http://www.odlaa.org/events/2005conf/ref/ODLAA2005PerryEdwards.pdf (last accessed 04/11/07)

Richardson C and Swan K (2003) 'Examining social presence in online courses in relation to students' perceived learning and satisfaction', *Journal of Asynchronous Learning* Volume 7, Issue 1 pp. 68-88

Siemens G (2006) *Knowing Knowledge*, Lulu.com

Swan K (2000) 'Building Knowledge Building Communities: consistency, contact and communication in the virtual classroom', *Journal of Educational Computing Research*, Volume 24, Issue 4 pp. 359-383

Swan K (2002) 'Building Learning Communities in Online Courses: the importance of interaction', *Education, Communication & Information,* Volume 2, Issue 1 pp. 23-49

Terras M (2006) 'Disciplined: Using Educational Studies to Analyse 'Humanities Computing'', *Literary and Linguistic Computing*, Volume 21, pp. 229-246.

The Middle English Grammar Corpus - a tool for studying the writing and speech systems of medieval English

Martti Mäkinen
martti.makinen@uis.no
University of Stavanger, Norway

The Middle English Grammar Project

The Middle English Grammar Project (MEG), shared by the Universities of Glasgow and Stavanger, is working towards the description of Middle English orthography, morphology and phonology. MEG is among the first attempts to span the gap between Jordan's *Handbuch der mittelenglischen Grammatik: Lautlehre* (1925) and now. Our aim is to combine the advances in Middle English dialectology in the latter half of the 20th century and the computing power currently available in the service of writing an up-to-date grammar of Middle English.

Middle English dialects and dialectology

The study of Middle English dialects took a giant leap forward by Angus McIntosh's insight that Middle English texts represent distinct regional varieties in their spelling systems, and therefore the spelling variants of these texts could be studied in their own right and not merely as reflections of the then speech systems, i.e. dialects (McIntosh 1963). This multitude of regional spellings arose when English had been replaced by French and Latin in all important aspects for nearly two centuries after the Norman Conquest: the re-introduction of English into literary and utilitarian registers from the thirteenth century onwards was not governed by any nationwide standard and thus English was written according to each scribe's perception of the 'correct' spelling. McIntosh's vision led to a project that grew into *A Linguistic Atlas of Late Mediaeval English* (LALME; 1986).

Aims of MEG

The Middle English Grammar project builds on the work of the LALME team and aims at producing a description of Middle English orthography, phonology and morphology, from 1100 to 1500. Here we use the term grammar in a wide, philological sense. Eventually, the grammar is meant as an replacement to Richard Jordan's *Handbuch der mittelenglischen Grammatik: Lautlehre*, and to provide a reference point for the students and scholars of Middle English in the form of a broad description of the Middle English usages accompanied by more specific county studies, and eventually also all the base material we accumulate for this task (Black, Horobin, and Smith 2002: 13).

The Middle English Grammar: How?

The first task of MEG is to compile a corpus of Middle English texts localized in LALME. The corpus is called *The Middle English Grammar Corpus*, or MEG-C (the first installment forthcoming 2007). Secondly, the corpus texts need appropriate lemmatization and annotation in order to be usable in the course of MEG.

Linguistic data is collected by transcribing extracts from either the original manuscripts, or good-quality microfilms. The prioritised material are the texts that were localized in LALME, although later texts that were not analysed for LALME will be taken into account as well. LALME covers years 1350-1450 (-1500); the material for the studies in 1100-1350 will be drawn from *A Linguistic Atlas for Early Middle English* (LAEME) (Laing and Lass, forthcoming 2007).

The manuscript texts are represented by 3,000-word extracts (or in *toto*, if shorter), which should be sufficiently for studies on orthography, phonology and morphology. The planned corpus will sample c. 1,000 texts, therefore the projected size of the corpus is 2.5-3 M words.

The conventions of transcription have been derived from those of the LAEME and *A Linguistic Atlas of Older Scots* projects (LAOS), with certain modifications. The most important questions that have been addressed during the transcription process have been whether to emphasise fidelity to the original vs. wieldy transcripts, and should the transcripts offer an interpretative reading of the manuscript text rather than the scribe's actual pen strokes. According to the principles chosen, the transcriptions attempt to capture the graphemic and broad graphetic details, but not necessarily each detail on the level of individual handwriting (Black, Horobin, and Smith 2002: 11).

MEG-C: lemmatization, annotation, publication

The second practical task is to lemmatize and to annotate the Corpus. Previous historical English corpora (*Helsinki Corpus*, *Middle English Medical Texts*) show the limitations the lack of lemmas set to the corpus user when tackling the variety of spellings attested to by Middle English texts. The lemmas in MEG-C will have an *Oxford English Dictionary* headword. There will also be another cue in the source language (the direct source languge before Middle English, usually Old English, French/Anglo-Norman or Latin). These two reference points on either side of Middle English will provide the user the means to search for occurrences of a lexical item even when the full range of spelling variation in Middle English is not known.

As regards the annotation of words of a text, they are divided into bound morphemes and other spelling units (this system is partly derived from Venezky (1970)). Each word is divided into a word initial sequence containing Onset and Nucleus, and they are followed by a series of Consonantal and Vowel Spelling Units. Each spelling unit is also given the equivalents in the source language and in Present Day English, thus enabling the search for e.g. all the ME reflexes of OE [a:] or or the spelling variants in Middle English that correspond to Present Day English word initial spelling sh-.

For the task of annotation and lemmatization the corpus is rendered into a relational database. The database plan has tables for different extralinguistic information, and the actual texts will be entered word by word, i.e. in the table for corpus texts, there will be one record for each word. The annotation plan we are intending to carry out should result in a corpus where one can search for any combination of extralinguistic factors and spelling units with reference points embedded in the actual Middle English texts and also in the source language and PDE spelling conventions.

The first installment of MEG-C will be published in 2007, containing roughly 30 per cent of the texts in the planned corpus in ASCII format. It will be on the Internet, accessible for anyone to use and download. Our aim with publication is two-fold: firstly, we will welcome feedback of any kind, and especially from scholars who know the texts well; secondly, we want to encourage and to see other scholars use the corpus.

References

Black, Merja, Simon Horobin and Jeremy Smith, 2002. 'Towards a new history of Middle English spelling.' In P. J. Lucas and A.M. Lucas (eds), *Middle English from Tongue to Text*. Frankfurt am Main: Peter Lang, 9-20.

Helsinki Corpus = *The Helsinki Corpus of English Texts* (1991). Department of English, University of Helsinki. Compiled by Matti Rissanen (Project leader), Merja Kytö (Project secretary); Leena Kahlas-Tarkka, Matti Kilpiö (Old English); Saara Nevanlinna, Irma Taavitsainen (Middle English); Terttu Nevalainen, Helena Raumolin-Brunberg (Early Modern English).

Horobin, Simon and Jeremy Smith, 1999. 'A Database of Middle English Spelling.' *Literary and Linguistic Computing* 14: 359-73.

Jordan, Richard, 1925. *Handbuch der mittelenglischen Grammatik*. Heidelberg: Winter's Universitätsbuchhandlung.

LAEME = Laing, Margaret, and Lass, Roger, forthcoming 2007. *A Linguistic Atlas of Early Middle English*. University of Edinburgh. Nov. 22nd, 2007. http://www.lel.ed.ac.uk/ihd/laeme/laeme.html

Laing, M. (ed.) 1989. *Middle English Dialectology: essays on some principles and problems by Angus McIntosh, M.L. Samuels and Margaret Laing*. Aberdeen: Aberdeen University Press.

LALME = McIntosh, M.L., Samuels, M.L. and Benskin, M. (eds.) 1986. *A Linguistic Atlas of Late Mediaeval English. 4 vols*. Aberdeen: Aberdeen University Press. (with the assistance of M. Laing and K. Williamson).

LAOS = Williamson, Keith, forthcoming 2007. *A Linguistic Atlas of Older Scots*. University of Edinburgh. Nov. 22nd, 2007. http://www.lel.ed.ac.uk/research/ihd/laos/laos.html

McIntosh, A. 1963 [1989]. 'A new approach to Middle English dialectology'. *English Studies* 44: 1-11. repr. Laing, M. (ed.) 1989: 22-31.

Middle English Medical Texts = Taavitsainen, Irma, Pahta, Päivi and Mäkinen, Martti (compilers) 2005. *Middle English Medical Texts*. CD-ROM. Amsterdam: John Benjamins.

Stenroos, Merja, 2004. 'Regional dialects and spelling conventions in Late Middle English: searches for (th) in the LALME data.' In M. Dossena and R. Lass (eds), *Methods and data in English historical dialectology*. Frankfurt am Main: Peter Lang: 257-85.

Stenroos, Merja, forthcoming 2007. 'Sampling and annotation in the Middle English Grammar Project.' In Meurman-Solin, Anneli and Arja Nurmi (eds) *Annotating Variation and Change* (Studies in Variation, Contacts and Change in English 1). Research Unit for Variation, Change and Contacts in English, University of Helsinki. http://www.helsinki.fi/varieng/journal/index.html

Venezky, R., 1970. *The Structure of English Orthography*. The Hague/Paris: Mouton.

Designing Usable Learning Games for the Humanities: Five Research Dimensions

Rudy McDaniel
rudy@mail.ucf.edu
University of Central Florida, USA

Stephen Fiore
sfiore@ist.ucf.edu
University of Central Florida, USA

Natalie Underberg
nunderbe@mail.ucf.edu
University of Central Florida, USA

Mary Tripp
mtripp@mail.ucf.edu
University of Central Florida, USA

Karla Kitalong
kitalong@mail.ucf.edu
University of Central Florida, USA

J. Michael Moshell
jm.moshell@cs.ucf.edu
University of Central Florida, USA

A fair amount of research suggests that video games can be effective tools for learning complex subject matter in specific domains (Cordova & Lepper, 1996; Ricci, Salas, & Cannon-Bowers, 1996; Randel et al., 1992; Fiore et al., 2007; Garris & Ahlers et al., 2002). Although there has been much work in situating "serious gaming" (Sawyer, 2002) as a legitimate vehicle for learning in all types of disciplines, there is no central source for research on crafting and designing what is considered to be a *usable* game for the humanities. In this paper, we discuss research issues related to the design of "usable" humanities-based video games and situate those research questions along five interrelated dimensions.

We first provide our definition of usability as used in this context. A usable humanities game is a game which is both functionally capable in terms of an interface and its human interactions as well as appropriately designed to support the types of learning objectives attached to its scenarios or levels. This definition is formed from the convergence of traditional interface usability and learning effectiveness. We support our delineation of research questions using theoretical work from scholars writing about gaming as well as applied examples from our own research and game development prototypes. Drawing from three years of experience building a variety of games, and literature culled from our various fields, we discuss some of the unique research questions that designing for the humanities pose for scholars and game designers. We separate these questions across five dimensions, each with representative humanities learning questions:

1. **Participation**: how can we encourage diverse groups of students and researchers to interact together in virtual space? How can we design a space that is appealing and equitable to both genders and to a diverse range of demographic profiles?

2. **Mechanics**: how do you design algorithms that are applicable to the types of tasks commonly sought in humanities courses? For instance, how might one develop an algorithm to "score" morally relative gameplay decisions in an ethical dilemma?

3. **Ontology**: how is a player's self-image challenged when shifting from a real to what Gee (2003) calls a projected identity, and how is this process changed through the use of varying perspectives such as first-person or third-person perspective?

4. **Hermeneutics**: how do we probe the hidden layers that exist between game worlds, source content, and human computer interfaces? What "voices" are encouraged in virtual worlds, and what voices are repressed within this rhetorical space?

5. **Culture**: how are the cultural facets of primary source content mediated through the process of digitization?

To consider these dimensions, we craft a theoretical base formed from a wide range of researchers such as Gee (2003), Prensky (2001), Squire (2002), and Jenkins (2006a, 2006b). From Gee, we draw inspiration from his discussion of well-designed games and his exploration of the implicit learning occurring in several different genres of digital video games. Much of Gee's work has involved applying insights from the cognitive sciences to traditional humanities domains such as literature in order to explore identity, problem solving skills, verbal and nonverbal learning, and the transfer of learned abilities from one task to another. Marc Prensky notes that musical learning games in the humanities have been used for hundreds of years -- Bach's *Well Tempered Clavier* and *The Art of the Fugue* are his "learning games," simple to complex musical exercises that build skill. Prensky's work also informs our historical analysis as well as insights from the pioneers working in the field of serious gaming for military applications.

Jenkins' work in applying the interests of gaming fans as critical lenses provides insight for both formative guidelines and post-task measures of "success" in learning game environments. These gaming discourse communities often form wildly active and influential fan groups, and these groups cultivate their own forms of expression and understanding through complex jargon, virtual initiations, and ritualistic rules and procedures in virtual interaction. Gaming environments and virtual worlds have also been shown to offer rich sources of material for investigating notions of gender, race, ethnicity, and cultural identity (Berman & Bruckman, 2001; Squire, 2002).

Building on the work of these scholars, others have extended these general notions of digital game based learning to account for specific curricula or learning objectives such as media project management for humanities computing (McDaniel et al., 2006). To build these humanities learning games, we have assembled an interdisciplinary team composed of faculty members from Digital Media, English, and Philosophy. Individuals from this group have worked in a variety of capacities, as script writers, artists, programmers, and producers. Undergraduate and graduate students, in both classroom and research lab roles, have worked and contributed to each of these games in varying capacities. Five different games have been produced through these collaborations:

1. *Discover Babylon* (Lucey-Roper, 2006), a game produced by the Federation of American Scientists, the Walters Art Museum in Baltimore, and UCLA's Cuneiform Digital Library Initiative (CDLI). One of our team members developed the storyline for this game.

2. The *Carol Mundy Underground Railroad* game (Greenwood-Ericksen et al., 2006) examines issues of African-American history and culture and leads a player on an adventure from a Southern plantation to safety in the North through the Underground Railroad's system of safehouses. This was a "modded" game built atop Neverwinter Nights.

3. The *Turkey Maiden* (Underberg, forthcoming) is an educational computer game project based on a version of Cinderella collected by folklorist Ralph Steele Boggs in 1930s Ybor City, Florida. This variant of the Cinderella story, called "The Turkey Maiden" (from Kristin Congdon's anthology *Uncle Monday and Other Florida Tales*, 2001) forms the narrative structure of the game, which has been further developed by integrating specific tasks that the heroine Rosa ("Cinderella") must successfully complete to advance in the game that are based in lessons to be learned by the player about Florida history and culture.

4. *Chaucer's Medieval Virtual World Video Game* is a virtual medieval world game based on Chaucer's Canterbury Tales. This tale emphasizes the battle between men's perceived authority and women's struggles for power. This game uses Chaucer's narrative gap as a springboard for a virtual medieval quest. In addition to experiencing historical scenarios, the knight will disguise his identity and experience the world from various gender and social classes, the Three Estates of Clergy, Nobility, and Peasantry as well as the three feminine estates of virgin, widow, and wife.

5. *The Medium* is a prototype ethics game designed using the *Torque* game engine. This three year project was funded by the University of Central Florida's Office of Information Fluency and is in its earliest stages of design. The game pairs a time travel theme with a switchable first and third person perspective and also includes an environmentalist subtext.

We will use brief examples culled from these games to

support our theoretical assertions and discuss ways in which usable humanities games can act as a springboard for bringing together subject matter experts, technologists, and learners. In their 2006 report on Cyberinfrastructure for the Humanities and Social Sciences, the American Council of Learned Societies writes the following in regards to digital networked technologies for the humanities: "A cyberstructure for humanities and social science must encourage interactions between the expert and the amateur, the creative artist and the scholar, the teacher and the student. It is not just the collection of data—digital or otherwise—that matters: at least as important is the activity that goes on around it, contributes to it, and eventually integrates with it" (14). Our goal is to foster this type of humanistic, communicative environment using new technologies for virtual worlds, usability testing, and game-based learning environments. We hope that the scholarly community working the field of digital humanities can help us to explore and refine both theoretical models and applied technologies related to this goal.

References

American Council of Learned Societies. 18 July, 2006. "Our Cultural Commonwealth: The Report of the American Council of Learned Societies' Commission on Cyberinfrastructure for the Humanities and Social Sciences." 13 October 2007. <http://www.acls.org/cyberinfrastructure/OurCulturalCommonwealth.pdf>.

Berman, J. & Bruckman, A. (2001). The Turing Game: Exploring Identity in an Online Environment. *Convergence*, 7(3), 83-102.

Congdon, K. (2001). *Uncle Monday and other Florida Tales.* Jackson, MS: University Press of Mississippi.

Cordova, D. I., & Lepper, M. R. (1996). Intrinsic motivation and the process of learning: beneficial effects of hypercontextualization, personalization, and choice. *Journal of Educational Psychology*, 88(4), 715-730.

Fiore, S. M., Metcalf, D., & McDaniel, R. (2007). Theoretical Foundations of Experiential Learning. In M. Silberman (Ed.), *The Experiential Learning Handbook* (pp. 33-58): John Wiley & Sons.

Garris, R., R. Ahlers, et al. (2002). "Games, Motivation, and Learning: A Research and Practice Model." *Simulation Gaming 33*(4), 441-467.

Gee, J. P. (2003). *What Video Games Have to Teach Us About Learning and Literacy*. New York: Palgrave Macmillan.

Greenwood-Ericksen, A., Fiore, S., McDaniel, R., Scielzo, S., & Cannon-Bowers, J. (2006). "Synthetic Learning Environment Games: Prototyping a Humanities-Based Game for Teaching African American History." *Proceedings of the 50th Annual Meeting of the Human Factors and Ergonomics Society*. Santa Monica, CA: Human Factors and Ergonomics Society.

Jenkins, H. (2006). *Convergence Culture.* New York: New York University Press.

Jenkins, H. (2006). *Fans, Bloggers, and Gamers.* New York: New York University Press.

Lucey-Roper M. (2006), *Discover Babylon:* Creating A Vivid User Experience By Exploiting Features Of Video Games And Uniting Museum And Library Collections, in J. Trant and D. Bearman (eds.). *Museums and the Web 2006: Proceedings,* Toronto: Archives & Museum Informatics, published March 1, 2006 at <http://www.archimuse.com/mw2006/papers/lucey-roper/lucey-roper.html>.

McDaniel, R., Fiore, S. M., Greenwood-Erickson, A., Scielzo, S., & Cannon-Bowers, J. A. (2006). Video Games as Learning Tools for Project Management. *The Journal of the International Digital Media and Arts Association,* 3(1), 78-91.

Prensky, M. (2001). *Digital Game-Based Learning.* New York: McGraw-Hill.

Randel, J. M., Morris, B. A., Wetzel, C. D., & Whitehill, B. V. (1992). The effectiveness of games for educational purposes: a review of recent research. *Simulation & Gaming, 23*(3), 261-276.

Ricci, K. Salas, E., & Cannon-Bowers, J. (1996). "Do Computer-Based Games Facilitate Knowledge Acquisition and Retention?" *Military Psychology 8*(4), 295-307.

Sawyer, B. (2002). *Serious Games: Improving Public Policy through Game-based Learning and Simulation.* Retrieved June 24 2006, from http://www.seriousgames.org/images/seriousarticle.pdf.

Squire, K. (2002). Cultural Framing of Computer/Video Games. *Game Studies*, 2(1).

Underberg, Natalie (forthcoming). "The Turkey Maiden Educational Computer Game." In *Folklife in Education Handbook*, Revised Edition, Marshall McDowell, ed.

Picasso's Poetry: The Case of a Bilingual Concordance

Luis Meneses
ldmm@cs.tamu.edu
Texas A&M University, USA

Carlos Monroy
cmonroy@cs.tamu.edu
Texas A&M University, USA

Richard Furuta
furuta@cs.tamu.edu
Texas A&M University, USA

Enrique Mallen
mallen@shsu.edu
Sam Houston State University, USA

Introduction

Studying Picasso as writer might seem strange, considering that the Spanish artist is mostly known for his paintings. However, in the Fall of 2006 we began working on Picasso's writings. Audenaert, et.al. [1], describe the characteristics of Picasso's manuscripts, and the challenges they pose due to their pictorial and visual aspects. With over 13,000 artworks up-to-date catalogued, the On-line Picasso Project [5] includes a historical narrative of relevant events in the artist's life. In this paper we discuss the contents of the texts—from the linguistic standpoint—and the implementation of a bilingual concordance of terms based on a red-black tree. Although concordances have been widely studied and implemented within linguistics and humanities, we believe that our collection raises interesting challenges; first because of the bilingual nature of Picasso's poems, since he wrote both in Spanish and French, and second, because of the connection between his texts and his paintings. The work reported in this paper focuses on the first issue.

Integrating Texts Into the On-line Picasso Project

A catalogue raisonné is a scholarly, systematic list of an artist's known works, or works previously catalogued. The organization of the catalogues may vary—by time period, by medium, or by style—and it is useful to consult any prefatory matter to get an idea of the overall structure imposed by the cataloguer. Printed catalogues are by necessity constrained to the time in which they are published. Thus, quite often catalogue raisonnés are superseded by new volumes or entirely new editions, which may (or may not) correct an earlier publication [2]. Pablo Picasso's artistic creations have been documented extensively in numerous catalogs. Chipp and Wofsy [3], started publishing a catalogue raisonné of Picasso's works that contains an illustrated synthesis of all catalogues to date on the works of Pablo Picasso.

In the Fall of 2007 Picasso's texts were added to the collection along with their corresponding images, and a concordance of terms both in Spanish and French was created. The architecture created for Picasso's poems is partially based on the one we developed for the poetry of John Donne [7]. As often happens in the humanities, each collection has its own characteristics, which makes a particular architecture hard if not impossible to reuse directly. For example, Donne's texts are written in English; Picasso in contrast wrote both in Spanish and French.

The Concordance of Terms

A concordance, according to the Oxford English Dictionary, is "an alphabetical arrangement of the principal words contained in a book, with citations of the passages in which they occur." When applied to a specific author's complete works, concordances become useful tools since they allow users to locate particular occurrences of one word, or even more interestingly, the frequency of such words in the entire oeuvre of an author. Apparently, the first concordances in English ever put together were done in the thirteenth century, and dealt with the words, phrases, and texts in the Bible. Such concordances were intended for the specialized scholar of biblical texts and were never a popular form of literature. As might be expected, these were soon followed by a Shakespeare concordance.

A concordance of the literary works of Pablo Picasso has more in common with a Biblical concordance than with a Shakespearian concordance, due to the manner in which the Spanish artist/poet composed his poems. Many critics have pointed out the heightened quality of words in Picasso's texts, a value that surpasses their own sentential context. One gets the impression that words are simply selected for their individual attributes and are then thrown together in the poems. Picasso himself appears to admit using this technique when he is quoted as saying that "words will eventually find a way to get along with each other." For this reason, readers of Picasso's poems become well aware of the frequent recurrence of certain "essential words," which one is then eager to locate precisely in each of the poems to determine significant nuances.

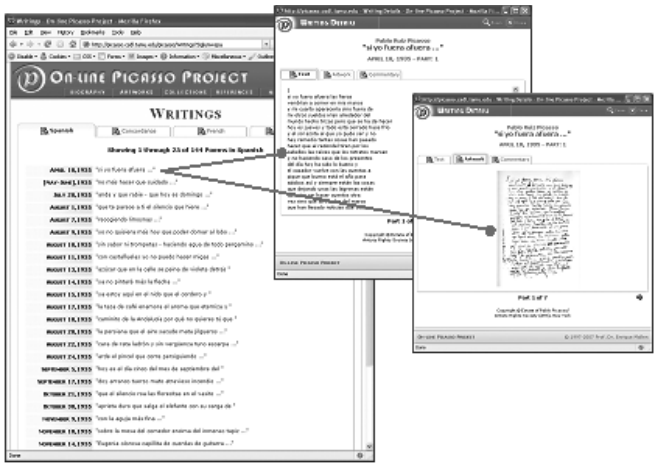


Figure 1. A tabbed interface presents users with French and Spanish Texts in chronological order. On the right are text and image presentations

By narrowing down the number of these "essential words," the concordance also allows users to delimit the "thematic domain" elaborated in Picasso's writings. Clearly many words deal with physical duress and the confrontation between good and evil, as manifestations of concrete human suffering during the Spanish Civil War and the German occupation of France in World War II. By identifying the range of words employed, users can clearly determine the political and cultural environment that surrounds Picasso's artistic creations during this period.

Nevertheless, one must not forget that Picasso's main contribution to the world is that of a plastic artist. A Concordance will allow users to identify each of the words Picasso used and link them to specific graphic images in his artworks. It has been argued that Picasso's poetry is quite "physical" (he often refers to objects, their different colors, textures, etc.). Even in his compositional technique, one gets a sense that the way he introduces "physical" words into his poems emulates the manner in which he inserted "found objects" in his cubist collages. Many critics have pointed out, on the other hand, that Picasso's art, particularly during his Cubist period, is "linguistic" in nature, exploring the language of art, the arbitrariness of pictorial expression, etc.

Mallen [6] argues that Cubism explored a certain intuition Picasso had about the creative nature of visual perception. Picasso realized that vision involves arbitrary representation, but, even more importantly, that painting also does. Once understood as an accepted arbitrary code, painting stopped being treated as a transparent medium to become an object in its own right. From then on, Picasso focused his attention on artistic creation as the creative manipulation of arbitrary representations. Viewed in this light, painting is merely another language, governed by similar strict universal principles as we find in verbal language, and equally open to infinite possibilities of expression. A concordance allows users to see these two interrelated aspects of Picasso's career fleshed out in itemized form.

Concordances are often automatically generated from texts, and therefore fail to group words by classes (lexical category, semantic content, synonymy, metonymy, etc.) The Concordance we are developing will allow users to group words in such categories, thus concentrating on the network of interrelations between words that go far beyond the specific occurrences in the poems.

Picasso is a bilingual poet. This raises several interesting questions connected to what has been pointed out above. One may wonder, for instance, if Picasso's thematic domain is "language-bound," in other words, whether he communicates certain concepts in one language but not in the other. A Concordance will allow users to set correspondences between words in one language and another. Given Picasso's strong Spanish heritage, it would be expected that concrete ideas (dealing with food, customs, etc) will tend to be expressed exclusively in Spanish, while those ideas dealing with general philosophical and religious problems will oscillate between the two languages. The possible corroboration of this point is another objective of the planned Concordance.

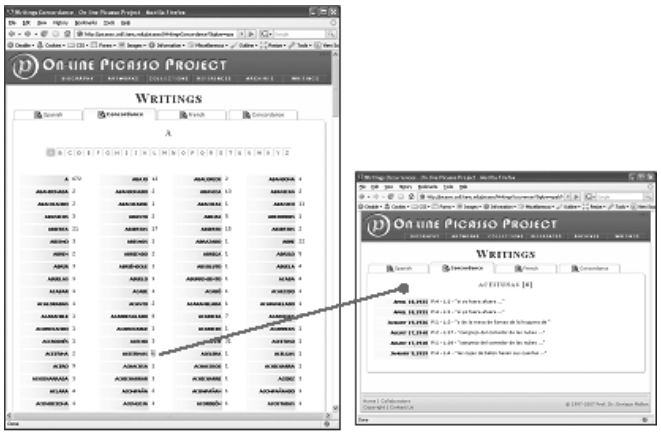


Figure 2. Concordance with term frequency. On the right, term-in-context display.

The Red Black Tree Implementation

Term concordances require the extraction of terms from a large corpus along with the metadata related to their occurrences, an operation often computationally expensive. The Digital Donne for instance, is a pre-computed concordance. On the other hand, repositories of texts are under constant revision as errors are detected and corrected. When the corpus is modified, part or the entire concordance has to be rebuilt. To solve this problem, the Picasso's concordance is computed on the fly, without requiring any previous processing.

The repository of poems has initially been divided into poems, stanza, and lines, then stored in a database. Using standard join operations, the poems are reconstructed, allowing the terms to be retrieved along additional metadata such as title, section, and line number. Once the poems have been reconstructed, each poem line is broken down into terms, which are defined as a series of characters delimited by a textual space. Each

occurrence of each term in turn, is inserted into the data structure, as well as its occurrence metadata.

Our algorithm consists of an augmented data structure composed of a Red Black Tree [4,8], where each node represents one term found in Picasso's writings and is used as pointer to a linked list.A Red Black Tree is a self balanced binary tree, that can achieve insert, delete, and search operations in O(log n) time. Because only insertions and random access is not required on the linked list, term occurrences can be traversed sequentially in O(n) time. Picasso's literary legacy can be browsed and explored using titles as surrogates, which are ordered by date and by the term concordance. Each entry provides the number of occurrences and its corresponding list, which can be used as index to browse the poems.

The retrieval process for terms and their occurrences is carried out by selecting specific terms or choosing from an alphabetical index of letters. To achieve this, a subtree is extracted from the data structure and it is traversed, obtaining every occurrence of a term along with additional metadata including a unique poem identifier, page and line number. Extensible Stylesheet Language Transformations (XSLTs) are used to transform the resulting XML output, and extract the specific term occurrences within a line of the poem and produce a surrogate, which is composed of a portion of the text. Additionally, this surrogate gives access to the corresponding poems through hyperlinks.

A new component of our implementation is a Spanish-French thesaurus that correlates terms in both languages, along with their meanings and commentary. Because our concordance is created on the fly, we have to devise a mechanism to support this. Additionally, this approach still remains to be tested with different corpuses in other languages, especially where terms are separated uniquely and spaces between them play a different role in language constructs. The term extraction algorithm is efficient using spaces as delimiters—a common case both in Spanish and French. However, other languages might include composite words.

Acknowledgements

This material is based upon work supported by the National Science Foundation under Grant No. IIS-0534314.

References

1. Audenaert, N., Karadkar, U., Mallen, E., Furuta, R., and Tonner, S. "Viewing Texts: An Art-Centered Representation of Picasso's Writings," *Digital Humanities* (2007): 14-16.

2. Baer, B. *Picasso Peintre-Graveur. "Catalogue Raisonné de L'oeuvre Gravé et des Monotypes."* 4 Vols. Berne, Editions Kornfeld, 1986-1989.

3. Chipp, H. and Wofsy, A. *The Picasso Project*. San Francisco: Alan Wofsy Fine Arts. 1995-2003.

4. Cormen, T., Leiserson, C., Rivest, R., and Stein, C., *Introduction to Algorithms*, The MIT Press, 2nd Edition, 2001.

5. Mallen, Enrique. "The On-line Picasso Project." http://picasso.tamu.edu accessed October 2007.

6. Mallen, Enrique. *The Visual Grammar of Pablo Picasso*. Berkeley Insights in Linguistics & Semiotics Series. New York: Peter Lang. 2003.

7. Monroy, C., Furuta, R., and Stringer, G., "Digital Donne: Workflow, Editing Tools, and the Reader's Interface of a Collection of 17th-century English Poetry." *Proceedings of JCDL 2007*, Vancouver, B.C. 2007.

8. Red-Black Tree. Accessed 2007-11-15, http://en.wikipedia.org/wiki/Red_black_tree

Exploring the Biography and Artworks of Picasso with Interactive Calendars and Timelines

Luis Meneses
ldmm@cs.tamu.edu
Texas A&M University, USA

Richard Furuta
furuta@cs.tamu.edu
Texas A&M University, USA

Enrique Mallen
mallen@shsu.edu
Sam Houston State University, USA

Introduction

Scholars and general users of digital editions face a difficult and problematic scenario when browsing and searching for resources that are related to time periods or events. Scrolling continuously through a long list of itemized search results does not constitute an unusual practice for users when dealing with this type of situation. The problem with this searching mechanism is that a notion of the corresponding dates or keywords associated with the event are required and constitute a precondition to a successful search.

An ordered list is unable to provide the required affordances and constraints that users need and desire to conduct scholarly research properly. It is a common practice among users to utilize the search mechanism present in most web browsers, and then perform another search among the obtained results to "narrow down" or limit the results to a smaller working set that is easier to manage. The use of an external search mechanism in a digital edition is a strong indicator that improved interfaces must be designed, conceived and implemented, just to achieve the sole purpose of facilitating scholarly research.

Background

The Online Picasso Project (OPP) is a digital collection and repository maintained by the Center for the Study of Digital Libraries at Texas A&M University, and the Foreign Languages Department at Sam Houston State University. As of November 2007, it contains 13704 catalogued artworks, 9440 detailed biographical entries, a list of references about Picasso's life and works, and a collection of articles from various sources regarding the renowned Spanish artist.

How does the OPP provide its content? Java Servlets are used to retrieve the documents and metadata from a MySQL database. As a result, a Apache Tomcat web server outputs a XML which is linked to XSLTs and CSS. The later are used to perform a dynamic transformation into standards-compliant HTML, achieving a clear separation between content and presentation.

The OPP includes an interface that allows scholars and users in general to browse through the significant events in his life, artworks, and a list of museums and collections that hold ownership to the various art objects created by the artist during his lifetime. The implemented navigation scheme works well for experienced scholars who have a deep knowledge of Picasso's life and works. The amount of available information can be overwhelming to the project audience, composed primarily of art scholars and historians, because of its magnitude and painstaking detail.

The Humanities rely profoundly on dates to create a strong relationship between events and documents. It is obvious to assume that key events influenced Picasso in such a way, that they caused significant changes in artistic style and expression. The OPP contains a vast amount of information that could be used in conjunction with the proposed interfaces, in order to help answer this type of inquiries. The calendars and timelines in the OPP propose an alternate method for exploring an existing document collection since they use date-related metadata as a discriminating factor, instead of an ordering criterion. Dates are used to provide mechanisms and interfaces that allow rich exploration of the artist's legacy in order to get a more whole and concise understanding of his life

Calendars

The calendar interfaces were developed to provide a timetable for the creation of artworks and occurrence of events cataloged in the OPP. Their purpose is to provide with a quick glance, relevant biographical and artistic dates. Additionally, the calendars provide means for formulating direct comparisons between dates within a single year. months and seasons.

The calendar interfaces have 5 display possibilities to filter results, which apply to the artworks and to the narrative:

1. Show start date and end date: used to display "exact matches".
2. Show start date or end date:
3. Show start date only:
4. Show End date only
5. Show Ranges of dates

Surrogates are provided in the form of artwork thumbnails and textual description of the events. Clicking on the specified date, month or season produces a web page, where detailed descriptions can be accessed.

Colors were added to the dates containing items, to show the distribution of the artworks and events. The design decision to include additional stratification schemes relates to the research goal of providing an enhanced browsing mechanism. The inclusion of this feature does not implicate any greater additional processing of the data, but it provides a richer environment for browsing the document collection.

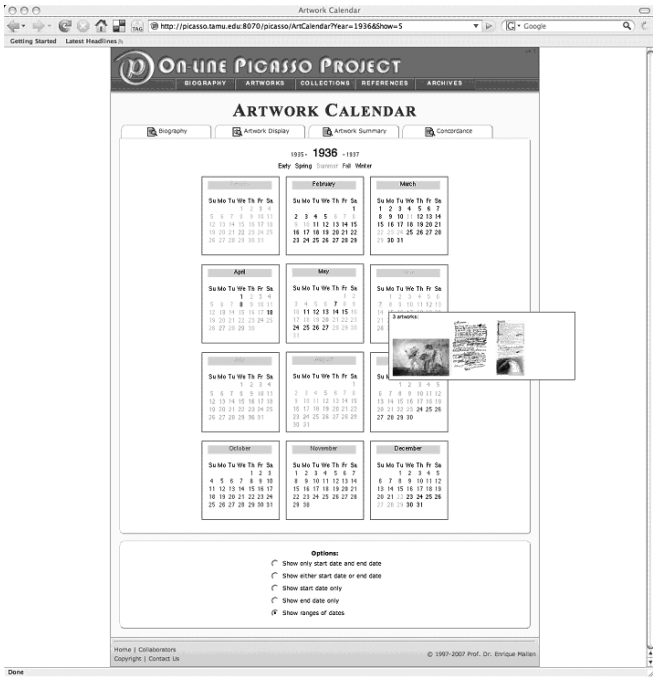


Figure 1: Artwork surrogates – ranges of dates

The use of calendar interfaces provides new possibilities for scholars and users in general: the discovery of relationships between documents, which standard HTML interfaces do not facilitate. The main advantages derived from their use include:

1. The possibility of visualizing an entire year in Picasso's biography and artistic career.

Through the use of a Calendar-based interface, artworks can be visually identified to their specific dates of creation. This provides a visualization mechanism that allows the user to navigate through a potentially large number of artworks in one screen. The number of artworks that can be accessed, depends on how esthetically prolific the artist was in that specific year.

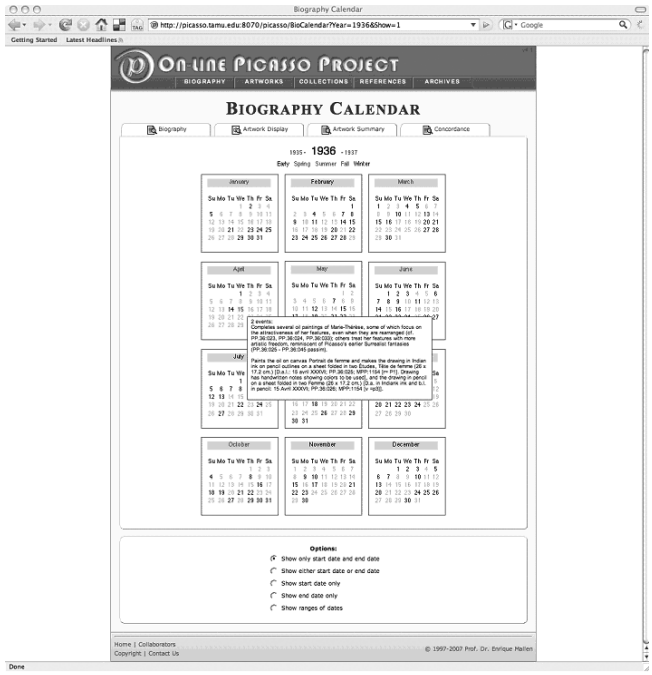


Figure 2: Event surrogates – start and end dates

Additionally, this interface allows the users of the project to quickly determine periods where the artist was more prolific. Dates where Picasso was "active" are clearly visible and identifiable. This data stratification gives users an additional layer of information.

For the case of biographical events, a similar scenario is created. Users can navigate through an entire year of events, and the information is presented in a way that affords quick navigation and encourages interaction. Visually, periods where more events occurred in Picasso's life are easily identifiable.

2. The possibility of moving to specific day, month or season within a year in one single interaction with the interface.

Through the use of information surrogates, users have the possibility of moving to a specific day, month or season within a year with a single click. The actions produced by scrolling through multiple screens are eliminated, and users can view the artifacts produced on a specific date with ease. Consequently, comparisons between artworks can be achieved fluidly due to the enhancements in the browsing environment. Similarly, users can read about the specific events in Picasso's biography by visually selecting concrete dates.

The deployment of the Calendar interfaces, produce visualizations that enable scholars to determine time periods or specific dates in Picasso's life. They serve as a tool that helps identify when the artist was more prolific, as well the opposite: when less artworks were produced. This analysis could also be accomplished by browsing through the artwork collection but it requires additional interaction from the user, which at the end equals more time and effort. The interfaces also constitute a new way of browsing the document collection:

visually and in strict correlation to time. They also facilitate further exploration of artworks and events in certain days, months, seasons and years.

Timelines

This new browsing mechanism in the OPP, based on the Simile Timeline, was introduced by placing artworks as markers in a time frame. It was designed to allow users to examine the artworks produced, along with the recorded events of a given year. Further modifications were necessary, since the Simile Timeline was designed to support single events occurring in one day.

Initially, the timeline was designed to focus only on Picasso's artworks. This design choice gave users great freedom to explore large amounts of information in a manipulatable visual space. However, the biographical events were being excluded. These events included in the OPP, are particularly important since provide a historical framework, which is crucial to the understanding of the artist's legacy and are tightly bound to his work rhythm.

Moreover, some of the artworks produced by Picasso have a certain degree of uncertainty in their dates of creation, since their start and end dates were not documented. The timelines provide a mechanism for dealing with uncertainty, where the artworks are represented with a time bar with a lower level of saturation in their color. This gives a visual clue that the start and end dates are not fixed, and are subject to speculation.

Additional information such as changes in style and location were injected to the timeline, which were extracted from the artist's biography. Their purpose is to provide an additional layer of information that can be used to interpret the events that lead to the creation and mood of certain artifacts, and thus enhancing the browsing environment.

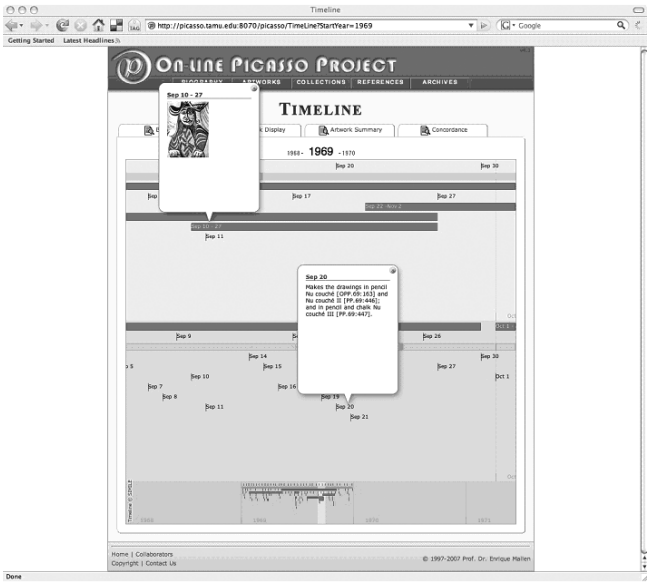


Figure 3: Exploring artwork - event correlation

The advantages gained through the use of timelines include:

1. The possibility of grasping visually time-extensions in Picasso's output.

Picasso worked on several artworks at times, which share a similar theme. Even though they share common characteristics, they are not identical. Each of these artworks has variations, which differentiate them.

On the other hand, the timelines allow users to freely explore all the artworks and events within a given year, and point out their similarities and differences, and affords further examination regarding the evolution of a common and shared theme.

2. The possibility of visually comparing works ordered in chronological order.

The timelines provide a mechanism that filters artworks according to their year of creation. The enhanced navigational scheme provided, allows scholars to view artifacts in chronological order. The addition of surrogates allows users to point out a specific item, and then compare them in relation to others through and their time correlation.

3. The possibility of seeing correlations between change of location and artwork creation.

The deployed timelines allow the exploration of correlations between location changes and the creation of specific artworks. Changes in location are marked in the timelines, and clearly denote a point in time where exposure to a new or recurring context occurred.

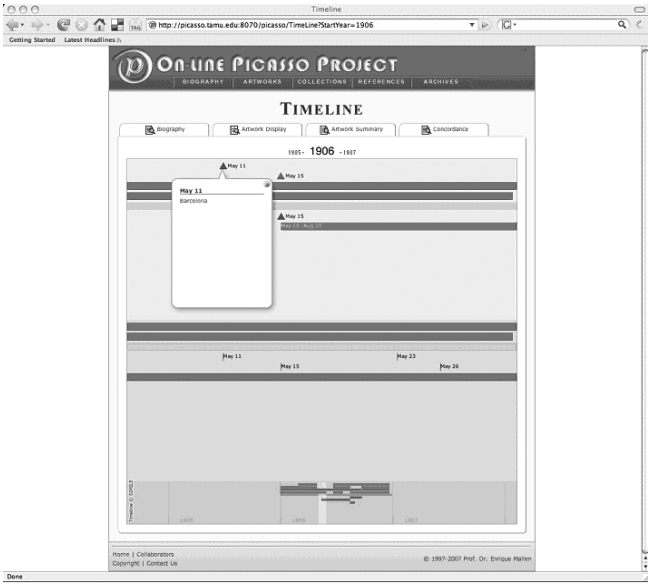


Figure 4: Changes in location

4. The possibility of comparing different stylistic periods as they relate to concrete artworks and specific locations.

The timelines produce a visualization that puts changes in thematic periods and in locations in a common context, along with the artworks that were elaborated in that time span. This tool is augmented with the navigational ease of clicking through a series of artworks, to compare their characteristics and perform a deeper analysis if necessary.

The interfaces have been deployed taking into account that additional functionality could be introduced with ease. As a consequence, information regarding Picasso's writings and poems will be included into the next iteration of the timelines and calendars. This will allow a deeper understanding of his legacy, since it could potentially provide a greater understanding of his artworks and biography. The writings and poems constitute a compendium of his thoughts and insights, extremely valuable because they were written by the artist himself.

The timeline interfaces in the OPP, narrow the gap and visually correlate biographical entries with artworks. They provide scholars a bigger picture of Picasso's artistic landscape and how events they could have affected his artworks. The dynamic nature of the web-accessible interfaces facilitate the insertion of new documents and metadata and thus altering the graphical space, which is not feasible on static and printed editions.

References

The Online Picasso Project. Available at http://picasso.tamu.edu. [Accessed May 17, 2007]

SIMILE Project. Available at http://simile.mit.edu/. [Accessed on May 24, 2007]

Monroy, R. Furuta, and E. Mallen, "Visualizing and Exploring Picasso's World," in *Proceedings of the 3rd ACM/IEEE-CS joint conference on Digital Libraries*, 2003, IEEE Computer Society, pp. 173- 175.

Kumar, R. Furuta, and R. Allen, "Metadata visualization for digital libraries: interactive timeline editing and review," in *Proceedings of the third ACM conference on Digital Libraries,* 1998, ACM Press, pp. 126 – 133.

Monroy, R. Furuta, E. Urbina, and E. Mallen, "Texts, Images, Knowledge: Visualizing Cervantes and Picasso," in *Proceedings of the Visual Knowledges Conference,* 2003, John Frow, www.iash.ed.ac.uk/vkpublication/monroy.pdf.

N. Audenaert, U. Karadkar, E. Mallen, R. Furuta, and S. Tonner, "Viewing Texts: An Art-centered Representation of Picasso's Writings," In *Proceedings of Digital Humanities 2007*, 2007.

C. Monroy, R. Furuta, and E. Mallen, "Creating, Visualizing, Analyzing, and Comparing Series of Artworks," 2003, Unpublished paper.

Computer Assisted Conceptual Analysis of Text: the Concept of Mind in the Collected Papers of C.S. Peirce

Jean-Guy Meunier
meunier.jean-guy@uqam.ca
Université du Quebec à Montréal, Canada

Dominic Forest
dominic.forest@umontreal.ca
Université de Montréal, Canada

Computer assisted reading and analysis of text (CARAT) has recently explored many variants of what has become fashionable to call "text mining" strategies. Text mining strategies are theoretically robust on large corpus. However, since they mainly operate at a macro textual level, their use is still the object of resistance by the expert readers that aim at fine and minute conceptual analysis. In this paper, we present a computer assisted strategy for assisting conceptual analysis based on automatic classification and annotation strategies. We also report on experiment using this strategy on a small philosophical corpus.

Conceptual analysis is an expert interpretation methodology for the systematic exploration of semantic and inferential properties of set of predicates expressing a particular concept in a text or in a discourse (Desclés, 1997; Fodor, 1998; Brandom, 1994; Gardenfors, 2000; Rastier, 2005). Computer assisted reading and analysis of text (CARAT) is the computer assistance of this conceptual analysis.

The strategy of CACAT

Our text analysis strategy rests on the following main hypothesis:

> *The expression of a canonical concept in a text presents linguistics regularities some of which can be identified using classification algorithms*

This hypothesis itself unwraps into three sub hypothesis:

> *Hypothesis 1: conceptual analysis can be realized by the contextual exploration of the canonical forms of a concept*

This is realized through the classical concordance strategy and variances on a pivotal term and its linguistic variants (e.g. mind, mental, mentally, etc.) (Pincemin *et al.,* 2006; McCarthy, 2004; Rockwell, 2003).

Hypothesis 2: the exploration of the contexts of a concept is itself realized through some mathematical classification strategy.

This second hypothesis postulates that contexts of a concept present regularities that can be identified by mathematical clustering techniques that rest upon similarities found among contextual segments (Jain *et al.* 1999; Manning and Schütze, 1999).

Hypothesis 3: Classes of conceptual of similar conceptual contexts can be annotated so as to categorize their semantic content.

This last hypothesis allows to associate to each segment of a class of contexts some formal description of their content be it semantic, logical, pragmatic, rhetorical, etc. (Rastier *et al.,* 2005; Djioua and Desclés, 2007; Meyers, 2005; Palmer *et al.*, 2005; Teich *et al.*, 2006). Some of these annotations can be realized through algorithms; others can only be done manually.

Experiment

From these three hypotheses emerges an experiment which unwraps in five phases. This experiment was accomplished using C.S. Peirce's *Collected Papers* (volumes I-VIII) (Peirce, 1931, 1935, 1958). More specifically, this research aimed at assisting conceptual analysis of the concept of "Mind" in Peirce's writings.

Phase 1: Text preparation

In this methodology, the first phase is text pre-processing. The aim of this first phase is to transform the initial corpus according to phases 2 and 3 requirements. Various operations of selection, cleaning, tokenisation, and segmentation are applied. In the experiment we report on, no lemmatisation or stemming was used. The corpus so prepared was composed of 74 450 words (tokens) with a lexicon of 2 831 word types.

Phase 2: Key Word In Context (KWIC) extraction (concordance)

Using the corpus pre-processed in phase 1, a concordance is made with the pivotal word "Mind". The KWIC algorithm generated 1 798 contextual segments of an average of 7 lines each. In order to be able to manually evaluate the results of the computer-assisted conceptual analysis, we decided to select in the project only a random sampling of the 1 798 contextual segments. The sampling algorithm delivered 717 contextual segments. This sample is composed of 3 071 words (tokens) and 1 527 type words.

Phase 3: KWIC clustering

The concordance is in itself a subtext (of the initial corpus). A clustering technique was applied to the concordance results. In this project, a hierarchical agglomerative clustering algorithm was applied. It generated 83 clusters with a mean 8.3 segments per class. It is possible to represent spatially the set of words in each class. Figure 1 illustrates such a regrouping for cluster 1.

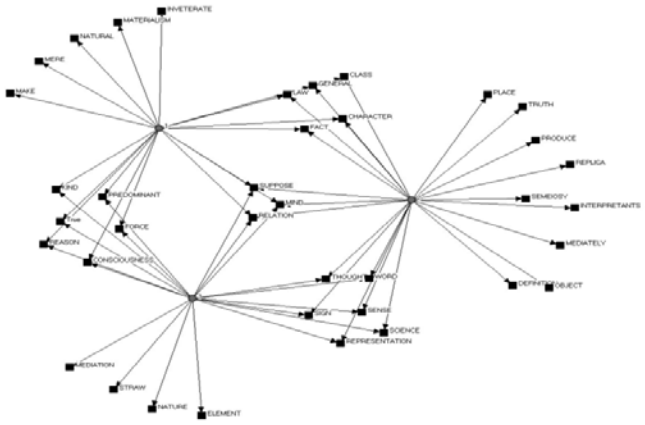


Figure 1. Graphical representation of cluster 1 lexicon.

It is often on this type of representation that many numerical analyses start their interpretation. One traditional critic presented by expert analysts is their great generality and ambiguity. This kind of analysis and representation give hints on the content of documents, but as such it is difficult to use for fine grained conceptual analysis. It must hence be refined. It is here that the annotation phase comes into play.

Phase 4: Annotation

The annotation phase allows the expert reader to make more explicit the type of information contained in each clusters (generated in phase 3). For instance, the interpreter may indicate if each cluster is a THEME, a DEFINITION, a DESCRIPTION, an EXPLANATION, an ILLUSTRATION, an INFERENCE, or what is it MODALITY (epistemic, epistemological, etc.). The variety of annotation types is in itself a research object and depends on various textual and linguistic theories.

Annotation results

In this abstract, size constraints do not allow us here to present detailed results of classification and annotation processes. We shall only present a sample on a few segments of three classes.

Annotations of cluster 1: The first cluster contained 17 segments all of which have received an annotation. Here are samples of annotation for two segments of cluster 1. The annotation is preceded by the citation itself from the original text.

[SEGMENT NO 512]

"Finally laws of mind divide themselves into laws of the universal action of mind and laws of kinds of psychical manifestation."

ANNOTATION: DEFINITION: the law of mind is a general action of the mind and a psychological manifestation

[SEGMENT NO 1457]

"But it differs essentially from materialism, in that, instead of supposing mind to be governed by blind mechanical law, it supposes the one original law to be the recognized law of mind, the law of association, of which the laws of matter are regarded as mere special results."

ANNOTATION: EXPLICATION: The law of mind is not a mechanical materialism.

Phase 5: Interpretation

The last phase is the interpretative reading of the annotations. Here, the interpreter situates the annotated segments into his own interpretative world. He may regroup the various types of annotation (DEFINITIONS, EXPLANTIONS, etc.) and hence build a specific personal data structure on what he has annotated. From then on, he may rephrase these in his own language and style but most of all situate them in some theoretical, historical, analytical, hermeneutic, epistemological, etc. perspective. It is the moment where the interpreter generates his synthesis of the structure he believes underlies the concept.

We present here a sample of the synthesis of conceptual analysis assisted by the CARAT process on cluster 1 (the concept of "mind" in C.S. Peirce's writings – cluster 1).

The law of Mind: association

The Peircian theory of MIND postulates that a mind is governed by laws. One of these laws, a fundamental one, is associative (segment 512). This law describes a habitus acquired by the mind when it functions (segment 436).

Association is connectivity

This functioning is one of relation building through connections. The connectivity is of a specific nature. It realizes a synthesis (à la Kant) which is a form of "intellectual" generalisation (segment 507).

It is physically realized

Such a law is also found in the biological world. It is a law that can be understood as accommodation (segment 1436). In fact, this law is the specific form of the Mind's dynamic. It is a fundamental law. But it is not easy for us to observe it because we are victim of a interpretative tradition (segment 1330) that understands the laws of mind as laws of nature. This is a typical characteristic of an "objective idealism" (segments 1762 and 1382). The laws of mind do not belong to mechanist materialism (segments 90 and 1382).

And there exist a variety of categories

There exist subdivisions of this law. They are related to the generalisation process that is realised in infanthood, education, and experience. They are intimately related to the growth of consciousness (segments 375 and 325).

Conclusion

This research project explores a Computer-Assisted Reading and Analysis of Text (CARAT) methodology. The classification and annotation strategies manage to regroup systematically segments of text that present some content regularity. This allows the interpreter to focus directly on the organized content of the concept under study. It helps reveal its various dimensions (definitions, illustrations, explanations, inferences, etc.).

Still, this research is ongoing. More linguistic transformations should be applied so as to find synonymic expressions of a concept. Also, various types of summarization, extraction and formal representation of the regularities of each class are to be explored in the future.

But the results obtained so far reinstate the pertinence of the concordance as a tool for conceptual analysis. But it situates it in a mathematical surrounding that aim at unveiling the various dimensions of a conceptual structure. Most of all, we believe that this methodology may possibly interest expert readers and analysis for it gives a strong handle and control on their interpretation process although assisting them throughout the process.

References

Brandom, R.B. (1994). *Making it Explicit.* Cambridge: Harvard University Press.

Desclés, Jean-Pierre (1997) "Schèmes, notions, predicats et termes". *Logique, discours et pensée, Mélanges offerts à Jean-Blaize Grize,* Peter Lang, 9-36. 47.

Djioua B. and Desclés, J.P. (2007), "Indexing Documents by Discourse and Semantic Contents from Automatic Annotations of Texts", FLAIRS 2007, Special Track "Automatic Annotation and Information Retrieval : New Perspectives", Key West, Florida, May 9-11.

Fodor, J. (1998) *Concepts: Where Cognitive Science Went Wrong*. Oxford: OUP.

Gardenfors, P. (2000) *Conceptual Spaces*. Cambridge (Mass.): MIT Press.

Jain, et al. (1999). Data Clustering: A Review. *ACM Computing Surveys*, 31(3):264–323.

Manning, C. and Schutze, H. (1999) *Foundations of Statistical Natural Language Processing*, Cambridge Mass. : MIT Press.

Meyers, Adam (2005) Introduction to Frontiers in CorpusAnnotation II Pie in the Sky *Proceedings of the Workshop on Frontiers in Corpus Annotation II: Pie in the Sky*, pages 1–4, New York University Ann Arbor, June 2005.

McCarthy, W. (2004) *Humanities Computing*, Palgrave MacMillan Blackwell Publishers.

Palmer, M., Kingsbury, P., Gildea, D. (2005) "The Proposition Bank: An Annotated Corpus of Semantic Roles". *Computational Linguistics*, vol. 31, no 1, pp. 71-106.

Peirce, C.S. (1931-1935, 1958), *Collected Papers of Charles Sanders Peirce, vols. 1–6*, Charles Hartshorne and Paul Weiss (eds.), vols. 7–8, Arthur W. Burks (ed.), Harvard University Press, Cambridge, MA, 1931–1935, 1958.

Pincemin, B. et al. (2006). Concordanciers: thème et variations, in J.-M. VIPREY (éd.), *Proc. of JADT 2006*, pp. 773-784.

Rastier, F. (2005) Pour une sémantique des textes théoriques. *Revue de sémantique et de pragmatique*, 17, 2005, pp. 151-180.

Rastier, F. et al. (eds) (1995) *L'analyse thématique des données textuelles: l'exemple des sentiments*. Paris: Didier Érudition.

Rockwell, G. (2003) What is text analysis, really? *Literary and Linguistic Computing*, 18(2): 209–219.

Topic Maps and Entity Authority Records: an Effective Cyber Infrastructure for Digital Humanities

Jamie Norrish
Jamie.Norrish@vuw.ac.nz
New Zealand Electronic Text Centre, New Zealand

Alison Stevenson
alison.stevenson@vuw.ac.nz
New Zealand Electronic Text Centre, New Zealand

The implicit connections and cross-references between and within texts, which occur in all print collections, can be made explicit in a collection of electronic texts. Correctly encoded and exposed they create a framework to support resource discovery and navigation by following links between topics. This framework provides opportunities to visualise dense points of interconnection and, deployed across otherwise separate collections, can reveal unforeseen networks and associations. Thus approached, the creation and online delivery of digital texts moves from a digital library model with its goal as the provision of access, to a digital humanities model directed towards the innovative use of information technologies to derive new knowledge from our cultural inheritance.

Using this approach the New Zealand Electronic Text Centre (NZETC) has developed a delivery system for its collection of over 2500 New Zealand and Pacic Island texts using TEI XML, the ISO Topic Map technology[1] and innovative entity authority management. Like a simple back-of-book index but on a much grander scale, a topic map aggregates information to provide binding points from which everything that is known about a given subject can be reached. The ontology which structures the relationships between dierent types of topics is based on the CIDOC Conceptual Reference Model[2] and can therefore accommodate a wide range of types. To date the NZETC Topic Map has included only those topics and relationships which are simple, veriable and object based. Topics currently represent authors and publishers, texts and images, as well as people and places mentioned or depicted in those texts and images. This has proved successful in presenting the collection as a resource for research, but work is now underway to expand the structured mark-up embedded in texts to encode scholarly thinking about a set of resources. Topic-based navigable linkages between texts will include 'allusions' and 'influence' (both of one text upon another and of an abstract idea upon a corpus, text, or fragment of text).[3]

Importantly, the topic map extends beyond the NZETC collection to incorporate relevant external resources which expose structured metadata about entities in their collection (see Figure 1).

Cross-collection linkages are particularly valuable where they reveal interdisciplinary connections which can provide fertile ground for analysis. For example the National Library of New Zealand hosts a full text archive of the Transactions and Proceedings of the Royal Society containing New Zealand science writing 1868-1961. By linking people topics in the NZETC collection to articles authored in the Royal Society collection it is possible to discern an interesting overlap between the 19th century community of New Zealand Pakeha artists and early colonial geologists and botanists.

In order to achieve this interlinking, between collections, and across institutional and disciplinary boundaries, every topic must be uniquely and correctly identied. In a large, full text collection the same name may refer to multiple entities,[4] while a single entity may be known by many names.[5] When working across collections it is necessary to be able to condently identify an individual in a variety of contexts. Authority control is consequently of the utmost importance in preventing confusion and chaos.

The library world has of course long worked with authority control systems, but the model underlying most such systems is inadequate for a digital world. Often the identier for an entity is neither persistent nor unique, and a single name or form of a name is unnecessarily privileged (indeed, stands in as the entity itself). In order to accommodate our goals for the site, the NZETC created the Entity Authority Tool Set (EATS),[6] an authority control system that provides unique, persistent, sharable[7] identiers for any sort of entity. The system has two particular benets in regards to the needs of digital humanities researchers for what the ACLS described as a robust cyber infrastructure.[8]

Firstly, EATS enables automatic processing of names within textual material. When dealing with a large collection, resource constraints typically do not permit manual processing -- for example, marking up every name with a pointer to the correct record in the authority list, or simply recognising text strings as names to begin with. To make this process at least semi-automated, EATS stores names broken down (as much as possible) into component parts. By keeping track of language and script information associated with the names, the system is able to use multiple sets of rules to know how to properly glue these parts together into valid name forms. So, for example, William Herbert Ellery Gilbert might be referred to in a text by "William Gilbert", "W. H. E. Gilbert", "Gilbert, Wm.", or a number of other forms; all of these can be automatically recognised due to the language and script rules associated with the system. Similarly Chiang Kai-shek, being a Chinese name, should be presented with the family name first, and, when written in Chinese script, without a space between the name parts (蒋介石).

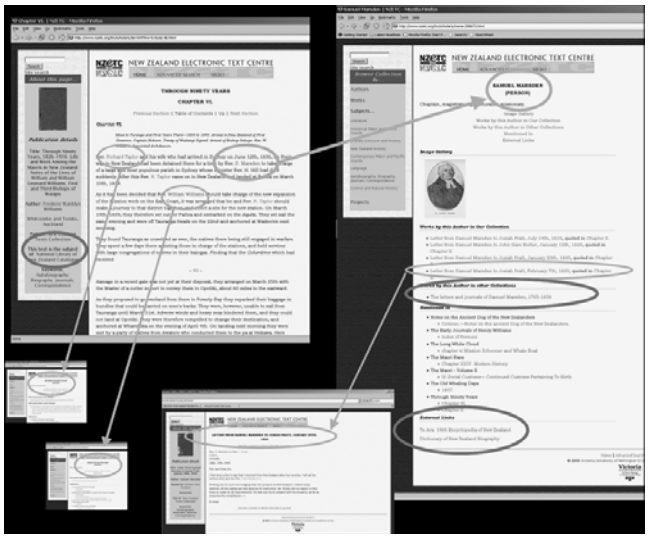


Figure 1: A mention of Samuel Marsden in a given text is linked to a topic page for Marsden which in turn provides links to other texts which mention him, external resources about him and to the full text of works that he has authored both in the NZETC collection and in other online collections entirely separate from the NZETC.

The ability to identify entities within plain text and add structured, machine-readable mark-up contributes to the growth of electronic text corpora suitable for the types of computational analysis oered by projects such as the MONK environment.9 This is, however, distinct from the problem of identifying substrings within a text that might be names, but that are not found within EATS. This problem, though signicant, does not fall within the main scope of the EATS system.[10] Similarly, disambiguating multiple matches for the same name is generally best left to the determination of a human being: even date matches are too often problematic.[11]

Secondly, the system is built around the need to allow for an entity to carry sometimes conflicting, or merely dierent, information from multiple sources, and to reference those sources.[12] Having information from multiple sources aids in the process of disambiguating entities with the same names; just as important is being able to link out to other relevant resources. For example, our topic page for William Colenso links not only to works in the NZETC collection, but also to works in other collections, where the information on those other collections is part of the EATS record.

It is, however, barely sucient to link in this way directly from one project to another. EATS, being a web application, can itself be exposed to the net and act as a central hub for information and resources pertaining to the entities within the system. Since all properties of an entity are made as assertions by an organisation, EATS allows multiple such organisations to use and modify records without touching anyone else's data; adding data harvesting to the mix allows for centralisation of information (and, particularly, pointers to further information) without requiring much organisational centralisation.

One benefit of this approach is handling entities about which there is substantial dierence of view. With topics derived from research (such as ideas and events) there are likely to be dierences of opinion as to both the identification of entities and the relationships between them. For example one organisation may see one event where another sees two. To be able to model this as three entities, with relationships between them asserted by the organisations, a potentially confusing situation becomes clear, without any group having to give up its own view of the world. The EATS system can achieve this because all information about an entity is in the form of a property assertion made by a particular authority in a particular record (see figure 2).

The technologies developed and deployed by the NZETC including EATS are all based on open standards. The tools and frameworks that have been created are designed to provide durable resources to meet the needs of the academic and wider community in that they promote interlinking between digital collections and projects and are themselves interoperable with other standards-based programs and applications including web-based references tools, eResearch virtual spaces and institutional repositories.

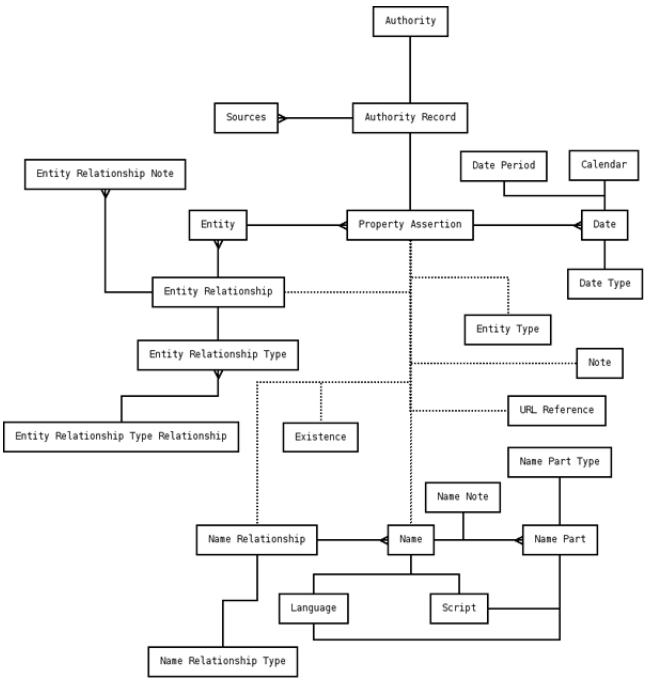


Figure 2: The EATS objects and basic relationships

Only once both cultural heritage institutions and digital humanities projects adopt suitable entity identiers and participate in a shared mapping system such as EATS, can there exist unambiguous discovery of both individual resources and connections between them. The wider adoption of this type of entity authority system will contribute substantially to the creation of the robust cyber infrastructure that will, in the words of the ACLS "allow digital scholarship to be cumulative, collaborative, and synergistic."

Notes

1 For further information see Conal Tuhoy's \Topic Maps and TEI | Using Topic Maps as a Tool for Presenting TEI Documents" (2006) http://hdl.handle.net/10063/160.

2 The CIDOC CRM is an ISO standard (ISO 21127:2006) which provides denitions and a formal structure for describing the implicit and explicit concepts and relationships used in cultural heritage documentation. For more information see http://cidoc.ics.forth.gr/.

3 The initial project work is being undertaken in collaboration with Dr Brian Opie from Victoria University of Wellington and is centred around the work and influences of William Golder, the author of the first volume of poetry printed and published in New Zealand.

4 For example, the name Te Heuheu is used in a number of texts to refer to multiple people who have it as part of their full name.

5 For example the author Iris Guiver Wilkinson wrote under the pseudonym Robin Hyde.

6 For more analysis of the weakness of current Library standards for authority control and for more detail information on EATS see Jamie Norrish's \EATS: an entity authority tool set" (2007) at http://www.nzetc.org/downloads/eats.pdf.

7 Being a web-based application the identiers are also dereferencable (ie resolve a web resource about the entity) and therefore can be used as a resource by any web project.

8 "Our Cultural Commonwealth: The final report of the American Council of Learned Societies Commission on Cyberinfrastructure for the Humanities & Social Sciences" (2006) http://www.acls.org/cyberinfrastructure/OurCulturalCommonwealth.pdf.

9 Metadata Oer New Knowledge http://www.monkproject.org/.

10 EATS can be provided with name data on which to make various judgements (such as non-obvious abbreviations like Wm for William), and it would be trivial to get a list of individual parts of names from the system, for identication purposes, but there is no code for actually performing this process.

11 That said, the NZETC has had some success with automated ltering of duplicate matches into dierent categories, based on name and date similarity (not equality); publication dates provide a useful cut-o point for matching.

12 A source may be a primary text or an institution's authority record.

2D and 3D Visualization of Stance in Popular Fiction

Lisa Lena Opas-Hänninen
lisa.lena.opas-hanninen@oulu.fi
University of Oulu, Finland

Tapio Seppänen
tapio@ee.oulu.fi
University of Oulu, Finland

Mari Karsikas
mmantysa@mail.student.oulu.fi
University of Oulu, Finland

Suvi Tiinanen
stiinane@mail.student.oulu.fi
University of Oulu, Finland

Introduction

In analyzing literary style through statistical methods, we often show our results by plotting graphs of various types, e.g. scatterplots, histograms, and dendrograms. While these help us to understand our results, they always necessarily "flatten" the data into a 2- dimensional format. For the purposes of visualization of complex data, we might be better off trying to look at it three dimensionally if a suitable vector representation of the data is found first. This paper investigates stance in three groups of popular fiction, namely romance fiction, detective fiction written by female authors and detective fiction written by male authors, and presents the results using both 2 dimensional and three dimensional visualization.

Romance fiction and detective novels are both characterized by the fact that they all have the same basic story. In romance fiction it is the story of how the hero and heroine meet, how their relationship develops and how the happy ending is achieved. In detective fiction it is the story of the murder, the unraveling of the case, the misleading clues and the final solution. The same story is being retold over and over again, just as in the oral storytelling tradition (Radway 1984:198). The reader is not left in suspense of the final outcome and each story is different from the others in the details of the events and the way the story is told. Through this the reader becomes involved in the story and partakes in the emotions, attitudes and thoughts of the protagonist. These feelings, emotions and moods are marked by syntactic and semantic features often referred to as markers of stance.

Stance refers to the expression of attitude and consists of two different types of expressions of attitude: evidentiality and affect (Biber and Finegan 1989). Evidentiality means that the reader becomes privy to the speaker's attitudes towards whatever knowledge the speaker has, the reliability of that knowledge and how the speaker came about that knowledge. Affect refers to the personal attitudes of the speaker, i.e. his/her emotions, feelings, moods, etc. Biber and Finegan (1989) investigated 12 different categories of features deemed to mark stance: certainty/doubt adverbs, certainty/doubt verbs, certainty/doubt adjectives, affective expressions, hedges, emphatics and necessity/possibility/predictive modals. They showed that different text types are likely to express stance in different ways. Opas and Tweedie (1999) studied stance in romance fiction and showed that three types of romance fiction can be separated by their expression of stance. This paper continues these studies, paying special attention to the visualization of the results.

Materials and methods

Our total corpus is 760 000 words. It consists of three parts: romance fiction, female-authored detective fiction and male-authored detective fiction, all published in the 1990s. The romance fiction comprises a total of 240 000 words from the Harlequin Presents series, the Regency Romance series and Danielle Steel's works, which are often classified as women's fiction or 'cross-overs'. The female-authored detective fiction part of our corpus includes Patricia Cornwell, Sue Grafton, P.D. James, Donna Leon, Ellis Peters, and Ruth Rendell. These texts make up 295 000 words. The rest of the corpus (229 000 words) is made up of male-authored detective fiction texts, including Colin Dexter, Michael Dibdin, Quintin Jardine, Ian Rankin and Peter Tremayne.

Principal components analysis was used to reduce the 12 markers of stance to three dimensions which describe 52.4% of the variation in the data.

Results

In a previous study Opas and Tweedie (2000) concluded that the detective stories seem to mark evidentiality, i.e. the characters express their certainty and doubt. The romance stories, on the other hand, seem to mainly mark affect, i.e. the characters express their emotions and moods. In Biber and Finegan's terms (1989), the detective fiction texts show 'interactional evidentiality' and the romance fiction texts show 'emphatic expression of affect'.

The results in Figures 1 and 2 below show the female-authored detective stories in shades of red and yellow, the male-authored detective stories in shades of blue and the romance stories in shades of black. While in Figure 1 the female-authored texts are perhaps slightly more clearly separatable from the others, it still seems that all these texts are far more overlapping and less clearly distinguishable as three groups than the texts in previous studies were. However, broadly speaking, the romance texts are in the lower half of the graph, the female-authored detective texts in the upper half and the male-authored detective texts in the middle. What is surprising though is that no features seem to "pulling" texts downwards, towards the lower half of the graph; and that the feature that "pull" the texts upwards include both markers of certainty/doubt and affect.

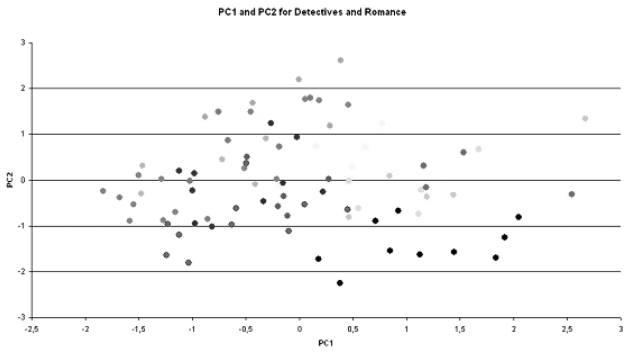


Figure 1. PC1 and PC2 for detective and romance fiction

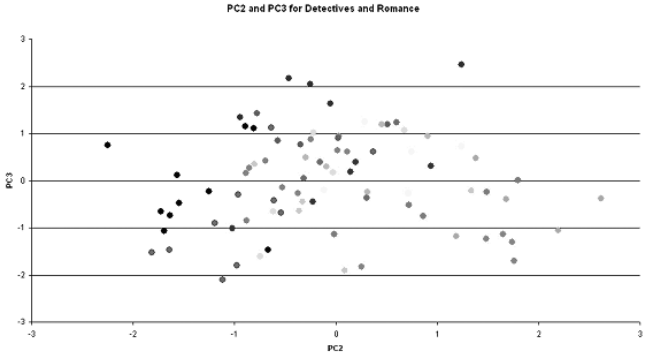


Figure 2. PC2 and PC3 for detective and romance fiction

Figure 2 seems to show similar results. Here, perhaps, the female-authored detective texts are even slightly more easily separated from the others, but the general impression of the male-authored texts and the romance texts overlapping remains. Yet again it seems that there are hardly any features accounting for the texts on the left-hand side of the graph and that the feature "pulling" the texts to the right include both features marking evidentiality and those marking affect.

These results are quite surprising. Either male-authored detective stories mark evidentiality and affect in the same manner as romance fiction, and here is seems that they don't show many features that would mark stance, or there is something more complex at work here. To help us understand these phenomena better, we would suggest visualizing the workings of the markers of stance in a 3 dimensional model. To this end, we have built a tool that takes the principal components analysis data, reduces dimensionality down to three components with the most energy and presents the data with these components. The software tool is implemented in the MATLAB® environment (The MathWorks, Inc., Massachusetts, USA) utilizing its 3D graphical functions. The tool is an interactive one, allowing the researcher to turn the 3D model and look for the angles that best show the clustering structure and the differences between the texts. We will demonstrate how tools such as this one significantly improve the researcher's ability to visualize the research results and to interpret them.

References

Biber, D. and E. Finegan (1989) Styles of Stance in English: Lexical and Grammatical Marking of Evidentiality and Affect. *Text* 9.1: 93-124.

Opas, L.L. and F.J. Tweedie (1999) The Magic Carpet Ride: Reader Involvement in Romantic Fiction. *Literary and Linguistic Computing* 14.1: 89-101.

Opas, L.L. and F.J. Tweedie (2000) Come into My World: Styles of stance in detective and romantic fiction. Poster. ALLC-ACH2000. Glasgow, UK.

Radway, J.A. (1984) *Reading the Romance*. Chapel-Hill: University of North Carolina Press.

TEI Analytics: a TEI Format for Cross-collection Text Analysis

Stephen Ramsay
sramsay@unlserve.unl.edu
University of Nebraska-Lincoln, USA

Brian Pytlik-Zillig
bpytlikz@unlnotes.unl.edu
University of Nebraska-Lincoln, USA

The Monk Project (http://www.monkproject.org/) has for the last year been developing a Web-based system for undertaking text analysis and visualization with large, full-text literary archives. The primary motivation behind the development of this system has been the profusion of sites like the Brown Women Writers Project, Perseus Digital Library, Wright American Fiction, Early American Fiction, and Documenting the American South, along with the vast literary text corpora oered by organizations like the Text Creation Partnership and Chadwyck-Healey. Every one of these collections represents fertile ground for text analysis. But if they could be somehow combined, they would constitute something considerably more powerful: a literary full text-corpus so immense as to be larger than anything that has yet been created in the history of computing.

The obstacles standing in the way of such a corpus are well known. While all of the collections mentioned above are encoded in XML and most of them are TEI-conformant, local variations can be so profound as to prohibit anything but the most rudimentary form of cross-collection searching. Tagset inclusions (and exclusions), local extensions, and local tagging and metadata conventions dier so widely among archives, that it is nearly impossible to design a generalized system that can cross boundaries without a prohibitively cumbersome set of heuristics. Even with XML and TEI, not all texts are created equal.

TEI Analytics, a subset of TEI in which varying text collections can be expressed, grew out of our desire to make MONK work with extremely large literary text corpora of the sort that would allow computational study of literary and linguistic change over broad periods of time.

TEI Analytics

Local text collections vary not because archive maintainers are contemptuous toward standards or interoperability, but because particular local circumstances demand customization. The nature of the texts themselves may require specialization, or something about the storage, delivery, or rendering framework used may favor particular tags or particular structures. Local environments also require particular metadata conventions (even within the boundaries of the TEI header). This is in part why the TEI Consortium provides a number of pre-fabricated customizations, such as TEI Math and TEI Lite, as well as modules for Drama, transcriptions of speech, and descriptions of manuscripts. Roma (the successor to Pizza Chef) similarly allows one to create a TEI subset, which in turn may be extended for local circumstances.

TEI Analytics, which is itself a superset of TEI Tite, is designed with a slightly dierent purpose in mind. If one were creating a new literary text corpus for the purpose of undertaking text analytical work, it might make the most sense to begin with one of these customizations (using, perhaps, TEI Corpus). In the case of MONK, however, we are beginning with collections that have already been tagged using some version of TEI with local extensions. TEI Analytics is therefore designed to exploit common denominators in these texts while at the same time adding new structures for common analytical data structures (like part-of-speech tags, lemmatizations, named-entities, tokens, and sentence markers). The idea is to create a P5-compliant format that is designed not for rendering, but for analytical operations such as data mining, principle component analysis, word frequency study, and n-gram analysis. In the particular case of MONK, such documents have a relatively brief lifespan; once documents are converted, they are read in by a system that stores the information using a combination of object-relational database technology and binary indexing. But before that can happen, the texts themselves need to be analyzed and re-expressed in the new format.

Implementation

Our basic approach to the problem involves schema harvesting. The TEI Consortium's Roma tool (http://tei.oucs.ox.ac.uk/Roma/) was first used to create a base W3C XML schema for TEI P5 documents, which we then extended using a custom ODD file.

With this basis in place, we were able to create an XSLT "meta-stylesheet" (MonkMetaStylesheet.xsl) that consults the target collection's W3C XML schema to determine the form into which the TEI P4 les should be converted. This initial XSLT stylesheet is a meta-stylesheet in the sense that it programatically authors another XSLT stylesheet. This second stylesheet (XMLtoMonkXML.xsl), which is usually thousands of lines long, contains the conversion instructions to get from P4 to the TEI Analytics's custom P5 implementation. Elements that are not needed for analysis are removed or re-named according to the requirements of MONK (for example, numbered s are replaced with un-numbered s). Bibliographical information is critical for text analysis, and both copyright and responsibility information must be maintained, but much of the information contained in the average <teiHeader> (like revision histories and records of workflow) are not relevant to the task. For this reason, TEI Analytics uses a radically simplied form of the TEI header.

Here is a sample template in the meta-stylesheet (in a somewhat abbreviated form):

```
<xsl:template match="xs:element[@
name=$listOfAllowableElements/*]">

<!-- elements are identified in the
MONK shema, and narrowed to list
of allowable elements-->

<xsl:element name="xsl:template">
<!-- begins writing of 'xsl:template'
elements in the final XMLtoMonkXML.xsl
stylesheet -->

<xsl:attribute name="match">

<!-- begins writing of the (approximately
122 unique) match attributes on
the 'xsl:template' elements -->

<xsl:choose>
<xsl:when test="$attributeName
= $attributeNameLowercase">
<xsl:value-of select="@name"/>
</xsl:when>
<xsl:otherwise>
<xsl:value-of select="concat(@
name,' | ',lower-case(@name))"/>
</xsl:otherwise>
</xsl:choose>
</xsl:attribute>

<!-- ends writing of the match
attributes on the 'xsl:template'
elements -->

<xsl:element name="{$attributeName}">

<!-- writes the unique contents of
each 'xsl:template' element in the
XMLtoMonkXML.xsl stylesheet -->

<xsl:for-each select="$associ
atedAttributeList/list">
<xsl:choose>
<xsl:for-each select="child::
item[string-length(.) &gt; 0]">

<!-- all strings (in the dynamically-
generated list of associated
attributes) greater than
zero are processed -->

<xsl:when>
<xsl:attribute name="test">
<xsl:value-of select="concat('@',.)"/>
</xsl:attribute>
<xsl:copy-of>
<xsl:attribute name="select">
<xsl:value-of select="concat('@',.)"/>

<!-- copies the element's
attributes, constrained to a list
of attributes desired by MONK -->

</xsl:attribute>
</xsl:copy-of>
</xsl:when>
</xsl:for-each>
<xsl:otherwise> </xsl:otherwise>

<!-- any zero-length strings (in
the dynamically-generated list of
associated attributes) are discarded -->
</xsl:choose>
</xsl:for-each>
<xsl:apply-templates/>
</xsl:element>

<!-- ends writing of 'xsl:template'
elements in the final XMLtoMonkXML.xsl
stylesheet -->

</xsl:template>
```

All processes are initiated by a program (called Abbot) that performs, in order, the following tasks:

1. Generates the XMLtoMonkXML.xsl stylesheet

2. Edits the XMLtoMonkXML.xsl stylesheet to add the proper schema declarations in the output files.

3. Converts the entire P4 collection to MONK's custom P5 implementation.

4. Removes any stray namespace declarations from the output files, and

5. Parses the converted les against the MONK XML schema.

These steps are expressed in BPEL (Business Process Execution Language), and all source files are retained in the processing sequence so that the process can be tuned, adjusted, and re-run as needed without data loss. The main conversion process takes, depending on the hardware, approximately 30 minutes for roughly 1,000 novels and yields les that are then analyzed and tagged using Morphadorner (a morphological tagger developed by Phil Burns, a member of the MONK Project at

Northwestern University). Plans are underway for a plugin architecture that will allow one to use any of the popular taggers (such as GATE or OpenNLP) during the analysis stage.

Conclusion

We believe that TEI Analytics performs a useful niche function within the larger ecology of TEI by making disparate texts usable within a single text analysis framework. Even without the need for ingestion into a larger framework, TEI Analytics facilitates text analysis of disparate source files simply by creating a consistent and unied XML representation. We also believe that our particular approach to the problem of XML conversion (a small stylesheet capable of generating massive stylesheets through schema harvesting) may be useful in other contexts|including, perhaps, the need to convert texts from P4 to P5.

The Dictionary of Words in the Wild

Geoffrey Rockwell
georock@mcmaster.ca
McMaster University

Willard McCarty
willard.mccarty@kcl.ac.uk
King's College London, UK

Eleni Pantou-Kikkou
King's College London, UK

Introduction

The Dictionary of Words in the Wild [1] is an experiment in social textuality and the perceptual dynamics of reading. The Dictionary is a social image site where contributors can upload pictures of words taken "in the wild" and tag them so they are organized alphabetically as an online visual dictionary. Currently the Dictionary has 2227 images of 3198 unique words and 24 contributor accounts. The images uploaded and tagged are of text, usually single words or phrases, that appear in the everyday environment. Images uploaded include pictures of signs, body tatoos, garbage, posters, graffiti, labels, church displays, gravestones, plastic bags, clothing, art, labels, and other sights. The site is structured with an application programming interface to encourage unanticipated uses.

In this paper we will,

- Give a tour through the online social site and its API,
- Discuss current scholarship on public textuality and the perceptual dynamics of reading,
- Reflect on the Dictionary in these contexts, and
- Conclude with speculations on its possible contributions to our understanding of textuality and reading.

Outline of the Dictionary

The following is a narrative of the demonstration part of the presentation.

The Dictionary was developed with Ruby on Rails by Andrew MacDonald with support from the TAPoR project under the direction of Geoffrey Rockwell. Users can get a free account in order to start uploading images. (We discovered once the project was visible for a few months that spambots were automatically creating accounts so we have introduced a CAPTCHAlike graphical challenge-response feature to weed out false accounts.)

When you upload a picture you are given the opportunity to crop it and are prompted to provide a list of words that appear in the text. You can also provide a brief description or discussion of the word image.

Once uploaded the image is filed in a database and the interface allows you to access the images in different ways:

- You can click on a letter and browse images with words starting with that letter. An image with "On" and "Off" will be filed twice, though at present the label is the first word tagged.

- You can search for a word and see the images that have been tagged with that word.

- You can type in a phrase and get a sequence of images.

- You can see what comments have been left by others on your images and respond to them.

The API allows the user to create other processes or text toys that can query the dictionary and get back an XML file with the URLs for word images like:

```
<phrase>
  <word href="href_for_image">Word</word>
  <word>Word_with_no_image</word>
</phrase>
```

One of the goals of the project is to support mashups that use the dictionary for new projects. This is where the Dictionary is different than other such projects, many of which use Flickr to pull images of letters or words.

Theoretical Discussion

The Dictionary is an exploratory project designed to encourage the gathering of images of words in the wild and to provoke thinking about our encounters with these words. It did not start with an articulated theory of text that it set out to illustrate, nor in the contributors' experience does the actual collecting of words tend to be governed by this or that theory. Rather, in its simplicity, the Dictionary encourages participants to document public textuality as they encounter and perceive it.

At least in the initial phase of the project, the designers have imposed no rules or guidelines for the collecting, editing and tagging of words. Although it is clear that certain requirements for entering words into the Dictionary would make subsequent research questions easier to pursue, the designers prefer not to impose them so as to discover what in fact participants find to be interesting and practical to record. Because recording of words is voluntary and would seem inevitably to be limited to a few individuals, the time and effort required must be kept to a minimum in order to have a collection sufficiently large to allow the research potential of the Dictionary to emerge.

The Dictionary is meant to provoke reflection on the actual verbal environment in its totality, on the moment-by-moment encounter with individual words and phrases where one finds them and on the experience of reading them as each reading unfolds.

Collecting of verbal images for the Dictionary presents collectors with a way of defamiliarizing familiar environments. Conventional techniques for framing a photograph and digital tools for cropping it give them limited but surprisingly powerful means of recording defamiliarized sights. Additional means are provided by a commentary feature, but the amount of time required to compose this commentary tends to discourage much of it. Theoretical reflection on the encounter with words in the wild would seem to require participation in the project to be adequate to the data. For this reason collecting is not a task to be done separately from theorizing. Whatever theory is to emerge will come from participant observation.

In some cases, we have found, what appears interesting to record is relatively static, requiring little compositional care or subsequent cropping. In many cases, however, the experience of reading is a dynamic event, as when part of a verbal message, not necessarily first in normal reading order, is perceived first, then submerged into the overall syntax of its verbal and/or visual environment. In other cases, the experience may include a specific detail of the environment in which the text of interest is embedded but not be informed significantly by other details. Occasionally one has sufficient time to frame a photograph to capture the experience adequately, but most often photographs must be taken quickly, as when unwelcome attention would be drawn to the act of photography or the situation otherwise advises stealth (an inscription on a t-shirt, for example). The Dictionary, one might say, is a record of psycholinguistic events as much or more than simply of environmental data.

In more theoretical terms, is the project aims to study how language acts as a semiotic system materially placed in the real word. In order to interpret this multidimensional, "semiotic" role of language, our analysis focuses on how dictionary users perceive different signs and attribute meanings to words by referring to these signs. We will argue that through this kind of visual dictionary contributors can interact and play with language by using visual artifacts (photos, images, graffiti etc) to express and define the meanings of words. We have strong theoretical reasons for regarding text as co-created by the reader in interaction with the verbal signs of the document being read. The Dictionary gives the reader of words in the wild a means of implementing the act of reading as co-creative, but with a significant difference from those acts that have previously been the focus of theoretical work. The collector of wild words, like the reader of literature, is obviously not so much a viewer as a producer of meaning, but unlike the literary reader, the collector is operating in a textual field whose real-world context is highly unlikely ever to be otherwise recorded. It so obviously goes without saying that it also goes by and vanishes without ever being studied.

The title of the project suggests a distinction between text in two different places: the kind at home and the kind in the wild. But the collector's gaze rapidly suggests that the distinction is only in part one of place. Text at home can also be 'wild' if it can be defamiliarized, e.g. the title of a book on its spine taken in poor light conditions inside a house. The wildness of words, is then 'in the eye of the beholder', though the domestic environment is usually so well regulated that opportunities for perceiving wildness are far more limited than in relatively undomesticated environments. Thus such opportunities tend to occur far less frequently in well-kept or wealthy neighbourhoods than in poorly kept ones, where rubbish is frequently encountered and advertising of all kinds is evident. This suggests ironically that poorer neighbourhoods are in respect of the sheer amount of reading more rather than less literate. But the correlation between wealth and verbosity is not so straightforward. Airport lounges, for example, are rich in examples of wild words. What would seem to matter in this correlation is the acquisitive desire of the population: those who have tend not to read for more.

Such theorizing, however, is clearly at a quite preliminary stage. The project calls for a more systematic ethnography of textuality and its everyday occurence. Insofar as it can be conceptualized, the long-term goal of the project can be put as a question: would it be possible to develop a panoptic topology of the appearance of the legible in everyday life, if even just for one person?

Similar Work

The Dictionary is one of a number of projects that use the internet to share images of textuality. For example, Typography Kicks Ass: Flickr Bold Italic [2] is Flash toy that displays messages left by people using letters from Flickr. The London Evening Standard Headline Generator [3] from thesurrealist.co.uk generates headlines from a Flickr set of images of headlines. IllegalSigns.ca [4] tracks illegal billboards in Toronto and has a Clickable Illegal Signs Map [5] that uses Google Maps. On Flickr one can find sets like Its Only Words [6] of images of texts.

What all these projects have in common is the photographic gaze that captures words or phrases in a context whether for aesthetic purposes or advocacy purposes. The Dictionary is no different, it is meant to provoke reflection on the wild context of text as it is encountered on the street.

The Future

The success of the project lies in how the participants push the simple assumptions encoded in the structure. The project would have failed had no one contributed, but with contributions come exceptions to every design choice. The types of text contributors want to collect and formally tag has led to the specification of a series of improvements that are being implemented with the support of TAPoR and SSHRC. The paper will conclude with some images and the future directions they have provoked:

• We need to parse phrases so that we remove punctuation. For example, "faith," won't find the image for "faith".

• We need to allow implicit words to be entered with parentheses where the word doesn't appear, but is implicit. An example would be http://tapor1-dev.mcmaster.ca/~dictwordwild/show/694 which is filed under "Average" even though the word doesn't appear.

• We need to allow short phrasal verbs and compounds to be entered with quotation marks so they are filed as one item. An example would be "come up" or "happy days".

• We need to allow images of longer passages to identified as "Sentences in the Sticks", "Phrases in the Fields" or "Paragraphs in the Pastures". These would not be filed under individual words, but the full text could be searched.

• We need to allow people to control capitalization so that, for example, "ER" (which stands for "Emergency Room") is not rendered as "Er".

• We need to let people add tags that are not words so images can be sorted according to categories like "Graffiti" or "Billboard".

Links

1. Dictionary of Worlds in the Wild: http://tapor1-dev.mcmaster.ca/~dictwordwild/

2. Typography Kicks Ass: http://www.typographykicksass.com/

3. The London Evening Standard Headline Generator: http://thesurrealist.co.uk/standard.php

4. IllegalSigns.ca: http://illegalsigns.ca/

5. Clickable Illegal Signs Map: http://illegalsigns.ca/?page_id=9

6. Its Only Words: http://www.flickr.com/photos/red_devil/sets/72157594359355250/

Bibliography

Davis, H. and Walton, P. (1983): *Language, Image, Media*, Blackwell, Oxford.

Kress, G. and Van Leeuwen, T. (1996): *Reading images: the grammar of visual design*, Routledge, London.

Kress, G. and Van Leeuwen, T. (2001): *Multimodal Discourse: the modes and media of contemporary communication*, Arnold, London.

McCarty, Willard. 2008 (forthcoming). "Beyond the word: Modelling literary context". Special issue of *Text Technology*, ed. Lisa Charlong, Alan Burke and Brad Nickerson.

McGann, Jerome. 2004. "Marking Texts of Many Dimensions." In Susan Schreibman, Ray Siemens and John Unsworth, eds. *A Companion to Digital Humanities*. Oxford: Blackwell. www.digitalhumanities.org/companion/, 16. (12/10/07). 198-217.

Scollon, R. and Scollon, S. (2003): *Discourses in Places: Language in the material world*, Routledge, London.

Van Leeuwen, T. (2005): *Introducing social semiotics*, Routledge, London.

The TTC-Atenea System: Researching Opportunities in the Field of Art-theoretical Terminology

Nuria Rodríguez
nro@uma.es
University of Málaga, Spain

The purpose of this presentation is to demonstrate the research opportunities that the TTC-ATENEA system offers to specialists in Art Theory. Principally, we will focus on those functionalities that are able to assist specialists in the interpretation of the theoretical discourses, a complex task due to the ambiguity and inaccuracy that distinguish this type of artistic terminology.

Introduction

The TTC-ATENEA system is currently being developed under my supervision in a Project supported by the Ministerio de Educación y Ciencia of Spain called *Desarrollo de un tesauro terminológico conceptual (TTC) de los discursos teórico-artísticos españoles de la Edad Moderna, complementado con un corpus textual informatizado (ATENEA)* (HUM05-00539). It is an interdisciplinary project integrated by the following institutions: University of Málaga (Spain), University of de Santiago de Compostela (Spain), University of Valencia (Spain), European University of Madrid (Spain), the Getty Research Institute (Los Angeles, EE.UU.), the University of Chicago (EE.UU.) and the Centro de Documentación de Bienes Patrimoniales of Chile. Also, it maintains a fruitful collaboration with the Department of Computational Languages and Sciences of the University of Málaga, specifically with the Project called *ISWeb: Una Infraestructura Básica para el Desarrollo de La Web Semántica y su Aplicación a la Mediación Conceptual* (TIN2005-09098-C05-01).

This system can be defined as a virtual net made up of two complementary components: an electronic text database (ATENEA), and the terminological conceptual thesaurus (TTC). The TTC is a knowledge tool for art specialists that consists of a compilation of described, classified and linked terms and concepts. The textual database consists of important art-historical texts (16th -18th centuries), encoded in XML-TEI (1), that supply the terms and concepts recorded in the thesaurus (2).

We want to remark the significant contribution of ATENEA database to the field of the virtual and digital libraries in Spain. Despite the fact that the building of digital and virtual libraries is an important area of development (3), ATENEA represents one of the first textual databases about specialized subject matter –with exception of the digital libraries on literary texts- (4). It is a very different situation from other contexts, where we find relevant specialized electronic textual collections. It

is the case, for example, of the digital libraries about Italian artistic texts developed and managed by Signum (Scuola Normale Superiore di Pisa) (5); or the collections included in the ARTFL Project (University of Chicago) (6).

Providing Solutios to the Ambiguity of Art-theoretical Terminology

The configuration of the ATENEA-TTC system was determined by the main purpose of the project itself--that is, to provide a satisfactory answer to the terminological/conceptual problems that encumber the task of conducting research in art theory, due to the high degree of semantic density of the terminology contained in historical documents. In this respect, it is important to point out that the objective of the Project is not only to distinguish the terminological polysemy, that is, the different concepts that a term denotes (i.e., *disegno* as product, *disegno* as abstract concept, *disegno* as formal component of the paintings…), but also the different interpretations that the same concept assumes in discourses belonging to different authors. For instances, the concept *disegno*, understood as an abstract concept, has not the same interpretation in Zuccaro´s treatise than in Vasari´s one, what implies that the term *disegno* does not mean exactly the same in such texts. Indeed, this is one of the more problematic aspects of textual interpretation, since the same term assumes different semantic nuances in each text or treatise, increasing, in that way, the possibilities for ambiguity and error.

In connection with this question, the eminent Professor Fernando Marías (7) has suggested the development of a "cronogeografía" of the art-theoretical concepts and terms. In his words, the so-called "cronogeografía" should be *a precise means of control of the emergence and use of the theoretical vocabulary as well as of the ideas behind this*. We consider that the methodology proposed by this Project could be an answer to Professor Marías´ suggestion.

It is clear that this approach requires, firstly, a comparative conceptual analysis among the artistic texts stored in ATENEA in order to identify the different interpretations of each concept and, as a result, the semantic variations assumed by every term. To make visible in the thesaurus these semantic-conceptual distinctions, we have proceeded to tag concepts and terms by mean of *parenthetical qualifiers* related to authors, which have been specifically developed to this end. These qualifiers specify the author who has used or defined them. Thus, the TTC records the conceptual-semantic variations among concepts and terms used by different authors, and also shows them graphically and visually. For example, the following representation generated by the system (figure 1):

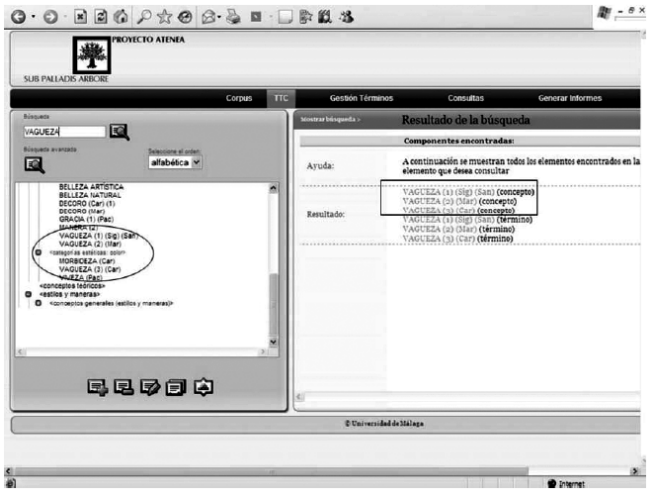


Fig. 1. Uses and interpretations of vagueza in the Spanish art –theory (17th century).

reveals us, at a glance, that *vagueza* has the same interpretation in Sigüenza and Santos´s discourses, but that this assumes other senses in Carducho and Martinez´s treatises.

This procedure implies other significant feature of the ATENEA-TTC system: that terms and concepts are described in the TTC according to how they has been used or defined in each particular text. This last point deserves to be emphasized. Effectively, there are other projects that also link terms to online dictionary entries, in which the user finds general definitions of the terms. Nevertheless, in the TTC terms and concepts are defined in reference to the specific texts in where they have been located. So, clicking on the selected term or concept, the user gets information about how such term or concept has been specifically used and defined in a particular text or by a particular author.

In addition to that, the TTC-ATENEA system provides full interaction between the ATENEA corpus and the TTC. The system easily allows the definition of connections between a term or a concept in an XML-TEI marked text and the registry of the TTC. Thus, terms and concepts in the text are linked to all their relevant information, stored and structured in the TTC records. In the same way, the system allows to establish a connection between terms and concepts registered in the TTC and the different places where they appear in the XML-TEI marked texts. As far as XML-TEI marked texts and TTC are stored in the same database, we can establish these links in a totally effective and efficient way without having to access to different data repositories to retrieve and link this information. Consequently, one of the most important potentialities of the system as a research tool derives from the interactions that users are able to establish among the different types of information compiled in each repository.

Finally, future extensions of the TTC-ATENEA will study the possibility of use not only a thesaurus but a full fledged OWL [7] ontology and its integration in the *Khaos Semantic*

Web platform [8] [9] [10] in order to achieve a formal explicit representation of the artistic epistemology contained in the texts.

Research Possibilities

The system allows users to retrieve the stored information from both terminological-conceptual and textual repositories. Nevertheless, as indicated above, the most important research possibilities derive from the interactions that users are able to establish between each repository. To illustrate this, we will give a brief example.

1. ATENEA Corpus as a starting point. Imagine that a specialist is interested in studying the meanings of the terms used by the Italian author F. Pacheco in his Spanish treatise (*Diálogos de la Pintura*, 1633). ATENEA Corpus enables the specialist to visualize the electronic transcription of the text, and explore it linguistically by means of different queries -concordances, frequency lists, co-occurrences…- since the system has been implemented with its own textual analyzer. These functionalities do not constitute any novelty; all we know that. Consequently, the most relevant feature resides in the fact that the system allows user to browse directly from a term or concept located in a text to the TTC record where such term or concept is described according to its use in Carducho´s text (figure 2).

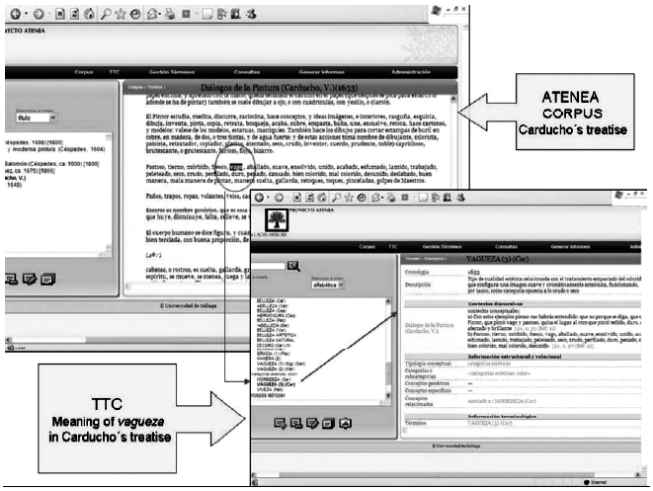


Fig. 2. Connection between a concept in an XML-TEI marked text (ATENEA) and the TTC record, where it is described.

In this way, the specialist is able to identify the precise senses that the terms used by Carducho assume, and, therefore, analyze with more accuracy his theory about the painting.

Once the specialist has accessed to the TTC records, the system offers him other research possibilities given that the recorded information is linked to other sections of the TTC as well as to other texts stored in ATENEA. Thus, the specialist is able to go more deeply in his investigation and to get complementary information browsing through the site. The presentation will show some clarifying examples of these other possibilities.

2.TTC as a starting point. Imagine now that the specialist is interested in the use of a particular term. In this case, the most convenient is to search the term in the TTC. Made the query, the specialist retrieves the different concepts associated to the term, as in a conventional specialized dictionary. However, as the term is marked with the *author-parenthetical qualifiers*, the specialist also finds out in which discourses the term has been identified denoting each concept (see figure1). In fact, various hyperlinks enable him to go from the TTC record to the exact point of the texts -stored in ATENEA- where the term appears denoting such concepts (figure 3).

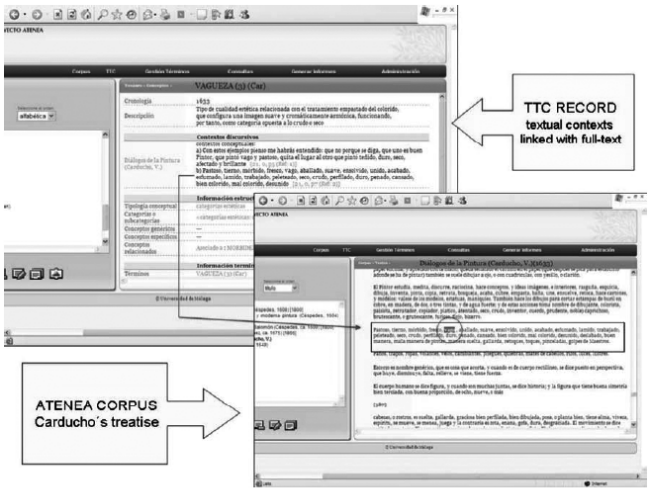


Fig. 3. Connection between a concept registered in the TTC and the places where they appear in the XML-TEI marked texts.

Once in ATENEA Corpus, the specialist has the opportunity to use its functionalities in order to specify his investigation.

The presentation will show in full detail these possibilities of usage derived from this interactivity-based approach, and will include time for Q&A. We believe that the questions taken into consideration in our exposition will be of interest to other humanistic disciplines, since the difficulties of textual interpretations due to the terminological ambiguity is a common place in the field of the Humanities.

Notes

(1) http://www.tei-c.org. The DTD used in this Project has been created from the *TEI-Lite DTD*. We have selected those elements that best adjusted to our requirements and we have added others considered necessaries.

(2) For further information on this system see [1], [2] and [3].

(3) The most paradigmatic example is, without doubt, the *Biblioteca Virtual Miguel de Cervantes* [http://www.cervantes.virtual.com]. For more information, see the directories of the Junta de Andalucía [http://www.juntadeandalucia.es/cultura/bibliotecavirtualandalucia/recursos/otros_recursos.cmd], Ministerio de Cultura [http://www.mcu.es/roai/es/comunidades/registros.cmd] or the list posted on the web of the University of Alicante [http://www.ua.es/es/bibliotecas/referencia/electronica/bibdigi.html]. Also, see [4] and [5].

(4) Other Spanish projects about specialized digital library are: the *Biblioteca Virtual del Pensamiento Político Hispánico "Saavedra Fajardo"*, whose purpose is to develop a textual corpus of documents and sources about the Spanish theories on politic [http://saavedrafajardo.um.es/WEB/HTML/web2.html]; and the *Biblioteca Digital Dioscórides* (Complutense University of Madrid) that provides digitalized texts from its biomedicine-historical collection [http://www.ucm.es/BUCM/foa/presentacion.htm].

(5) http://www.signum.sns.it

(6) http://humanities.uchicago.edu/orgs/ARTFL/

(7) See [6]

References

[1] Rodríguez Ortega, N. (2006). *Facultades y maneras en los discursos de Francisco Pacheco y Vicente Carducho .Tesauro terminológico-conceptual.* Málaga: Universidad de Málaga.

[2] Rodríguez Ortega, N. (2006). "Use of computing tools for organizing and Accessing Theory and Art Criticism Information: the TTC-ATENEA Project", *Digital Humanities`06 Conference*, 4-9 de julio (2006), Universidad de la Sorbonne, París.

[3] Rodríguez Ortega, N. (2005). "Terminological/Conceptual Thesaurus (TTC): Polivalency and Multidimensionality in a Tool for Organizing and Accessing Art-Historical Information", *Visual Resource Association Bulletin*, vol. 32, n. 2, pp. 34-43.

[4] *Bibliotecas Virtuales FHL* (2005). Madrid, Fundación Ignacio Larramendi.

[5] García Camarero, E. y García Melero, L.A. (2001). *La biblioteca digital*. Madrid, Arco/libros.

[6] Marías, F. (2004). "El lenguaje artístico de Diego Velázquez y el problema de la `Memoria de las pinturas del Escorial´", en A*ctas del Simposio Internacional Velázquez 1999*, Junta de Andalucía, Sevilla, pp. 167-177.

[7] OWL Web Ontology Language. http://www.w3.org/TR/owl-features/

[8] Navas-Delgado I., Aldana-Montes J. F. (2004). "A Distributed Semantic Mediation Architecture". *Journal of Information and Organizational Sciences*, vol. 28, n. 1-2. December, pp. 135-150.

[9] Moreno, N., Navas, I., Aldana, J.F. (2003). "Putting the Semantic Web to Work with DB Technology". *IEEE Bulletin of the Technical Committee on Data Engineering*. December, pp. 49-54.

[10] Berners-Lee, T., Hendler, J., Lassila, O. (2001). "The semantic web". *Scientific American* (2001).

Re-Engineering the Tree of Knowledge: Vector Space Analysis and Centroid-Based Clustering in the Encyclopédie

Glenn Roe
glenn@diderot.uchicago.edu
University of Chicago, USA

Robert Voyer
rlvoyer@diderot.uchicago.edu
University of Chicago, USA

Russell Horton
russ@diderot@uchicago.edu
University of Chicago, USA

Charles Cooney
cmcooney@diderot.uchicago.edu
University of Chicago, USA

Mark Olsen
mark@barkov.uchicago.edu
University of Chicago, USA

Robert Morrissey
rmorriss@uchicago.edu
University of Chicago, USA

The *Encyclopédie* of Denis Diderot and Jean le Rond d'Alembert is one of the crowning achievements of the French Enlightenment. Mobilizing many of the great – and the not-so-great – *philosophes* of the eighteenth century, it was a massive reference work for the arts and sciences, which sought to organize and transmit the totality of human knowledge while at the same time serving as a vehicle for critical thinking. The highly complex structure of the work, with its series of classifications, cross-references and multiple points of entry makes it not only, as has been often remarked, a kind of forerunner of the internet[1], but also a work that constitutes an ideal test bed for exploring the impact of new machine learning and information retrieval technologies. This paper will outline our ongoing research into the ontology of the *Encyclopédie* , a data model based on the classes of knowledge assigned to articles by the editors and representing, when organized hierarchically, a system of human understanding. Building upon past work, we intend to move beyond the more traditional text and data mining approaches used to categorize individual articles and towards a treatment of the entire encyclopedic system as it is elaborated through the distribution and interconnectedness of the classes of knowledge. To achieve our goals, we plan on using a vector space model and centroid-based clustering to plot the relationships of the *Encyclopédie*'s epistemological categories, generating a map that will hopefully serve as a

corollary to the taxonomic and iconographic representations of knowledge found in the 18th century.[2]

Over the past year we have conducted a series of supervised machine learning experiments examining the classification scheme found in the *Encyclopédie*, the results of which were presented at Digital Humanities 2007. Our intent in these experiments was to classify the more than 22,000 unclassified articles using the known classes of knowledge as our training model. Ultimately the classifier performed as well as could be expected and we were left with 22,000 new classifications to evaluate. While there were certainly too many instances to verify by hand, we were nonetheless encouraged by the assigned classifications for a small sample of articles. Due to the limitations of this exercise, however, we decided to leverage the information given us by the editors in exploring the known classifications and their relationship to each other and then later, to consider the classification scheme as a whole by examining the general distribution of classes over the entire work as opposed to individual instances. Using the model assembled for our first experiment - trained on all of the known classifications - we then reclassified all of the classified articles. Our goal in the results analysis was twofold: first, we were curious as to the overall performance of our classification algorithm, i.e., how well it correctly labeled the known articles; and secondly, we wanted to use these new classifications to examine the outliers or misclassified articles in an attempt to understand better the presumed coherency and consistency of the editors' original classification scheme.[3]

In examining some of the reclassified articles, and in light of what we know about Enlightenment conceptions of human knowledge and understanding – ideas for which the *Encyclopédie* and its editors were in no small way responsible – it would seem that there are numerous cases where the machine's classification is in fact more appropriate than that of the editors. The machine's inability to reproduce the editors' scheme with stunning accuracy came somewhat as a surprise and called into question our previous assumptions about the overall structure and ordering of their system of knowledge. Modeled after Sir Francis Bacon's organization of the Sciences and human learning, the *Système Figuré des connaissances humaines* is a typographical diagram of the various relationships between all aspects of human understanding stemming from the three "root" faculties of Reason, Memory, and Imagination.[4] It provides us, in its most rudimentary form, with a view from above of the editors' conception of the structure and interconnectedness of knowledge in the 18th century. However, given our discovery that the editors' classification scheme is not quite as coherent as we originally believed, it is possible that the *Système figuré* and the expanded *Arbre généalogique des sciences et arts*, or tree of knowledge, as spatial abstractions, were not loyal to the complex network of contextual relationships as manifested in the text. Machine learning and vector space analysis offer us the possibility, for the first time, to explore this network of classifications as a whole, leveraging the textual content of the entire work rather than relying on external abstractions.

The vector space model is a standard framework in which to consider text mining questions. Within this model, each article is represented as a vector in a very high-dimensional space where each dimension corresponds to the words in our vocabulary. The components of our vectors can range from simple word frequencies, to n-gram and lemma counts, in addition to parts of speech and *tf-idf* (term frequency-inverse document frequency), which is a standard weight used in information retrieval. The goal of *tf-idf* is to filter out both extremely common and extremely rare words by offsetting term frequency by document frequency. Using *tf-idf* weights, we will store every article vector in a matrix corresponding to its class of knowledge. We will then distill these class matrices into individual class vectors corresponding to the centroid of the matrix.[5]

Centroid or mean vectors have been employed in classification experiments with great success.[6] While this approach is inherently lossy, our initial research suggests that by filtering out function words and lemmatizing, we can reduce our class matrices in this way and still retain a distinct class core. Using standard vector space similarity measures and an open-source clustering engine we will cluster the class vectors and produce a new tree of knowledge based on semantic similarity. We expect the tree to be best illustrated as a weighted undirected graph, with fully-connected sub-graphs. We will generate graphs using both the original classifications and the machine's decisions as our labels.

Due to the size and scale of the *Encyclopédie*, its editors adopted three distinct modes of organization - dictionary/alphabetic, hierarchical classification, and cross-references - which, when taken together, were said to represent encyclopedic knowledge in all its complexity.[7] Using the machine learning techniques outlined above, namely vector space analysis and centroid-based clustering, we intend to generate a fourth system of organization based on semantic similarity. It is our hope that a digital representation of the ordering and interconnectedness of the *Encyclopédie* will highlight the network of textual relationships as they unfold within the work itself, offering a more inclusive view of its semantic structure than previous abstractions could provide. This new "tree of knowledge" can act as a complement to its predecessors, providing new points of entry into the *Encyclopédie* while at the same time suggesting previously unnoticed relationships between its categories.

References

[1] E. Brian, "L'ancêtre de l'hypertexte", in *Les Cahiers de Science et Vie* 47, October 1998, pp. 28-38.

[2] R. Darnton. "Philosophers Trim the Tree of Knowledge." *The Great Cat Massacre and Other Essays in French Cultural History*. London: Penguin, 1985, pp. 185-207.

[3] R. Horton, R. Morrissey, M. Olsen, G. Roe, and R. Voyer. "Mining Eighteenth Century Ontologies: Machine Learning and Knowledge Classification in the *Encyclopédie*." under consideration for *Digital Humanities Quarterly*.

[4] For images of the *Système figuré des connaissance humaines* and the *Arbre généalogique des sciences et des arts principaux* see, http://www.lib.uchicago.edu/efts/ARTFL/projects/encyc/systeme2.jpg; and http://artfl.uchicago.edu/cactus/.

[5] G. Salton, A. Wong, and C. S. Yang. "A vector space model for automatic indexing." Communications of the ACM. November 1975.

[6] E. Han and G. Karypis. "Centroid-Based Document Classification." *Principles of Data Mining and Knowledge Discovery*. New York: Springer, 2000, pp. 424-431.

[7] G. Blanchard and M. Olsen. "Le système de renvois dans l'Encyclopédie: une cartographie de la structure des connaissances au XVIIIème siècle." *Recherches sur Diderot et sur l'Encyclopédie*, April 2002.

Assumptions, Statistical Tests, and Non-traditional Authorship Attribution Studies -- Part II

Joseph Rudman
jr20@andrew.cmu.edu
Carnegie Mellon University

> Statistical inferences are based only in part upon the observations. An equally important base is formed by prior assumptions about the underlying situation. Even in the simplist cases, there are explicit or implicit assumptions about randomness and independence....
>
> *Huber*

Introduction

Controversy surrounds the methodology used in non-traditional authorship attribution studies -- those studies that make use of the computer, stylistics, and the computer. [Rudman 2003] [Rudman 1998] One major problem is that many of the most commonly used statistical tests have assumptions about the input data that do not hold for the primary data of these attribution studies -- the textual data itself (e.g. normal distributions, randomness, independence). "...inappropriate statistical methods....In particular, asymtotic normality assumptions have been used unjustifiably, leading to flawed results." [Dunning, p.61] "Assumptions such as the binomiality of word counts or the independence of several variables chosen as markers need checking." [Mosteller and Wallace]

This paper looks at some of the more frequently used tests (e.g. chi-square, Burrows Delta) and then at the questions, "Are there assumptions behind various tests that are not mentioned in the studies?" and "Does the use of statistical tests whose assumptions are not met invalidate the results?" Part I of this paper was presented at the "10th Jubilee Conference of the International Federation of Classification Societies," 28 July 2006 in Ljubljana, Slovenia. The idea behind Part I was to get as much input as possible from a cross-section of statisticians from around the world. Part II takes the statisticians' input and continues looking at the problem of assumptions made by various statistical tests used in non-traditional attribution studies.

Because of the complicated intertwining of disciplines in non-traditional authorship attribution, each phase of the experimental design should be as accurate and as scientifically rigorous as possible. The systematic errors of each step must be computed and summed so that the attribution study can report an overall systematic error -- Mosteller and Wallace come close to doing this in their "Federalist" study -- and it seems that no one else tries.

It is the systematic error that drives the focus of this paper -- the assumptions behind the statistical tests used in attribution studies. I am not concerned with assumptions made by practitioners that are not an integral part of the statistics -- e.g. Morton, using the cusum technique, assumes that style markers are constant across genre -- this has been shown to be false but has nothing to do with the cusum test itself. [Sanford et al.]

There are many statistical tests along with their many assumptions that are used in non-traditional attribution studies - e.g the Efron-Thisted tests are based on the assumption that things (words) are well mixed in time [Valenza] -- obviously not true in attribution studies. Because of time constraints, I want to limit the number of tests discussed to three: the chi-square, Burrows' Delta, and the third being more of a `catagory' -- machine learning.

This paper looks at each test and attempts to explain why the assumptions exist -- how they are determined -- how integral assumptions are to the use of the test.

Chi-Square test

The chi-square test, in all of its manifestations, is ubiquitous in non-traditional authorship studies. It also is the test that has recieved the most criticism from other practitioners. Delcourt lists some of the problems [Delcourt (from Lewis and Burke)]:

1) Lack of independence among the single events or measures

2) Small theoretical frequencies

3) Neglect of frequencies of non-occurence

4) Failure to equalize the sum of the observed frequencies and the sum of the theoretical frequencies

5) Indeterminate theoretical frequencies

6) Incorrect or questionable categorizing

7) Use of non-frequency data

8) Incorrect determination of the number of degrees of freedom

Does the chi-square test always demand independence and randomness, and ever a normal distribution? Gravetter and Wallnau say that although the chi-square test is non-parametric, "...they make few (if any) assumptions about the population distribution." [Gravetter and Wallnau, 583]

The ramifications of ignoring these assumptions are discussed.

Burrows' Delta

This section discusses the assumptions behind Burrows' Delta.

The assumptions behind Burrows' delta are articulated by Shlomo Argamon:

1) Each indicator word is assumed to be randomly distributed

2) Each indicator word is assumed to be statistically independent of every other indicator word's frequency

The ramifications of ignoring these assumptions are discussed.

Do the assumptions really make a difference in looking at the results? Burrows' overall methodology and answers are to be highly commended.

Machine Learning -- Data Mining

Almost all machine learning statistical techniques assume independence in the data.

David Hand et al. say, "...of course...the independence assumption is just that, an assumption, and typically it is far too strong an assumption for most real world data mining problems." [Hand et al., 289]

Malerba et al. state, "Problems caused by the independence assumption are particularly evident in at least three situations: learning multiple attributes in attribute-based domains, learning multiple predicates in inductive logic programming, and learning classification rules for labeling."

The ramifications of ignoring these assumptions are discussed.

Conclusion

The question at hand is not, "Does the violation of the assumptions matter?", but rather, "How much does the violation of assumptions matter?" and, "How can we either correct for this or calculate a systematic error?"

How are we to view attribution studies that violate assumptions, yet show some success with studies using only known authors? The works of McNemar, Baayen, and Mosteller and Wallace are discussed.

The following are answers often given to questions about assumptions:

1) The statistics are robust

The tests are so robust that the assumptions just do not matter.

2) They work until they don't work

You will know when this is and then you can re-think what to do.

3) Any problems are washed out by statistics

There is so much data that any problems from the violation of assumptions are negligible.

These answers and some solutions are discussed.

Preliminary References

Argamon, Shlomo. "Interpreting Burrows' Delta: Geometric and Probabilistic Foundations." (To be published in *Literary and Linguistic Computing*.) Pre-print courtesy of author, 2006.

Baayen, R. Harald. "The Randomness Assumption in Word Frequency Statistics." Research in Humanities Computing 5. Ed. Fiorgio Perissinotto. Oxford: Clarendon Press, 1996, 17-31.

Banks, David. "Questioning Authority." *Classification Society of North America Newsletter* 44 (1996): 1.

Box, George E.P. et al. *Statistics for Experimenters: An Introduction to Design, Data Analysis, and Model Building.* New York: John Wiley and Sons, 1978.

Burrows, John F. "`Delta': A Measure of Stylistic Difference and a Guide to Likely Authorship." *Literary and Linguistic Computing*17.3 (2002): 267-287.

Cox, C. Phillip. A Handbook of Introductory Statistical Methods. New York: John Wiley and Sons, 1987.

Delcourt, Christian. "Stylometry." *Revue Belge de Philologie et d'histoire* 80.3 (2002): 979-1002.

Dunning, T. "Accurate Methods for the Statistics of Suprise and Coincidence." Computational Linguistics , 19.1 (1993): 61-74.

Dytham, Calvin. *Choosing and Using Statistics: A Biologist's Guide*. (Second Edition) Malden, MA: Blackwell, 2003.

Gravetter, Fredrick J., and Larry B. Wallnau. *Statistics for the Behavioral Sciences*. (5th Edition) Belmont, CA: Wadsworth Thomson Learning, 2000.

Hand, David, et al. *Principles of Data Mining*. Cambridge, MA: The MIT Press, 2001.

Holmes, David I. "Stylometry." In *Encyclopedia of Statistical Sciences* (Update of Volume 3). Eds. Samuel Kotz, Campbell, and David Banks. New York: John Wiley and Sons, 1999.

Hoover, David L. "Statistical Stylistics and Authorship Attribution: An Empirical Investigation." *Literary and Linguistic Computing* 16.4 (2001): 421-444.

Huber, Peter J. *Robust Statistics.* New York: John Wiley and Sons, 1981.

Khmelev, Dmitri V., and Fiona J. Tweedie. "Using Markov Chains for Identification of Writers." *Literary and Linguistic Computing* 16.3 (2001): 299-307.

Lewis, D., and C.J. Burke. "The Use and Misuse of the Chi-squared Test." *Psychological Bulletin* 46.6 (1949): 433-489.

Malerba, D., et al. " A Multistrategy Approach to learning Multiple Dependent Concepts." In *Machine Learning and Statistics: The Interface*. Eds. G. Nakhaeizadeh and C.C. Taylor. New York: John Wiley and Sons, 87-106.

McNemar, Quinn. Psychological Statistics (3rd Edition). New York: John Wiley and Sons, 1962.

Mosteller, Fredrick, and David L. Wallace. *Applied Bayesian and Classical Inference: The Case of the "Federalist Papers."* New York: Springer-Verlag, 1984.

Rudman, Joseph. "Cherry Picking in Non-Traditional Authorship Attribution Studies." *Chance* 16.2 (2003): 26-32.

Rudman, Joseph. "The State of Authorship Attribution Studies: Some Problems and Solutions." Computers and the Humanities 31 (1998): 351-365.

Sanford, Anthony J., et al. "A Critical Examination of Assumptions Underlying the Cusum Technique of Forensic Linguistics." *Forensic Linguistics* 1.2 (1994): 151-167.

Valenza, Robert J. "Are the Thisted-Efron Authorship Tests Valid?" *Computers and the Humanities* 25.1 (1991): 27-46.

Wachal, Robert Stanly. "Linguistic Evidence, Statistical Inference, and Disputed Authorship." Ph.D Dissertation, University of Wisconsin, 1966.

Williams, C.B. "A Note on the Statistical Analysis of Sentence Length as a Criterion of Literary Style." *Biometrica* 31 (1939-1940): 356-351.

Does Size Matter? A Re-examination of a Time-proven Method

Jan Rybicki
jrybicki@ap.krakow.pl
Pedagogical University, Krakow, Poland

Previous stylometric studies (Rybicki 2006, 2006a, 2007) of patterns in multivariate diagrams of correlation matrices derived from relative frequencies of most frequent words in character idiolects (Burrows 1987) in a number of originals and translations (respectively, Sienkiewicz's trilogy of historical romances and its two English translations; Rybicki's Polish translations of novels by John le Carré; and the three English versions of Hamlet, Q1, Q2, F, and its translations into nine different languages) have all yielded interesting similarities in the layout of data points in the above-mentioned diagrams for corresponding originals and translations. The repetitiveness of the observed phenomenon in many such pairs immediately raised some questions as to its causes – the more so as the most-frequent-word lists for any original and its translations, consisting primarily of functions words, modals, pronouns, prepositions and articles (if any in a given language) contains few direct counterparts in any two languages.

The primary suspect for a possible bias was the difference in size between the parts of individual characters, since that particular feature invariably remains proportionally similar between original and translation (Rybicki 2000). The import of this element remained unnoticed in many other studies employing Burrows's time-proven method (later developed, evaluated and applied by a number of scholars, including Hoover 2002) since they were either conducted on equal samples or, more importantly, although some did address the problem of translation, they never dealt with so many translations at a time as did e.g. the Hamlet study (Rybicki 2007).

The emerging suspicion was that since the sizes of the characters studied do not vary significantly between the twelve versions of Hamlet, this proportion might heavily influence the multivariate graphs produced – the more so as, in most cases, the most talkative characters occupied central positions in the graphs, while the lesser parts were usually limited to the peripheries. This (together with similar effects observed in earlier studies) had to lead to a reassessment of the results. Also, since most studies were conducted in traditional male-dominated writing, resulting in female characters speaking little in proportion even to their importance in the plot, these peripheries usually included separate groups of women; while this was often seen as the authors' ability to stylistically differentiate "male" and "female" idiom, the size bias could distort even this quite possible effect.

A number of tests was then performed on character idiolects and narrative fragments of various sizes and various configurations of characters from selected English novels and plays – ranging from Ruth and David Copperfield; from Conan Doyle to Poe; narration in Jane Austen, or plays by Ben Jonson and Samuel Beckett – to investigate how the impact of size distorts the image of stylistic differences presented in multivariate diagrams.

One possible source of the bias is the way relative frequencies of individual words are calculated. Being a simple ratio of word frequency to size of sample, it may be as unreliable as another very similar formula, the type-to-token ratio (or, indeed, one as complex as Yule's K), has been found to be unreliable as a measure of vocabulary richness in samples of different size (Tweedie & Baayen 1997). Also, due to the nature of multivariate analyses used (both Factor Analysis and Multimensional Scaling), it is not surprising that the less stable statistics of the less talkative characters would have their data points pushed to the peripheries of the graph. Finally, since the list of the most frequent words used in the analysis is most heavily influenced by the longest parts in any text, this might also be the reason for the "centralising" bias visible in data points for such characters.

It should be stressed that the above reservations do not, in any way, invalidate the entire method; indeed, results for samples of comparative size remain reliable and unbiased. Also, it should not be forgotten that size of a character's part is in itself an individuating feature of that character.

References

Burrows, J.F. (1987). *Computation into Criticism: A Study of Jane Austen's Novels and an Experiment in Method*. Oxford: Clarendon Press.

Hoover, D.L. (2002). *New Directions in Statistical Stylistics and Authorship Attribution. Proc. ALLC/ACH.*

Rybicki, J. (2000). *A Computer-Assisted Comparative Analysis of Henryk Sienkiewicz's Trilogy and its Two English Translations*. PhD thesis, Kraków: Akademia Pedagogiczna.

Rybicki, J. (2006). *Burrowing into Translation: Character Idiolects in Henryk Sienkiewicz's Trilogy and its Two English Translations. LLC* 21.1: 91-103.

Rybicki, J. (2006a). *Can I Write like John le Carré*. Paris: Digital Humanities 2006.

Rybicki, J. (2007). *Twelve Hamlets: A Stylometric Analysis of Major Characters' Idiolects in Three English Versions and Nine Translations.* Urbana-Champaign: Digital Humanities 2007.

F. Tweedie, R. Baayen (1997). Lexical 'constants' in stylometry and authorship studies, <http://www.cs.queensu.ca/achallc97/papers/s004.html>

The TEI as Luminol: Forensic Philology in a Digital Age

Stephanie Schlitz
sschlitz@bloomu.edu
Bloomsburg University, USA

A significant number of digital editions have been published in recent years, and many of these serve as exemplars for those of us working within the digital editing community. A glance at *Electronic Text Editing*, for example, published in 2006, indicates that such projects address a wide berth of editorial problems, ranging from transcriptional practices to document management to authenticity. And they offer a wealth of information to editors working on various types of digital editions.

In this paper, I consider an editorial problem that has yet to be resolved. My discussion centers on the difficulties that arise in editing a single, albeit somewhat unusual, Icelandic saga: *Hafgeirs saga Flateyings*. This saga is preserved in an unsigned, eighteenth-century manuscript, Additamenta 6, folio (Add. 6, fol.). Today housed in the collection of the Arni Magnusson Institute for Icelandic Studies in Reykjavík, Iceland, the manuscript was originally held as part of the Arnamagnæan Collection in Copenhagen, Denmark. According to the flyleaf: "Saga af Hafgeyre flateying udskreven af en Membran der kommen er fra Island 1774 in 4to exarata Seculo xij" (*Hafgeirs saga Flayetings was copied from a twelfth-century manuscript written in quarto, which came* [to Copenhagen] *from Iceland in 1774*).

While such a manuscript might appear unremarkable, as a number of paper manuscripts were copied during the late eighteenth century in Copenhagen, then the capital of Iceland and the seat of Icelandic manuscript transmission during this period, only twelve Old Norse/Old Icelandic manuscripts of those catalogued in the Copenhagen collections are dated to the twelfth century, while a mere eighteen are dated to 1200 (Kalund 512). The dating on the flyleaf is therefore unusual, and as it turns out, suspect as well, since no catalog entry exists to record the existence of the alleged source manuscript. Moreover, according to Jorgensen, the motif sequences found in *Hafgeirs saga* bear a striking resemblance to those found in the well-known mythical-heroic saga *Hálfdanars saga Brönufóstra* (157). And in a fascinating argument based primarily on this fact, Jorgensen argues that Add. 6, fol. is a forgery, claiming that Þorlákur Magnússon Ísfiord, an Icelandic student studying and working in Copenhagen during the 1780s, composed and sold *Hafgeirs saga* as a copy of an authentic medieval Icelandic saga (163).

In spite of its questionable origin, *Hafgeirs saga* stands as a remnant of Iceland's literary, linguistic, and textual history, and Add. 6, fol. can therefore be viewed as an important cultural artefact. As the editor of the *Hafgeirs saga* manuscript, my aim is to provide a 'reliable' (see "Guidelines for Editors of Scholarly Editions" Section 1.1) electronic edition of the text and the manuscript. But the question, at least until recently, was how? What is the best way to represent such a text? Encoding the work according to a markup standard such as the TEI Guidelines is surely a starting point, but doing so doesn't solve one of the primary concerns: How to represent the manuscript reliably (which presents a complex editorial problem of its own), while at the same time illuminating the textual and linguistic 'artefacts' that may offer readers insight into the saga's origin?

At the 2007 Digital Humanities Summer Institute, Matthew Driscoll gave a talk entitled "Everything But the Smell: Toward a More Artefactual Digital Philology." The talk provided a brief history of the shift toward 'new' philology, and, importantly, underscored the significance of the material or 'artefactual' aspect of new philology, which views manuscripts as physical objects and thus as cultural artefacts which offer insight into the "process to which they are witness" ("Everything But the Smell: Toward a More Artefactual Digital Philology"). The TEI, Driscoll pointed out, offers support for artefactual philology, and the descriptive framework of the P5 Guidelines, which defines Extensible Markup Language (XML) as the underlying encoding language, is ideal for this kind of work. Moreover, as Driscoll suggests, there is "no limit to the information one can add to a text - apart, that is, from the limits of our own imagination" ("Electronic Textual Editing: Levels of Transcription"). To be sure, this paper does not lapse into what McCarty refers to as the 'complete encoding fallacy' or the 'mimetic fallacy' (see Dahlström 24), but it does agree with Driscoll in arguing that P5-conformant editions, which can offer a significant layering of data and metadata, have the potential to move the reader beyond the aesthetics of sensory experience.

By definition, artefactual philology portends a kind of 'evidentiary' approach, one that can frame textual features, including linguistic and non-linguistic features (such as lacunae) for example, as kinds of evidence. Evidence of what? That is in the hands of the editors and the readers, but conceivably: linguistic development, the transmission process, literary merit, and so on. And when an evidentiary approach to philology is defined within a 'generative' approach to a scholarly edition (see Vanhoutte's "Generating" 164), a new direction in electronic editing becomes possible.

This paper explores this new direction. It shows how the *Hafgeirs saga* edition employs such a framework to address the problem of describing linguistic and non-linguistic artefacts, which are precisely the kinds of evidence that can bear witness to the composition date and transmission process. And it demonstrates how the display decisions result in an interactive and dynamic experience for the edition's readers.

For example, in addition to elements defined within existing modules of the TEI, the edition's schema defines four new elements[1]. Drawn from the perspectives of historical and socio- linguistics, these elements are intended to aid readers in evaluating the saga's composition date. Given the 'logic

of abundance' (see Flanders 135), encoding the metadata described in the new elements beside the descriptive data described by pre-defined elements (such as <del> for example) can be accomplished without sacrificing the role of the text as a literary and cultural artefact. Because the transformation of the source XML has been designed to display interactively the various encodings of the text[2], readers can view or suppress the various descriptions and generate their own novel versions of the text. Readers can display archaisms, for instance, and assess whether they are "affectation[s] of spurious age" (Einar 39) or features consistent with the textual transmission process, and they can view borrowings, for instance, and assess whether they preclude a medieval origin or are to be expected in a text ostensibly copied by a scribe living in Copenhagen during the eighteenth century. Or they can suppress these features and view the normalized transcription, the semi-diplomatic transcription, emendations, editorial comments, or any combination of these.

Ultimately, this paper synthesizes aspects of text editing, philology, and linguistics to explore a new direction in digital editing. In doing so, it frames P5 XML as a kind of *luminol* that, when transformed, can be used to illuminate new types of evidence: linguistic and philological data. And the goal is identical to that of a crime-scene investigator's: Not necessarily to solve the case, but to preserve and to present the evidence.

Notes

[1] The <borrowing> element describes a non-native word which has been adopted into the language. Distinct from the <foreign> element, a borrowing may have the following attributes: @sourcelang (source language), @borrowdate (date of borrowing), @borrowtype (type of borrowing; e.g. calque, loanword). The <modernism> element describes a word, phrase, usage, or peculiarity of style which represents an innovative or distinctively modern feature. The <neologism> element describes a word or phrase which is new to the language or one which has been recently coined. The <archaism> element describes an archaic morphological, phonological, or syntactic feature or an archaic word, phrase, expression, etc.

[2]My discussion of the display will continue my in-progress work with Garrick Bodine, most recently presented at TEI@20 on November 1, 2007: "From XML to XSL, jQuery, and the Display of TEI Documents."

References

Burnard, Lou, and Katherine O'Brien O'Keefe and John Unsworth eds. *Electronic Text Editing*. New York: The Modern Language Association of America, 2006.

Dalhström, Mats. "How Reproductive is a Scholarly Edition?" *Literary and Linguistic Computing*. 19:1 (2004): 17-33.

Driscoll, M.J. "Everything But the Smell: Toward a More Artefactual Digital Philology." Digital Humanities Summer Institute. 2007.

Driscoll, M.J. Text Encoding Initiative Consortium. 15 August 2007. "Electronic Textual Editing: Levels of Transcription." <http://www.tei-c.org/Activities/ETE/Preview/driscoll.xml>.

Einar Ól. Sveinsson. *Dating the Icelandic Sagas*. London: Viking Society for Northern Research, 1958.

Flanders, Julia. "Gender and the Eectronic Text." *Electronic Text*. Ed. Kathryn Sutherland. Clarendon Press: Oxford, 1997. 127-143.

"Guidelines for Editors of Scholarly Editions." *Modern Language Association*. 25 Sept. 2007. 10 Nov. 2007. <http://www.mla.org/cse_guidelines>.

Jorgensen, Peter. "Hafgeirs saga Flateyings: An Eighteenth-Century Forgery." *Journal of English and Germanic Philology*. LXXVI (1977): 155-164.

Kalund, Kristian. *Katalog over de oldnorsk-islandske handskrifter i det store kongelige bibliotek og i universitetsbiblioteket.* Kobenhavn: 1900.

Vanhoutte, Edward. "Traditional Editorial Standards and the Digital Edition." *Learned Love: Proceedings from the Emblem Project Utrecht Conference on Dutch Love Emblems and the Internet (November 2006).* Eds. Els Stronks and Peter Boot. DANS Symposium Publications 1: The Hague, 2007. 157-174.

A Multi-version Wiki

Desmond Schmidt
schmidt@itee.uq.edu.au
University of Queensland, Australia

Nicoletta Brocca
nbrocca@unive.it
Università Ca' Foscari, Italy

Domenico Fiormonte
fiormont@uniroma3.it
Università Roma Tre, Italy

If, until the middle of the nineties, the main preoccupation of the scholarly editor had been that of conserving as faithfully as possible the information contained in the original sources, in the last ten years attention has shifted to the user's participation in the production of web content. This development has occurred under pressure from the 'native' forms of digital communication as characterised by the term 'Web 2.0' (Tapscott and Williams, 2006). We are interested in harnessing the collaborative power of these new environments to create 'fluid, co-operative, distributed' (Robinson 2007: 10) digital versions of our textual cultural heritage. The approach taken by Hypernietzsche (Barbera, 2007) similarly tries to foster collaboration around cultural heritage texts. However, while Hypernietzsche focuses on annotation, what we have developed is a wiki-inspired environment for documents that were previously considered too complex for this kind of editing.

Before explaining how this can work in practice, it is worthwhile to reflect on why it is necessary to change our means of editing cultural heritage texts. G.T. Tanselle has recently argued (2006) that in the digital medium 'we still have to confront the same issues that editors have struggled with for twenty-five hundred years'. That may be true for analysis of the source texts themselves, but digitisation does change fundamentally both the objectives of editing and the function of the edition. For the representation of a text 'is conditioned by the modes of its production and reproduction' (Segre, 1981). The well-known editorial method of Lachmann, developed for classical texts, and that of the New Bibliography for early printed books, both assumed that the text was inherently corrupt and needed correction into the single version of a print edition. With the advent of the digital medium textual critics 'discovered that texts were more than simply correct or erroneous' (Shillingsburg, 2006: 81). The possibility of representing multiple versions or multiple markup perspectives has long been seen as an enticing prospect of the digital medium, but attempts to achieve this so far have led either to complexity that taxes the limitations of markup (Renear, 1997: 121) or to an overload of information and a 'drowning by versions' (Dalhstrom, 2000). It is now generally recognised that written texts can contain complexities and subtleties of structure that defeat the power of markup alone to represent them (Buzzetti, 2002).

In our talk we would like to present three examples of how overlapping structures can be efficiently edited in a wiki: of a modern genetic text in Italian, of a short classical text, the 'Sybilline Gospel', and of a short literary text marked up in various ways. Our purpose is to demonstrate the flexibility of the tool and the generality of the underlying algorithm, which is not designed for any one type of text or any one type of editing. However, because of the limited space, we will only describe here the Sibylline Gospel text. This is a particularly interesting example, because it not only has a complex manuscript tradition but it has also been deliberately altered throughout its history like agenetic text.

The Sibylla Tiburtina is an apocalyptic prophecy that describes the nine ages of world history up to the Last Judgement. The first version of this prophecy, which enjoyed a widespread and lasting popularity throughout the whole medieval era, was written in Greek in the second half of the 4th century. Of this lost text we have an edited Byzantine version, dating from the beginning of the 6th century, and several Latin versions, besides translations in Old French and German, in oriental languages (Syric, Arabic and Ethiopian) and in Slavic and Romanian.

The known Latin manuscripts number approximately 100, ranging in date from the mid 11th century to the beginning of the 16th. In the course of these centuries, with a particular concentration between the 11th and 12th centuries, the text was subjected to continuous revisions, in order to adapt it. This is demonstrated both by the changing names of eastern rulers mentioned in the ninth age, which involves the coming of a Last Roman Emperor and of the Antichrist, and by the introduction of more strictly theological aspects, especially in the so-called 'Sibylline Gospel', that is an account of Christ's life presented by the Sibyl under the fourth age. No critical edition of the Sibylla Tiburtina based on all its Latin versions has yet been produced, although this is the only way to unravel the genesis of the text and the history of its successive reworkings. A classical type of critical edition, however, would not be appropriate, nor would it make sense, to establish a critical text in the traditional sense as one 'cleaned of the errors accumulated in the course of its history and presented in a form most closely imagined to have been the original'. With this kind of representation one would have on the one hand the inconvenience of an unwieldy apparatus criticus and on the other, the serious inconsistency of a 'critical' text that never had any real existence.

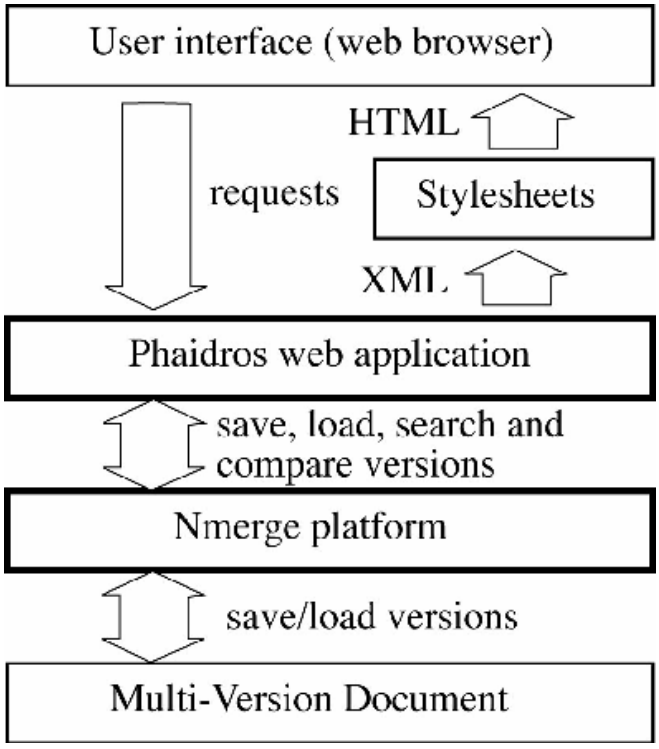
To handle cases like this, we are well advanced in the development of a multi-version document wiki application. The multi-version document (or MVD) concept is a further development of the variant graph model described at Digital Humanities 2006 and elsewhere (Schmidt and Fiormonte, 2007). In a nutshell the MVD format stores the text as a graph that accurately represents a single work in digital form, however many versions or markup perspectives it may be composed of. An MVD file is no mere aggregation of separate files; it is a single digital entity, within which versions may be efficiently searched and compared. It consists of three parts:

1) The variant-graph consisting of a list of the differences between the versions

2) A description of each of the versions, including a short name or siglum, e.g. 'L10', a longer name e.g. 'London, Brit. Lib. Cotton Titus D.III, saec. XIII' and a group name.

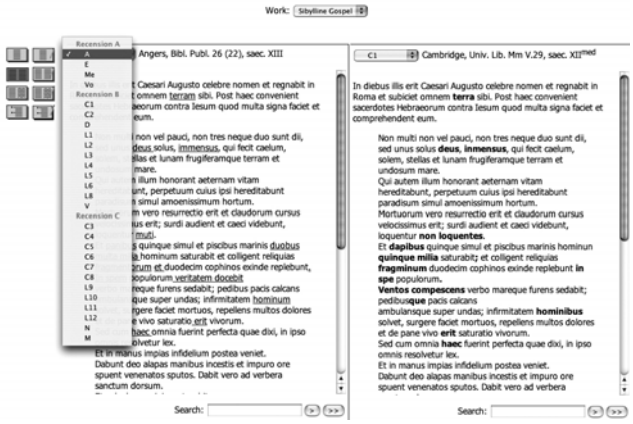
3) A list of groups. A group is a name for a group of versions or other groups. For example, in the Sybilline Gospel text there are three recensions, to which each of the manuscripts belong.

The wiki application consists of two JAVA packages (outlined in bold):



At the lower level the NMerge package implements all the functionality of the MVD format: the searching, comparison, retrieval and saving of versions. It can also export an MVD into a readable XML form. The text of each version is recorded in a simplified form of TEI-XML, but the MVD format does not rely on any form of markup, and is equally capable of handling binary file formats.

Building on this package, the Phaidros web application provides various forms for viewing and searching a multi-version document: for comparing two versions, for viewing the text alongside its manuscript facsimile, for reading a single version or for examining the list of versions. The user may also edit and save the XML content of each of these views. Since NMerge handles all of the overlapping structures, the markup required for each version can be very simple, as in a real wiki. In the drawing below the differences between the two versions are indicated by underlined or bold text.



This application is suitable for collaborative refinement of texts within a small group of researchers or by a wider public, and attempts to extend the idea of the wiki to new classes of text and to new classes of user.

References

Barbera, M. (2007) The HyperLearning Project: Towards a Distributed and Semantically Structured e-research and e-learning Platform. *Literary and Linguistic Computing*, 21(1), 77-82.

Buzzetti, D. (2002) Digital Representation and the Text Model. *New Literary History* 33(1), 61–88.

Dahlström, M. (2000) Drowning by Versions. Human IT 4. Retrieved 16/11/07 from http://www.hb.se/bhs/ith/4-00/md.htm

Renear, A. (1997) Out of Praxis: Three (Meta) Theories of Textuality. In *Electronic Text*, Sutherland, K. (Ed.) Clarendon Press, Oxford, pp. 107–126.

Robinson, P. (2007) Electronic editions which we have made and which we want to make. In A. Ciula and F. Stella (Eds), *Digital Philology and Medieval Texts*, Pisa, Pacini, pp. 1-12.

Schmidt D. and Fiormonte, D. (2007) Multi-Version Documents: a Digitisation Solution for Textual Cultural Heritage Artefacts. In G. Bordoni (Ed.), *Atti di AI*IA Workshop for Cultural Heritage*. 10th Congress of Italian Association for Artificial Intelligence, Università di Roma Tor Vergata, Villa Mondragone, 10 Sept., pp. 9-16.

Segre, C. (1981) Testo. In *Enciclopedia Einaudi* Vol. 14. Einaudi, Torino, pp. 269-291.

Shillingsburg, P. (2006) *From Gutenberg to Google*. Cambridge University Press, Cambridge.

Tanselle, G. (2006) Foreword in Burnard, L., O'Brien O'Keeffe, K. and Unsworth, J. (Eds.) *Electronic Textual Editing*. Text Encoding Initiative, New York and London.

Tapscott, D., and Williams, A. (2006) *Wikinomics: How Mass Collaboration Changes Everything.* Portfolio, New York.

Determining Value for Digital Humanities Tools

Susan Schreibman
sschreib@umd.edu
University of Maryland, USA

Ann Hanlon
ahanlon@umd.edu
University of Maryland, USA

Tool development in the Digital Humanities has been the subject of numerous articles and conference presentations (Arts and Humanities Research Council (AHRC), 2006; Bradley, 2003; McCarty, 2005; McGann, 2005; Ramsay, 2003, 2005; Schreibman, Hanlon, Daugherty, Ross, 2007; Schreibman, Kumar, McDonald, 2003; Summit on Digital Tools for the Humanities, 2005; Unsworth, 2003). While the purpose and direction of tools and tool development for the Digital Humanities has been discussed and debated in those forums, the value of tool development itself has seen little discussion. This is in part because tools are developed to aid and abet scholarship – they are not necessarily considered scholarship themselves. That perception may be changing, though. The clearest example of such a shift in thinking came from the recent recommendations of the *ACLS Commission on Cyberinfrastructure for the Humanities and Social Sciences*, which called not only for "policies for tenure and promotion that recognize and reward digital scholarship and scholarly communication" but likewise stated that "recognition should be given not only to scholarship that uses the humanities and social science cyberinfrastructure but also to scholarship that contributes to its design, construction and growth." On the other hand, the *MLA Report on Evaluating Scholarship for Tenure and Promotion* found that a majority of departments have little to no experience evaluating refereed articles and monographs in electronic format. The prospects for evaluating tool development as scholarship in those departments would appear dim. However, coupled with the more optimistic recommendations of the ACLS report, as well the MLA Report's findings that evaluation of work in digital form is gaining ground in some departments, the notion of tool development as a scholarly activity may not be far behind.

In 2005, scholars from the humanities as well as the social sciences and computer science met in Charlottesville, Virginia for a *Summit on Digital Tools for the Humanities*. While the summit itself focused primarily on the use of digital resources and digital tools for scholarship, the Report on Summit Accomplishments that followed touched on development, concluding that "the development of tools for the interpretation of digital evidence is itself research in the arts and humanities."

The goal of this paper is to demonstrate how the process of software or tool development itself can be considered the scholarly activity, and not solely a means to an end, i.e. a feature or interface for a content-based digital archive or repository. This paper will also deal with notions of value: both within the development community and as developers perceive how their home institutions and the community for which the software was developed value their work.

The data for this paper will be drawn from two sources. The first source is a survey carried out by the authors in 2007 on *The Versioning Machine*, the results of which were shared at a poster session at the 2007 Digital Humanities Conference. The results of this survey were intriguing, particularly in the area of the value of *The Versioning Machine* as a tool in which the vast majority of respondents found it valuable as a means to advance scholarship in spite of the fact that they themselves did not use it. As a result of feedback by the community to that poster session, the authors decided to conduct a focused survey on tool development as a scholarly activity.

After taking advice from several prominent Digital Humanities tool developers, the survey has been completed and will be issued in December to gather information on how faculty, programmers, web developers, and others working in the Digital Humanities perceive the value and purpose of their software and digital tool development activities. This paper will report the findings of a web-based survey that investigates how tools have been conceived, executed, received and used by the community. Additionally, the survey will investigate developers' perceptions of how tool development is valued in the academic community, both as a contribution to scholarship in their particular fields, and in relation to requirements for tenure and promotion.

The authors will use these surveys to take up John Unsworth's challenge, made at the 2007 *Digital Humanities Centers Summit* "to make our difficulties, the shortcomings of our tools, the challenges we haven't yet overcome, something that we actually talk about, analyze, and explicitly learn from." By examining not only how developers perceive their work, but how practitioners in the field use and value their tools, we intend to illuminate the role of tool development in the Digital Humanities and in the larger world of academia.

Bibliography

American Council of Learned Societies (ACLS), *Our Cultural Commonwealth: The Final Report of the American Council of Learned Societies Commission on Cyberinfrastructure for the Humanities & Social Sciences,* 2006, <http://www.acls.org/cyberinfrastructure/OurCulturalCommonwealth.pdf>

Arts and Humanities Research Council (AHRC), *AHRC ICT Methods Network Workgroup on Digital Tools Development for the Arts and Humanities,* 2006, <http://www.methnet.ac.uk/redist/pdf/wg1report.pdf>

Bradley, John. "Finding a Middle Ground between 'Determinism' and 'Aesthetic Indeterminacy': a Model for Text Analysis Tools." *Literary & Linguistic Computing* 18.2 (2003): 185-207.

Kenny, Anthony. "Keynote Address: Technology and Humanities Research." In *Scholarship and Technology in the Humanities: Proceedings of a Conference held at Elvetham Hall, Hampshire, UK, 9th-12th May 1990*, ed. May Katzen (London: Bowker Saur, 1991), 1-10.

McCarty, Willard. *Humanities Computing.* New York: Palgrave Macmillan, 2005.

McGann, Jerome. "Culture and Technology: The Way We Live Now, What Is to Be Done?" *New Literary History* 36.1 (2005): 71-82.

Modern Language Association of America, *Report of the MLA Task Force on Evaluating Scholarship for Tenure and Promotion*, December 2006, <http://www.mla.org/tenure_promotion_pdf>

Ramsay, Stephen. "In Praise of Pattern." *TEXT Technology* 14.2 (2005): 177-190.

Ramsay, Stephen. "Reconceiving Text Analysis: Toward an Algorithmic Criticism." *Literary and Linguistic Computing* 18.2 (2003): 167-174.

Schreibman, Susan and Ann Hanlon, Sean Daugherty, Tony Ross. "The Versioning Machine 3.1: Lessons in Open Source [Re]Development". Poster session at Digital Humanities (Urbana Champaign, June 2007)

Schreibman, Susan, Amit Kumar and Jarom McDonald. "The Versioning Machine". Proceedings of the 2002 ACH/ALLC Conference. *Literary and Linguistic Computing* 18.1 (2003) 101-107

Summit on Digital Tools for the Humanities (September 28-30, 2005), *Report on Summit Accomplishments*, May 2006. <http://www.iath.virginia.edu/dtsummit/SummitText.pdf>

Unsworth, John. "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, Saturday, January 18, 2003, Washington, D.C. <http://www.iath.virginia.edu/~jmu2m/carnegie-ninch.03.html>

Unsworth, John. "Digital Humanities Centers as Cyberinfrastructure", Digital Humanities Centers Summit, National Endowment for the Humanities, Washington, DC, Thursday, April 12, 2007 <http://www3.isrl.uiuc.edu/~unsworth/dhcs.html>

Recent work in the EDUCE Project

W. Brent Seales
seales@netlab.uky.edu
University of Kentucky, USA

A. Ross Scaife
scaife@gmail.com
University of Kentucky, USA

Popular methods for cultural heritage digitization include flatbed scanning and high-resolution photography. The current paradigm for the digitization and dissemination of library collections is to create high-resolution digital images as facsimiles of primary source materials. While these approaches have provided new levels of accessibility to many cultural artifacts, they usually assume that the object is more or less flat, or at least viewable in two dimensions. However, this assumption is simply not true for many cultural heritage objects.

For several years, researchers have been exploring digitization of documents beyond 2D imaging. Three-dimensional surface acquisition technology has been used to capture the shape of non-planar texts and build 3D document models (Brown 2000, Landon 2006), where structured light techniques are used to acquire 3D surface geometry and a high-resolution still camera is used to capture a 2D texture image.

However, there are many documents that are impossible to scan or photograph in the usual way. Take for example the entirety of the British Library's Cotton Collection, which was damaged by a fire in 1731 (Tite 1994, Prescott 1997, Seales 2000, Seales 2004). Following the fire most of the manuscripts in this collection suffered from both fire and water damage, and had to be physically dismantled and then painstakingly reassembled. Another example is the collection of papyrus scrolls that are in the Egyptian papyrus collection at the British Museum (Smith 1987, Andrews 1990, Lapp 1997). Because of damage to the outer shells of some of these scrolls the text enclosed therein have never been read - unrolling the fragile material would destroy them.

Certain objects can be physically repaired prior to digitization. However, the fact is that many items are simply too fragile to sustain physical restoration. As long as physical restoration is the only option for opening such opaque documents, scholarship has to sacrifice for preservation, or preservation sacrifice for scholarship.

The famous Herculaneum collection is an outstanding example of the scholarship/preservation dichotomy (Sider 2005). In the 1750s, the excavation of more than a thousand scorched papyrus rolls from the Villa dei Papiri (the Villa of the Papyri) in ancient Herculaneum caused great excitement among contemporary scholars, for they held the possibility

of the rediscovery of lost masterpieces by classical writers. However, as a result of the eruption of Mt. Vesuvius in A.D. 79 that destroyed Pompeii and also buried nearby Herculaneum, the papyrus rolls were charred severely and are now extremely brittle, frustrating attempts to open them.

A number of approaches have been devised to physically open the rolls, varying from mechanical and "caloric", to chemical (Sider 2005). Substrate breakages caused during the opening create incomplete letters that appear in several separate segments, which makes reading them very difficult. Some efforts have been made to reconstruct the scrolls, including tasks such as establishing the relative order of fragments and assigning to them an absolute sequence (Janko 2003). In addition, multi-spectral analysis and conventional image processing methods have helped to reveal significant previously unknown texts. All in all, although these attempts have introduced some new works to the canon, they have done so at the expense of the physical objects holding the text.

Given the huge amount of labor and care it takes to physically unroll the scrolls, together with the risk of destruction caused by the unrolling, a technology capable of producing a readable image of a rolled-up text without the need to physically open it is an attractive concept. Virtual unrolling would offer an obvious and substantial payoff.

In summary, there is a class of objects that are inaccessible due to their physical construction. Many of these objects may carry precise contents which will remain a mystery unless and until they are opened. In most cases, physical restoration is not an option because it is too risky, unpredictable, and labor intensive.

This dilemma is well suited for advanced computer vision techniques to provide a safe and efficient solution. The EDUCE project (Enhanced Digital Unwrapping for Conservation and Exploration) is developing a general restoration approach that enables access to those impenetrable objects without the need to open them. The vision is to apply this work ultimately to documents such as those described above, and to allow complete analysis while enforcing continued physical preservation.

Proof of Concept

With the assistance of curators from the Special Collections Library at the University of Michigan, we were given access to a manuscript from the 15th century that had been dismantled and used in the binding of a printed book soon after its creation. The manuscript is located in the spine of the binding, and consists of seven or so layers that were stuck together, as shown in figure 1. The handwritten text on the top layer is recognizable from the book of *Ecclesiastes*. The two columns of texts correspond to Eccl 2:4/2:5 (2:4 word 5 through 2:5 word 6) and Eccl 2:10 (word 10.5 through word 16). However, it was not clear what writing appears on the inner layers, or whether they contain any writing at all. We tested this manuscript using methods that we had refined over a series of simulations and real-world experiments, but this experiment was our first on a bona fide primary source.

Figure 1: Spine from a binding made of a fifteenth-century manuscript.

Following procedures which will be discussed in more detail in our presentation, we were able to bring out several layers of text, including the text on the back of the top layer. Figure 2 shows the result generated by our method to reveal the back side of the top layer which is glued inside and inaccessible. The left and right columns were identified as Eccl. 1:16 and 1:11 respectively.

To verify our findings, conservation specialists at the University of Michigan uncovered the back side of the top layer by removing the first layer from the rest of the strip of manuscript. The process was painstaking, in order to minimize damage, and it took an entire day. First, the strip was soaked in water for a couple of hours to dissolve the glue and enhance the flexibility of the material which was fragile due to age; this added a risk of the ink dissolving, although the duration and water temperature were controlled to protect against this happening. Then, the first layer was carefully pulled apart from the rest of the manuscript with tweezers. The process was very slow to avoid tearing the material. Remaining residue was scraped off gently.

The back side of the top layer is shown in figure 3. Most of the Hebrew characters in the images are legible and align well with those in the digital images of the manuscript. The middle rows show better results than the rows on the edges. That is because the edge areas were damaged in structure, torn and abraded, and that degraded the quality of the restoration.

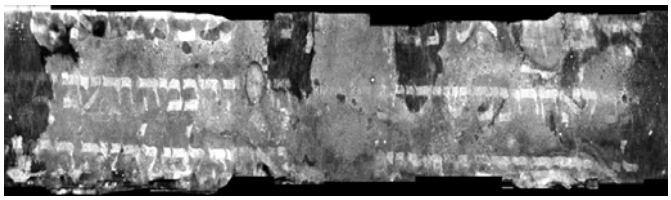


Figure 2: Generated result showing the back side of the top layer, identified as Eccl. 1:16 and 1:11.

Figure 3: Photo of the back side of the top layer, once removed from the binding.

Without applying the virtual unwrapping approach, the choices would be either to preserve the manuscript with the hidden text unknown or to destroy it to read it. In this case, we first read the text with non-invasive methods then disassembled the artifact in order to confirm our readings.

Presentation of Ongoing Work

In November 2007, representatives from the Sorbonne in Paris will bring several unopened papyrus fragments to the University of Kentucky to undergo testing following similar procedures to those that resulted in the uncovering of the Ecclesiastes text in the fifteenth-century manuscript. And in June 2008, a group from the University of Kentucky will be at the British Museum scanning and virtually unrolling examples from their collections of papyrus scrolls. This presentation at Digital Humanities 2008 will serve not only as an overview of the techniques that led to the successful test described above, but will also be an extremely up-to-date report of the most recent work of the project. We look forward to breaking down the dichotomy between preservation and scholarship for this particular class of delicate objects.

Bibliography

Andrews, C.A.R. *Catalogue of Demotic Papyri in the British Museum, IV: Ptolemaic Legal Texts from the Theban Area.* London: British Museum, 1990.

Brown, M. S. and Seales, W. B. "Beyond 2d images: Effective 3d imaging for library materials." In *Proceedings of the 5th ACM Conference on Digital Libraries* (June 2000): 27-36.

Brown, M. S. and Seales, W. B. "Image Restoration of Arbitrarily Warped Documents." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26:10 (October 2004): 1295-1306.

Janko, R. *Philodemus On Poems Book One*. Oxford University Press, 2003.

Landon, G.V. and Seales, W. B. "Petroglyph digitization: enabling cultural heritage scholarship." *Machine Vision and Applications*, 17:6 (December 2006): 361-371.

Lapp, G. *The Papyrus of Nu. Catalogue of Books of the Dead in the British Museum, vol. I.* London: British Museum, 1997.

Lin, Y. and Seales, W. B. "Opaque Document Imaging: Building Images of Inaccessible Texts." *Proceedings of the Tenth IEEE International Conference on Computer Vision (ICCV'05)*,

Prescott, A. "'Their Present Miserable State of Cremation': the Restoration of the Cotton Library." *Sir Robert Cotton as Collector: Essays on an Early Stuart Courtier and His Legacy.* Editor C.J. Wright. London: British Library Publications, 1997. 391-454.

Seales, W. B. Griffoen, J., Kiernan, K, Yuan, C. J., Cantara, L. "The digital atheneum: New technologies for restoring and preserving old documents." *Computers in Libraries* 20:2 (February 2000): 26-30.

Seales, W. B. and Lin, Y. "Digital restoration using volumetric scanning." In *Proceedings of the Fourth ACM/IEEE-CS Joint Conference on Digital Libraries* (June 2004): 117-124.

Sider, D. *The Library of the Villa dei Papiri at Herculaneum.* Getty Trust Publications: J. Paul Getty Museum, 2005.

Smith, M. *Catalogue of Demotic Papyri in the British Museum, III: the Mortuary Texts of Papyrus BM 10507.* London: British Museum, 1987.

Tite, C. G. *The Manuscript Library of Sir Robert Cotton.* London: British Library, 1994.

"It's a team if you use 'reply all'": An Exploration of Research Teams in Digital Humanities Environments

Lynne Siemens
siemensl@uvic.ca
University of Victoria, Canada

Introduction

This proposal reports on a research project exploring the nature of research teams in the Digital Humanities community. It is a start to understanding the type of supports and research preparation that individuals within the field require to successfully collaborate within research teams.

Context

Traditionally, research contributions in the humanities field have been felt to be, and documented to be, predominantly solo efforts by academics involving little direct collaboration with others, a model reinforced through doctoral studies and beyond (See, for example, Cuneo, 2003; Newell & Swan, 2000). However, Humanities Computing/Digital Humanities is an exception to this. Given that the nature of research work involves computers and a variety of skills and expertise, Digital Humanities researchers are working collaboratively within their institutions and with others nationally and internationallly to undertake the research. This research typically involves the need to coordinate efforts between academics, undergraduate and graduate students, research assistants, computer programmers, librarians, and other individuals as well as the need to coordinate financial and other resources. Despite this, there has been little formal research on team development within this community.

That said, efforts toward understanding the organizational context in which Digital Humanites research is situated is beginning in earnest. Two large-scale survey projects (Siemens et al., 2002; Toms et al., 2004) have highlighted issues of collaboration, among other topics, and Warwick (2004) found that the organizational context has had an impact on the manner in which Digital Humanities/Humanities Computing centres developed in the United States and England. Other studies are underway as well. In addition, McCarty (2005b) explores the ways that computers have opened the opportunities for collaboration within the humanities and has explored the associated challenges of collaboration and team research within the HUMANIST listserve (2005a). Finally, through efforts such as the University of Victoria's Digital Humanities/Humanities Computing Summer Institute and other similar ventures, the community is working to develop its collaborative capacity through workshops in topics like community-specific project management skills, which also includes discussion of team development and support.

This study draws upon these efforts as it explores and documents the nature of research teams within the Digital Humanities community to the end of identifying exemplary work patterns and larger models of research collaboration that have the potential to strengthen this positive aspect of the community even further.

Methods

This project uses a qualitative research approach with in-depth interviews with members of various multi-disciplinary, multi-location project teams in Canada, the United States, and the United Kingdom. The interview questions focus on the participants' definition of teams; their experiences working in teams; and the types of supports and research preparation required to ensure effective and efficient research results. The results will include a description of the community's work patterns and relationships and the identification of supports and research preparation required to sustain research teams (as per Marshall & Rossman, 1999; McCracken, 1988).

Preliminary Findings

At the time of writing this proposal, final data analysis is being completed, but clear patterns are emerging and, after final analysis, these will form the basis of my presentation.

The individuals interviewed currently are and have been a part of a diverse set of team research projects, in terms of research objective, team membership size, budget, and geographical dispersion, both within their own institution, nationally, and internationally. The roles they play are varied and include research assistant, researcher, computer programmer, and lead investigator. There are several commonalities among these individuals in terms of their skill development in team research and their definition of research teams and communities.

When final data analysis is complete, a series of exemplary patterns and models of research collaboration will be identified and outlined. These patterns and models will include the identification of supports and research preparation which can sustain research teams in the present and into the future.

The benefits to the Digital Humanities community will be several. First, the study contributes to an explicit description of the community's work patterns and relationships. Second, it also builds on previous efforts to understand the organizational context in which Digital Humanities/Humanities Computing centres operate in order to place focus on the role of the individual and teams in research success. Finally, it identifies possible supports and research preparation to aid the further development of successful research teams.

References

Cuneo, C. (2003, November). Interdisciplinary teams - let's make them work. *University Affairs,* 18-21.

Marshall, C., & Rossman, G. B. (1999). *Designing qualitative research* (3rd ed.). Thousand Oaks, California: SAGE Publications.

McCarty, W. (2005a). 19.215 how far collaboration? Humanist discussion group. Retrieved August 18, 2005, from http://lists.village.virginia.edu/lists_archive/Humanist/v19/0211.html

McCarty, W. (2005b). *Humanities computing.* New York, NY: Palgrave MacMillan.

McCracken, G. (1988). *The long interview* (Vol. 13). Newbury Park, CA: SAGE Publications.

Newell, S., & Swan, J. (2000). Trust and inter-organizational networking. *Human Relations*, 53(10), 1287-1328.

Siemens, R. G., Best, M., Grove-White, E., Burk, A., Kerr, J., Pope, A., et al. (2002). *The credibility of electronic publishing*: A report to the Humanities and Social Sciences Federation of Canada. Text Technology, 11(1), 1-128.

Toms, E., Rockwell, G., Sinclair, S., Siemens, R. G., & Siemens, L. (2004). The humanities scholar in the twenty-first century: How research is done and what support is needed, Joint International Conference of the Association for Computers and the Humanities and the Association for Literary & Linguistic Computing. Göteborg, Sweden. Published results forthcoming; results reported by Siemens, et al., in abstract available at http://www.hum.gu.se/allcach2004/AP/html/prop139.html.

Warwick, C. (2004). "No such thing as humanities computing?" an analytical history of digital resource creation and computing in the humanities, Joint International Conference of the Association for Computers and the Humanities and the Association for Literary & Linguistic Computing. Göteborg, Sweden. (Prepublication document available at http://tapor.humanities.mcmaster.ca/html/Nosuchthing_1.pdf.).

Doing Digital Scholarship

Lisa Spiro
lspiro@rice.edu
Rice University, USA

When I completed my dissertation *Bachelors of Arts: Bachelorhood and the Construction of Literary Identity in Antebellum America* in 2002, I figured that I was ahead of most of my peers in my use of digital resources, but I made no pretense of doing digital scholarship. I plumbed electronic text collections such as Making of America and Early American Fiction for references to bachelorhood, and I used simple text analysis tools to count the number of times words such as "bachelor" appeared in key texts. I even built an online critical edition of a section from Reveries of a Bachelor (http://etext.virginia.edu/users/spiro/Contents2.html), one of the central texts of sentimental bachelorhood. But in my determination to finish my PhD before gathering too many more gray hairs, I resisted the impulse to use more sophisticated analytical tools or to publish my dissertation online.

Five years later, the possibilities for digital scholarship in the humanities have grown. Projects such as TAPOR, Token-X, and MONK are constructing sophisticated tools for text analysis and visualization. Massive text digitization projects such as Google Books and the Open Content Alliance are making it possible to search thousands (if not millions) of books. NINES and other initiatives are building communities of digital humanities scholars, portals to content, and mechanisms for conducting peer review of digital scholarship. To encourage digital scholarship, the NEH recently launched a funding program. Meanwhile, scholars are blogging, putting up videos on YouTube, and using Web 2.0 tools to collaborate.

Despite this growth in digital scholarship, there are still too few examples of innovative projects that employ "digital collections and analytical tools to generate new intellectual products" (ACLS 7). As reports such as *A Kaleidoscope of American Literature and Our Cultural Commonwealth* suggest, the paucity of digital scholarship results from the lack of appropriate tools, technical skills, funding, and recognition. In a study of Dickinson, Whitman and Uncle Tom's Cabin scholars, my colleague Jane Segal and I found that although scholars are increasingly using digital resources in their research, they are essentially employing them to make traditional research practices more efficient and gain access to more resources, not (yet) to transform their research methodology by employing new tools and processes (http://library.rice.edu/services/digital_media_center/projects/the-impact-of-digital-resources-on-humanities-research). What does it mean to do humanities research in a Web 2.0 world? To what extent do existing tools, resources, and research methods support digital scholarship, and what else do scholars need?

To investigate these questions, I am revisiting my dissertation to re-imagine and re-mix it as digital scholarship. I aim not only to open up new insights into my primary research area--the significance of bachelorhood in nineteenth-century American culture--but also to document and analyze emerging methods for conducting research in the digital environment. To what extent do digital tools and resources enable new approaches to traditional research questions—and to what extent are entirely new research questions and methods enabled? I am structuring my research around what John Unsworth calls the "scholarly primitives," or core research practices in the humanities:

1. **Discovering:** To determine how much information is available online, I have searched for the nearly 300 resources cited in my dissertation in Google Books, Making of America, and other web sites. I found that 77% of my primary source resources and 22% of my secondary sources are available online as full-text, while 92% of all my research materials have been digitized (this number includes works available through Google Books as limited preview, snippet view, and no preview.) Although most nineteenth-century books cited in my dissertation are now freely available online, many archival resources and periodicals have not yet been digitized.

2. **Annotating**: In the past, I kept research notes in long, unwieldy Word documents, which made it hard to find information that I needed. New software such as Zotero enables researchers to store copies of the digital resources and to make annotations as part of the metadata record. What effect does the ability to share and annotate resources have on research practices? How useful is tagging as a mechanism for organizing information?

3. **Comparing**: Through text analysis and collation software such as Juxta and TAPOR, scholars can compare different versions of texts and detect patterns. Likewise, the Virtual Lightbox allows researchers to compare and manipulate digital images. What kind of new insights can be generated by using these tools? In the course of doing my research, I am testing freely available tools and evaluating their usefulness for my project.

4. **Referring**: With digital publications, we not only can refer to prior work, but link to it, even embed it. What is the best means for constructing a scholarly apparatus in digital scholarship, particularly in a work focused not only on making an argument, but also on examining the process that shaped that argument?

5. **Sampling**: With so much information available, what criteria should we use to determine what to focus on? Since not everything is digitized and search engines can be blunt instruments, what do we ignore by relying mainly on digital resources? In my blog, I am reflecting on the selection criteria used to produce the arguments in my revamped dissertation.

6. **Illustrating**: What kind of evidence do we use to build an argument in a work of digital scholarship, and how is that evidence presented? In my dissertation, I generalized about the significance of bachelorhood in American literature by performing close readings of a few key texts, but such a method was admittedly unsystematic. By using text analysis tools to study a much larger sample of primary texts, I can cite statistics such as word frequency in making my argument--but does this make my argument any more convincing?

7. **Representing**: How should a work of digital scholarship be presented? Ideally readers would be able to examine the evidence for themselves and even perform their own queries. At the same time, information should be offered so that it is clear and consistent with familiar academic discourse. How should I make available not only research conclusions, but also the detailed research process that undergirds these conclusions--the successful and unsuccessful searches, the queries run in text analysis software, the insights offered by collaborators? How will the digital work compare to the more traditional original dissertation? What kind of tools (for instance, ifBook's Sophie) will be used to author the work?

In addition to Unsworth's list, I offer two more:

8. **Collaborating**: Although humanities scholars are thought to be solitary, they collaborate frequently by exchanging bibliographic references and drafts of their essays. How do I engage the community in my research? I am encouraging others to comment on my (re-) work in progress (http://digitalhumanities.edublogs.org/) using Comment Press. Moreover, I am bookmarking all web-based sources for my study on delicious (http://del.icio.us/lms4w/digital_scholarship) and making available feeds from my various research sources through a PageFlakes portal (http://www.pageflakes.com/lspiro/). On my blog, "Digital Scholarship in the Humanities," I explore issues and ideas raised by my research (http://digitalscholarship.wordpress.com/). I am examining what it takes to build an audience and how visibility and collaboration affect my research practices.

9. **Remixing**: What would it mean to take an earlier work--my own dissertation, for example--use new sources and approaches, and present it in a new form? What constitutes a scholarly remix, and what are the implications for intellectual property and academic ethics? I also plan to experiment with mashups as a means of generating and presenting new insights, such as a Google Map plotting census statistics about antebellum bachelors or a visual mashup of images of bachelors.

This project examines the process of doing research digitally, the capabilities and limits of existing tools and resources, and the best means of authoring, representing and disseminating digital scholarship. I aim to make this process as open, visible, and collaborative as possible. My presentation will focus

on emerging research methodologies in the humanities, particularly the use of tools to analyze and organize information, the development of protocols for searching and selecting resources, and the dissemination of ideas through blogs and multimedia publication.

References

American Council of Learned Societies (ACLS). *Our Cultural Commonwealth: The Report of the American Council of Learned Societies Commission on Cyberinfrastructure for the Humanities and Social Sciences*. New York: American Council of Learned Societies, 2006.

Brogan, Martha. *A Kaleidoscope of Digital American Literature. Digital Library Federation*. 2005. 22 May 2007 <http://www.diglib.org/pubs/dlf104/>.

Unsworth, John. "Scholarly Primitives: what methods do humanities researchers have in common, and how might our tools reflect this?' 13 May 2000. 20 November 2007. <http://jefferson.village.virginia.edu/~jmu2m/Kings.5-00/primitives.html>

Extracting author-specific expressions using random forest for use in the sociolinguistic analysis of political speeches

Takafumi Suzuki
qq16116@iii.u-tokyo.ac.jp
University of Tokyo, Japan

Introduction

This study applies stylistic text classification using random forest to extract author-specific expressions for use in the sociolinguistic analysis of political speeches. In the field of politics, the style of political leaders' speeches, as well as their content, has attracted growing attention in both English (Ahren, 2005) and Japanese (Azuma, 2006; Suzuki and Kageura, 2006). One of the main purposes of these studies is to investigate political leaders' individual and political styles by analyzing their speech styles. A common problem of many of these studies and also many sociolinguistic studies is that the expressions that are analyzed are selected solely on the basis of the researcher's interests or preferences, which can sometimes lead to contradictory interpretations. In other words, it is difficult to determine whether these kind of analyses have in fact correctly identified political leaders' individual speech styles and, on this basis, correctly characterised their individual and political styles. Another problem is that political leaders' speech styles may also be characterised by the infrequent use of specific expressions, but this is rarely focused on.

In order to solve these problems, we decided to apply stylistic text classification and feature extraction using random forest to political speeches. By classifying the texts of an author according to their style and extracting the variables contributing to this classification, we can identify the expressions specific to that author. This enables us to determine his/her speech style, including the infrequent use of specific expressions, and characterise his/her individual and political style. This method can be used for the sociolinguistic analysis of various types of texts, which is, according to Argamon et al. (2007a), a potentially important area of application for stylistics.

Experimental setup

We selected the Diet addresses of two Japanese prime ministers, Nakasone and Koizumi, and performed two classification experiments: we distinguished Nakasone's addresses from those of his 3 contemporaries (1980-1989), and Koizumi's addresses from those of his 8 contemporaries (1989-2006). Because Nakasone and Koizumi were two of the most powerful prime ministers in the history of Japanese politics, and took a special interest in the content or style of

their own speeches (Suzuki and Kageura, 2007), their addresses are the most appropriate candidates for initial analysis. Suzuki and Kageura (2006) have demonstrated that the style of Japanese prime ministers' addresses has changed significantly over time, so we compared their addresses with those of their respective contemporaries selected by the standard division of eras in Japanese political history. We downloaded the addresses from the online database *Sekai to Nihon* (*The World and Japan*) (www.ioc.u-tokyo.ac.jp/~worldjpn), and applied morphological analysis to the addresses using ChaSen (Matsumoto et al., 2003). We united notational differences which were distinguished only by kanji and kana in Japanese. Table 1 sets out the number of addresses and the total number of tokens and types for all words, particles and auxiliary verbs in each category.

		all words		particles		auxiliary verbs	
	addresses	tokens	types	tokens	types	tokens	types
Nakasone	10	47419	3390	13851	70	4241	33
1980-1989	9	35979	2862	10656	64	3434	29
Koizumi	11	46979	3908	13677	62	3680	35
1989-2006	31	149234	5774	44712	80	13292	43

Table 1. Basic data on the corpora

As a classification method, we selected the random forest (RF) method proposed by Breiman (2001), and evaluated the results using out-of-bag tests. RF is known to perform extremely well in classification tasks in other fields using large amounts of data, but to date few studies have used this method in the area of stylistics (Jin and Murakami, 2006). Our first aim in using RF was thus to test its effectiveness in stylistic text classification.

A second and more important aim was to extract the important variables contributing to classification (which are shown as high Gini coefficients). The extracted variables represent the specific expressions distinguishing the author from others, and they can show the author's special preference or dislike for specific expressions. Examining these extracted expressions enables us to determine the author's speech style and characterise his/her individual and political styles. In an analogous study, Argamon et al. (2007b) performed gender-based classification and feature extraction using SVM and information gain, but as they are separate experiments and RF returns relevant variables contributing to classification, RF is more suitable for our purposes.

We decided to focus on the distribution of particles and auxiliary verbs because they represent the modality of the texts, information representing authors' personality and sentiments (Otsuka et al., 2007), and are regarded as reflecting political leaders' individual and political styles clearly in Japanese (Azuma, 2006, Suzuki and Kageura, 2006). We tested 8 distribution combinations as features (see Table 2). Though Jin (1997) has demonstrated that the distribution of particles (1st order part-of-speech tag) is a good indicator of the author in Japanese, the performances of these features, especially auxiliary verbs, have not been explored fully, and as the microscopic differences in features (order of part-of-speech and stemming) can affect the classification accuracy, we decided to test the 8 combinations.

Results and discussion

Table 2 shows the precision, recall rates and *F*-values. Koizumi displayed higher accuracy than Nakasone, partly because he had a more individualistic style of speech (see also Figure 2 and 3), and partly because a larger number of texts were used in his case. Many of the feature sets show high classification accuracy (more than 70%) according to the criteria of an analogous study (Argamon et al., 2007c), which confirms the high performance of RF. The results also show that the distribution of the auxiliary verbs and combinations can give better performance than that of the particles used in a previous study (Jin, 1997), and also that stemming and a deeper order of part-of-speech can improve the results.

		Nakasone			Koizumi		
		precision	recall rate	F_1-value*	precision	recall rate	F_1-value*
particles	(1st order)†	70.0	70.0	70.0	50.0	18.2	26.7
particles	(2nd order)‡	70.0	70.0	70.0	83.3	45.5	58.8
auxiliary verbs	(without stemming)	66.7	60.0	63.2	100.0	81.8	90.0
auxiliary verbs	(with stemming)	77.8	70.0	73.7	100.0	90.9	95.2
combination	(1st order, without stemming)	54.5	60.0	57.1	100.0	81.8	90.0
combination	(2nd order, without stemming)	61.5	80.0	69.6	100.0	81.8	90.0
combination	(1st order, with stemming)	70.0	70.0	70.0	100.0	90.9	95.2
combination	(2nd order, with stemming)	77.8	70.0	73.7	100.0	90.9	95.2

* F_β-value $= \frac{(\beta^2+1)PR}{\beta^2 P+R}$ (P: precision, R: recall rate)
† 1st order part-of speech tag ('particle')
‡ 2nd order part-of speech tag ('case particles', 'conjunctive particles' etc.)

Table 2. Precisions and recall rates

Figure 1 represents the top twenty variables with high Gini coefficients according to the classification of combinations of features (2nd order and with stemming). The figure indicates that several top variables had an especially important role in classification. In order to examine them in detail, we plotted in Figure 2 and 3 the transitions in the relative frequencies of the top four variables in the addresses of all prime ministers after World War 2. These include variables representing politeness ('masu', 'aru', 'desu'), assertion ('da', 'desu'), normativeness ('beshi'), and intention ('u'), and show the individual and political styles of these two prime ministers well. For example, 'aru' is a typical expression used in formal speeches, and also Diet addresses (Azuma, 2006), and the fact that Koizumi used this expression more infrequently than any other prime minister indicates his approachable speech style and can explain his political success. Also, the figure shows that we can extract the expressions that Nakasone and Koizumi used less frequently than their contemporaries as well as the expressions they used frequently. These results show the effectiveness of RF feature extraction for the sociolinguistic analysis of political speeches.

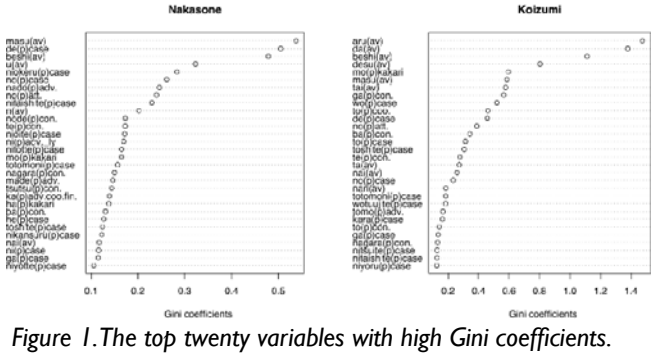


Figure 1. The top twenty variables with high Gini coefficients. The notations of the variables indicate the name of part-of-speech (p: particle, av: auxiliary verb) followed by (in the case of particles) the 2nd order part-of speech.

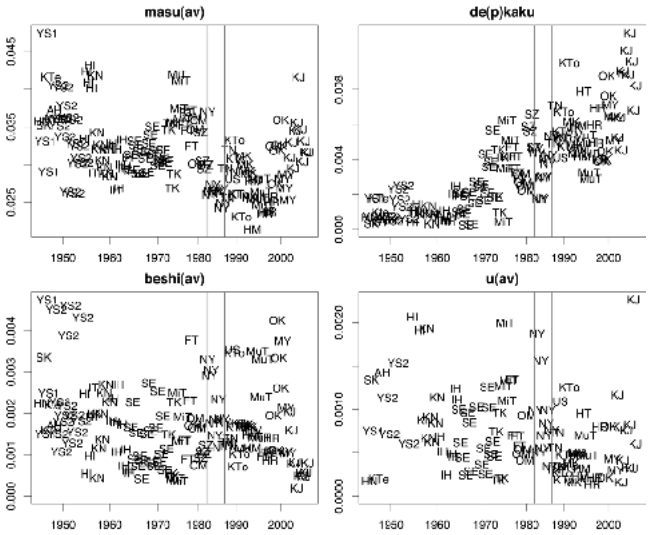


Figure 2. Transitions in the relative frequencies of the top four variables (Nakasone's case) in the addresses of all Japanese prime ministers from 1945 to 2006. The 'NY's between the red lines represent addresses by Nakasone, and other initials represent addresses by the prime minister with those initials (see Suzuki and Kageura, 2007).

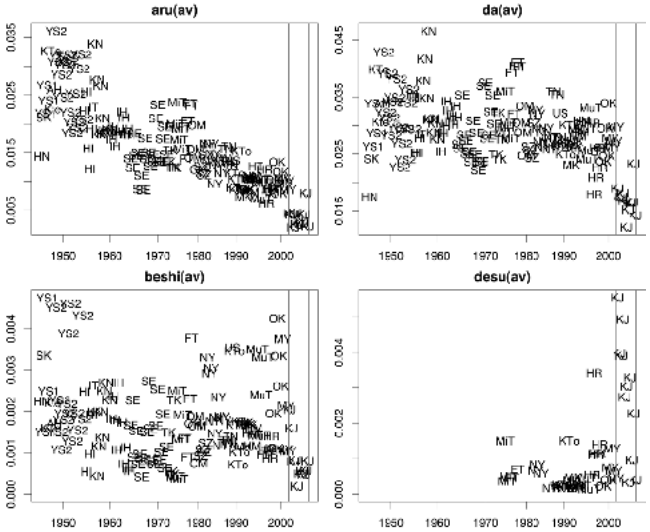


Figure 3. Transitions in the relative frequencies of the top four variables (Koizumi's case) in the addresses of all Japanese prime inisters from 1945 to 2006. The 'KJ's between the red lines represent addresses by Koizumi, and the other initials represent addresses by the prime minister with those initials (see Suzuki and Kageura, 2007).

Conclusion

This study applied text classification and feature extraction using random forest for use in the sociolinguistic analysis of political speeches. We showed that a relatively new method in stylistics performs fairly well, and enables us to extract author-specific expressions. In this way, we can systematically determine the expressions that should be analyzed to characterise their individual and political styles. This method can be used for the sociolinguistic analysis of various types of texts, which will contribute to further expansion in the scope of stylistics. A further study will include more concrete analysis of extracted expressions.

References

Ahren, K. (2005): People in the State of the Union: viewing social change through the eyes of presidents, *Proceedings of PACLIC 19, the 19th Asia-Pacific Conference on Language, Information and Computation*, 43-50.

Argamon, S. et al. (2007a): Stylistic text classification using functional lexical features, *Journal of the American Society for Information Science and Technology*, 58(6), 802-822.

Argamon, S. et al. (2007b): Discourse, power, and Ecriture feminine: the text mining gender difference in 18th and 19th century French literature, *Abstracts of Digital Humanities*, 191.

Argamon, S. et al. (2007c): Gender, race, and nationality in Black drama, 1850-2000: mining differences in language use in authors and their characters, *Abstracts of Digital Humanities*, 149.

Azuma, S. (2006): *Rekidai Syusyo no Gengoryoku wo Shindan Suru*, Kenkyu-sya, Tokyo.

Breiman, L. (2001): Random forests, *Machine Learning*, 45, 5-23.

Jin, M. (1997): Determining the writer of diary based on distribution of particles, *Mathematical Linguistics*, 20(8), 357-367.

Jin, M. and M. Murakami (2006): Authorship identification with ensemble learning, *Proceedings of IPSJ SIG Computers and the Humanities Symposium 2006*, 137-144.

Matsumoto, Y. et al. (2003): *Japanese Morphological Analysis System ChaSen ver.2.2.3*, chasen.naist.jp.

Otsuka, H. et al. (2007): *Iken Bunseki Enjin*, Corona Publishing, Tokyo, 127-172.

Suzuki, T. and K. Kageura (2006): The stylistic changes of Japanese prime ministers' addresses over time, *Proceedings of IPSJ SIG Computers and the Humanities Symposium 2006*, 145-152.

Suzuki, T. and K. Kageura (2007): Exploring the microscopic textual characteristics of Japanese prime ministers' Diet addresses by measuring the quantity and diversity of nouns, *Proceedings of PACLIC21, the 21th Asia-Pacific Conference on Language, Information and Computation*, 459-470.

Gentleman in Dickens: A Multivariate Stylometric Approach to its Collocation

Tomoji Tabata
tabata@lang.osaka-u.ac.jp
University of Osaka, Japan

This study proposes a multivariate approach to the collocation of *gentleman* in Dickens. By applying a stylo-statistical analysis model based on correspondence analysis (Tabata, 2004, 2005, 2007a, and 2007b) to the investigation of the collocation of the word *gentleman*, the present study visualizes the complex interrelationships among *gentleman*'s collocates, interrelationships among texts, and the association patterns between the *gentleman*'s collocates and texts in multi-dimensional spaces. By so doing, I shall illustrate how the collocational patterns of *gentleman* reflects a stylistic variation over time as well as the stylistic fingerprint of the author.

One thing that strikes readers of Dickens in terms of his style is the ways he combines words together (i.e. collocation). Dickens sometimes combines words in a quite unique manner, in a way that contradicts our expectation to strike humour, to imply irony or satire, or otherwise. Dickens also uses fixed/repetitive collocations for characterization: to make particular characters memorable by making them stand out from others. The collocation of word *gentleman*, one of the most frequent 'content' words in Dickens, is full of such examples, and would therefore be worthy of stylistic investigation.

The significance of collocation in stylistic studies was first suggested by J. R. Firth (1957) more than half a century ago. Thanks to the efforts by Sinclair (1991) and his followers, who have developed empirical methodologies based on large-scale corpora, the study of collocation has become widely acknowledged as an important area of linguistics. However, the vast majority of collocation studies (Sinclair, 1991; Kjellmer, 1994; Stubbs, 1995; Hunston and Francis, 1999, to name but a few) have been concerned with grammar/syntax, lexicography, and language pedagogy, not with stylistic aspects of collocation, by showing the company a word keeps. Notable exceptions are Louw (1993), Adolphs and Carter (2003), Hori (2004), and Partington (2006). The emergence of these recent studies seems to indicate that collocation is beginning to draw increasing attention from stylisticians.

Major tools used for collocation studies are KWIC concordances and statistical measures, which have been designed to quantify collocational strength between words in texts. With the wide availability of highly functional concordancers, it is now conventional practice to examine collocates in a span of, say, four words to the left, and four words to the right, of the node (target word/key word) with the help of statistics for filtering out unimportant information. This conventional approach makes it possible to detect a grammatical/syntactic, phraseological, and/or semantic relation between the node and its collocates. However, a conventional approach would not be suitable if one wanted to explore whether the company a word keeps would remain unchanged, or its collocation would change over time or across texts. It is unsuitable because this would involve too many variables to process with a concordancer to include a diachronic or a cross-textual/authorial perspective when retrieving collocational information from a large set of texts. Such an enterprise would require a multivariate approach.

Various multivariate analyses of texts have been successful in elucidating linguistic variation over time, variation across registers, variation across oceans, to say nothing of linguistic differences between authors (Brainerd, 1980; Burrows, 1987 & 1996; Biber and Finegan, 1992; Craig, 1999a, b, & c; Hoover, 2003a, b, & c; Rudman, 2005). My earlier attempts used correspondence analysis to accommodate low frequency variables (words) in profiling authorial/chronological/cross-register variations in Dickens and Smollett (Tabata, 2005, 2007a, & c). Given the fact that most collocates of content words tend to be low in frequency, my methodology based on correspondence analysis would usefully be applied to a macroscopic analysis of collocation of *gentleman*.

This study uses Smollett's texts as a control set against which the Dickens data is compared, in keeping with my earlier investigations. Dickens and Smollett stand in contrast in the frequency of *gentleman*. In 23 Dickens texts used in this study, the number of tokens for *gentleman* amounts to 4,547, whereas Smollett employs them 797 times in his seven works. In the normalised frequency scale per million words, the frequency in Dickens is 961.2, while the frequency in Smollett is 714.3. However, if one compares the first seven Dickens texts with the Smollett set, the discrepancy is even greater: 1792.0 versus 714.33 per million words. The word *gentleman* is significantly more frequent in early Dickens than in his later works.

Stubbs (2001: 29) states that "[t]here is some consensus, but no total agreement, that significant collocates are usually found within a span of 4:4". Following the conventional practice to examine collocation (Sinclaire, 1991; Stubbs, 1995 & 2001), the present study deals with words occurring within a span of four words prior to, and four words following, the node (*gentleman*) as variables (collocates) to be fed into correspondence analysis.

The respective frequency of each collocate is arrayed to form the frequency-profile for 29 texts (Smollett's The Adventure of an Atom is left out of the data set since it contains no instance of *gentleman*). The set of 29 collocate frequency profiles (collocate frequency matrix) is then transposed and submitted to correspondence analysis (CA), a technique for data-reduction. CA allows examination of the complex interrelationships between row cases (i.e., texts), interrelationships between column variables (i.e., collocates), and association between the row cases and column variables graphically in a multi-dimensional space. It computes the row coordinates (word

scores) and column coordinates (text scores) in a way that permutes the original data matrix so that the correlation between the word variables and text profiles are maximized. In a permuted data matrix, adverbs with a similar pattern of distribution make the closest neighbours, and so do texts of similar profile. When the row/column scores are projected in multi-dimensional charts, relative distance between variable entries indicates affinity, similarity, association, or otherwise between them.

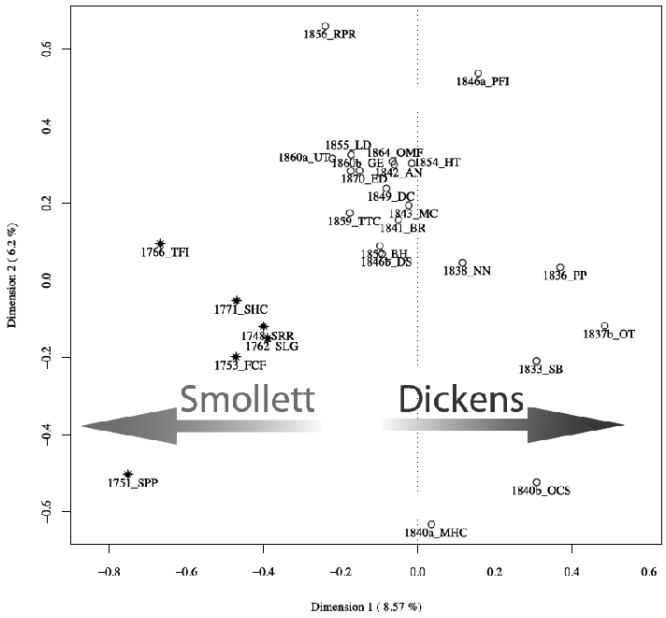


Figure 1. Correspondance analysis of the collocates of gentleman*: Text-map*

Figures 1 and 2 demonstrate a result of correspondence analysis based on 1,074 collocates of *gentleman* across 29 texts. The horizontal axis of Figure 1 labelled as Dimension 1, the most powerful axis, visualizes the difference between the two authors in the distribution of *gentleman*'s collocates. It is also interesting that the early Dickensian texts, written in 1830s and early 1840s, are lying towards the bottom half of the diagram. The same holds true for Smollett. Thus, the horizontal axis can be interpreted as indicating authorial variation in the collocation of *gentleman*, whereas the vertical axis can be interpreted as representing variation over time, although text entries do not find themselves in an exact chronological order. On the other hand, Figure 2 is too densely populated to identify each collocate, except for the outlying collocates, collocates with stronger "pulling power". However, it would be possible to make up for this by inspecting a diagram derived from much smaller number of variables (say, 100 variables, which will be shown shortly as Figure 4), whose overall configuration is remarkably similar to Figure 2 despite the decrease in the number of variables computed.

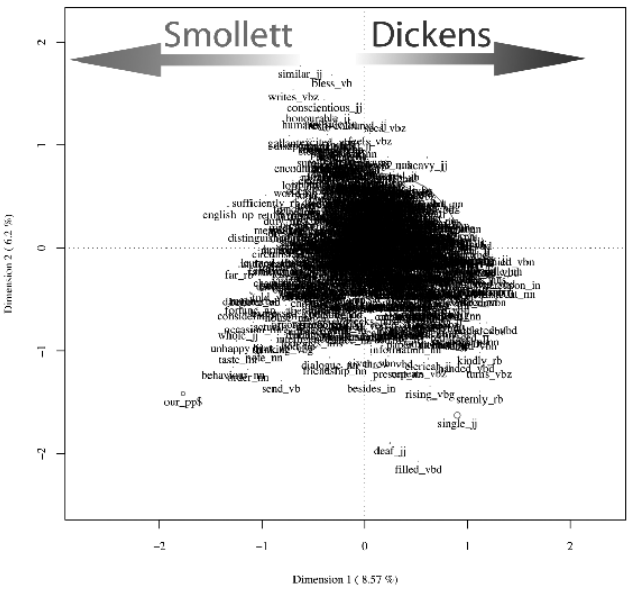


Figure 2. Correspondence analysis of the collocates of gentleman*: A galaxy of collocates*

The Dickens corpus is more than four times the size of the Smollett corpus, and the number of types as well as tokens of *gentleman*'s collocates in Dickens is more than four times as many as those in Smollett. It is necessary to ensure that a size factor does not come into play in the outcome of analysis. Figures 3 and 4 are derived from the variables of 100 collocates common to both authors. Despite the decrease in the number of variables from 1,074 to 100, the configuration of texts and words is remarkably similar to that based on 1,074 items. The Dickens set and the Smollett set, once again, can be distinguished from each other along Dimension 1. Moreover, in each of the two authors' sets, early works tend to have lower scores with later works scoring higher along Dimension 2. The results of the present analysis is consistent with my earlier studies based on different variables, such as *–ly* adverbs, superlatives, as well as high-frequency function words. It would be possible to assume that authorial fingerprints are as firmly set in the collocation of *gentleman* as in other component of vocabulary. These results seem to illustrate multivariate analysis of collocates could provide interesting new perspectives to the study of collocation.

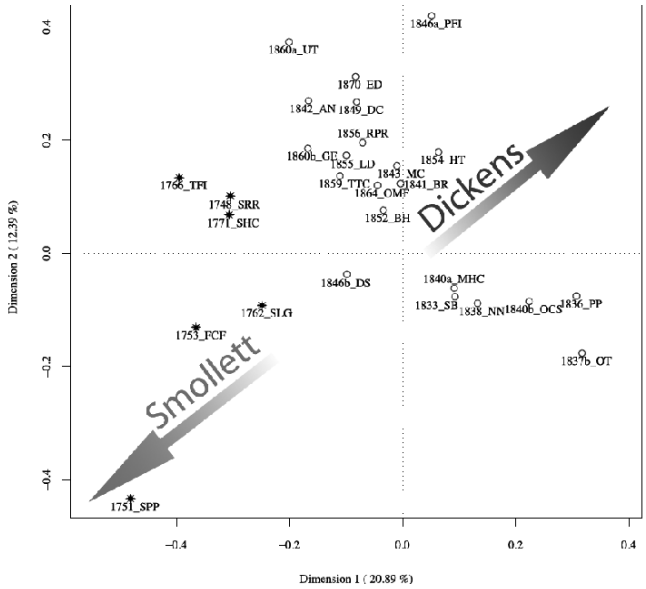


Figure 3. Correspondence analysis of 100 most common collocates of gentleman*: Text-map*

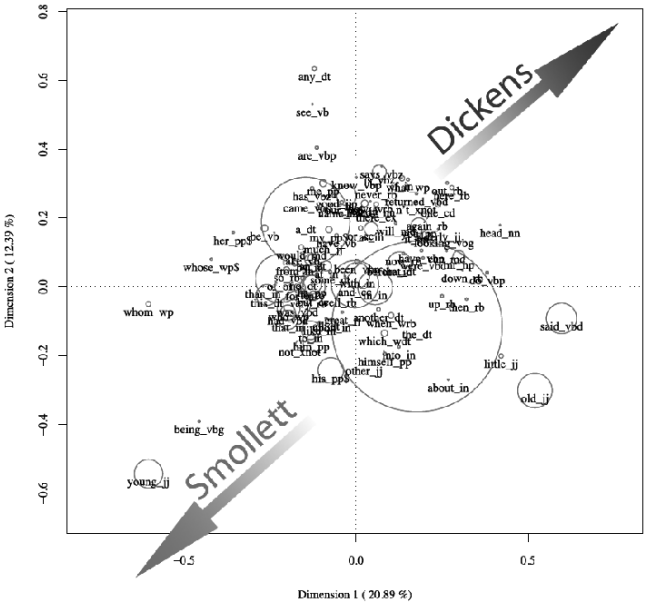


Figure 4. Correspondance analysis of 100 most common collocates of gentleman*: Word-map of 100 collocates*

References

Adolphs, S. and R. A. Carter (2003) 'Corpus stylistics: point of view and semantic prosodies in *To The Lighthouse*', *Poetica*, 58: 7–20.

Biber, D. and E. Finegan (1992) 'The Linguistic Evolution of Five Written and Speech-Based English Genres from the 17th to the 20th Centuries,' in M. Rissanen (ed.) *History of Englishes: New Methods and Interpretation in Historical Linguistics*. Berlin/New York: Mouton de Gruyter. 668–704.

Brainerd, B. (1980) 'The Chronology of Shakespeare's Plays: A Statistical Study,' *Computers and the Humanities*, 14: 221–230.

Burrows, J. F. (1987) *Computation into Criticism: A study of Jane Austen's novels and an experiment in method*. Oxford: Clarendon Press.

Burrows, J. F. (1996) 'Tiptoeing into the Infinite: Testing for Evidence of National Differences in the Language of English Narrative', in S. Hockey and N. Ide (eds.) *Research in Humanities Computing 4*. Oxford/New York: Oxford UP. 1–33.

Craig, D. H. (1999a) 'Johnsonian chronology and the styles of A Tale of a Tub,' in M. Butler (ed.) *Re-Presenting Ben Jonson: Text Performance, History*. London: Macmillan, 210–232.

Craig, D. H. (1999b) 'Authorial Attribution and Computational Stylistics: If You Can Tell Authors Apart, Have You Learned Anything About Them?' *Literary and Linguistic Computing*, 14: 103–113.

Craig, D. H. (1999c) 'Contrast and Change in the Idiolects of Ben Jonson Characters,' *Computers and the Humanities*, 33. 3: 221–240.

Firth, J. R. (1957) 'Modes of Meaning', in *Papers in Linguistics 1934–51*. London: OUP. 191-215.

Hoover, D. L. (2003a) 'Frequent Collocations and Authorial Style,' *Literary and Linguistic Computing*, 18: 261–286.

Hoover, D. L. (2003b) 'Multivariate Analysis and the Study of Style Variation,' *Literary and Linguistic Computing*, 18: 341–360.

Hori, M. (2004) *Investigating Dickens' style: A Collocational Analysis*. New York: Palgrave Macmillan.

Hunston, S. and G. Framcis (1999) *Pattern Grammar: A Corpus-Driven Approach to the Lexical Grammar of English*. Amsterdam: John Benjamins.

Kjellmer, G. (1994) *A Dictionary of English Collocations: Based on the Brown Corpus (3 vols.)*. Oxford: Clarendon Press.

Louw, W. (1993) 'Irony in the text or insincerity in the writer? The diagnostic potential of semantic prosodies', reprinted in G. Sampson and D. McCarthy (eds.) (2004) *Corpus Linguistics:*

Readings in a Widening Discipline. London and New York: Continuum. 229–241.

Partington, A. (2006) *The Linguistics of Laughter: A Corpus-Assisted Study of Laughter-Talk*. London/New York: Routledge.

Rudman, J. (2005) 'The Non-Traditional Case for the Authorship of the Twelve Disputed "Federalist" Papers: A Monument Built on Sand?", *ACH/ALLC 2005 Conference Abstracts*, Humanities Computing and Media Centre, University of Victoria, Canada, 193–196.

Sinclair, J. (1991) *Corpus, Concordance, Collocation*. Oxford: OUP.

Stubbs, M. (1995) 'Corpus evidence for norms of lexical collocation', in G. Cook and B. Seidlhofer (eds.) *Principle and Practice in Applied Linguistics: Studies in Honour of H. G. Widdowson*. Oxford: OUP. 243–256.

Stubbs, M. (2001) *Words and Phrases: Corpus studies of lexical semantics*. Oxford: Blackwell.

Tabata, T. (2004) 'Differentiation of Idiolects in Fictional Discourse: A Stylo-Statistical Approach to Dickens's Artistry', in R. Hiltunen and S. Watanabe (eds.) *Approaches to Style and Discourse in English*. Osaka: Osaka University Press. 79–106.

Tabata, T. (2005) 'Profiling stylistic variations in Dickens and Smollett through correspondence analysis of low frequency words', *ACH/ALLC 2005 Conference Abstracts*, Humanities Computing and Media Centre, University of Victoria, Canada, 229–232.

Tabata, T. (2007a) 'A Statistical Study of Superlatives in Dickens and Smollett: A Case Study in Corpus Stylistics', *Digital Humanities 2007 Conference Abstracts, The 19th Joint International Conference of the Association for Computers and the Humanities and the Association for Literary and Linguistic Computing, University of Illinois, Urbana-Champaign, June 4–June 8, 2007*, 210–214.

Tabata, T. (2007b) 'The *Cunningest, Rummest, Superlativest* Old Fox: A multivariate approach to superlatives in Dickens and Smollett ', *PALA 2007—Style and Communication—Conference Abstracts, The 2007 Annual Conference of the Poetics and Linguistics Association, 31 Kansai Gaidai University, Hirakata, Osaka, Japan, July–4 August 2007*, 60–61.

Video Game Avatar: From Other to Self-Transcendence and Transformation

Mary L. Tripp
mtripp@mail.ucf.edu
University of Central Florida, USA

The aim of this paper is to establish a model for the relationship between the player of a video game and the avatar that represents that player in the virtual world of the video game. I propose that there is an evolution in the identification of the player of a video game with the avatar character that performs embodied actions in the virtual world of the game. This identification can be described through an examination of theories from a variety of subject areas including philosophy, literary studies, video game theory, and educational theory. Specifically, theories in hermeneutics, literary immersion, embodiment, empathy, narrative, game ego, play, and learning theory are synthesized to produce a broad picture of the player/avatar relationship as it develops over time. I will identify stages in the process of feeling immersed in a game. This feeling of immersion can, but may not necessarily will, occur as the player engages in game play over a period of time. I will identify this process in stages: "other"; "embodied empathy"; "self-transcendence"; and "transformation." At the final stage, the game play can offer a critique of the player's worldview, to call into question the player's approach and even presuppositions about the world. I suggest here, the player no longer sees the avatar as an "other" a character to be understood in some way, and not as some virtual representation of the self, something apart from himself, and not even as some sort of virtual embodiment of himself. I suggest that the player transcends even the avatar as any sort of representation, that the avatar disappears from the consciousness of the player altogether. A result of this critique is a transformation that takes place within the player that may actually transform one's self-understanding, thus providing the player with an authentic learning experience.

There is a sense of embodiment that goes along with controlling an avatar in a virtual world that, I believe, is not present when becoming immersed in a film or a book. In a video game, the player's actions have direct and immediate consequences in terms of reward or punishment as well as movement through the virtual world. The narrative effect helps drive the character through this reward and punishment. The idea of affordances is often used in terms of game design and the use of emotional affordances to help propel the player into the virtual environment. Thus, with a video game, there is a more complete immersion into the constructed imaginary world—an embodied immersion and an emotional immersion.

I would like to clarify the definition of this feeling of immersion that happens as a video game progresses. Various terms are used in the fields of literary analysis, psychology, video game

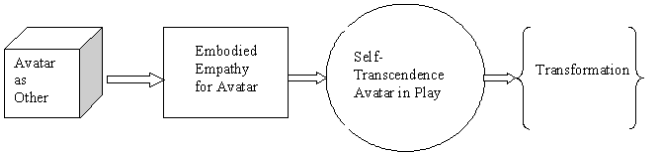
theory, and philosophy. Terms like literary transport, flow, presence, immersion, identification, and self-transcendence are often used to describe the feeling of the self moving out of the real world and into a constructed imaginary world. I use Janet Murray's offers a commonly accepted definition used in the gaming industry for this term, a sense of being physically submerged in water (Murray 1997, p. 99). My goal here is to illuminate the idea that there is a *process* to gaining this feeling of immersion. The player proceeds step-by-step into the game world. The work of Gallagher in philosophy of mind acknowledges that there is a growth process that takes place in developing an understanding of other people, through primary and secondary interaction. Gallese offers an alternative argument that the biological basis of this growth in understanding happen initially as a neural response from the mirror neuron system, then as a more mature empathetic response. Ultimately, I would like to establish a sequence that combines the work of Gallese and Gallagher with the work of digital narrative theorists Ryan and Murray. The sequence of immersion could be stated as: other→empathy→self-transcendence→transformation.

The first stage of approaching the avatar is learning the mechanics of the game. In this stage the player sees the avatar as "other," as an awkward and theoretical representation of the self. Initially, the avatar is a picture or symbol for the self. The player learns to manipulate the keyboard, mouse, or controller. The avatar is "other," a foreign virtual object that must be awkwardly manipulated by the player. In this stage there is little personal identification with the avatar, except on a theoretical level. The avatar cannot function efficiently according to the will of the player, but operates as a cumbersome vehicle for the player's intentions.

I propose that the player in the second stage the player begins to view the avatar as a character in a world, and the player begins to empathize with the character in a novel or in a film. Let me clarify, here, that at this stage the avatar is still viewed as Other, but a component of empathy now emerges through the use of narrative. The player now has learned to effectively manipulate the character, to move the character through the virtual world. I believe, here, embodiment plays an important biological role in helping to establish this feeling of empathy for the character. There is also an important component of narrative that drives the player toward empathy with the avatar. The work of Marie-Laure Ryan is an important theoriest in this area.

In the third stage, which I call "self-transcendence," the player experiences full identification with the avatar, not as empathizing with another character, but embodying the actions and world of the avatar as if it were his own. On this point, I will refer to Gadamer's idea of the self-transcendence of play as well as his article "The Relevance of the Beautiful" and several authors working on the concept of immersion and game ego in video game theory (Murray, Wilhelmsson, Ryan). The literature on video game theory uses the terms immersion, flow, and presence in a similar way, but I feel Gadamer's term "self-transcendence" more aptly fits the description I will offer.

At the point of self-transcendence, transformation can happen. Learning theory and the philosophy of hermeneutics, especially as stated in the work of Gallagher in `Hermeneutics and Education` and the hermeneutic theories of Gadamer (specifically his fusion of horizons), can establish how understanding happens and how transformation of the self can take place more completely in the virtually embodied world of video games.



View of the avatar character by the player of a video game as game play progresses.

References

Brna, Paul. "On the Role of Self-Esteem, Empathy and Narrative in the Development of Intelligent Learning Environments" Pivec, Maja ed. *Affective and Emotional Aspects of Human-Computer Interaction: Game-Based and Innovative Learning Approaches.* IOS Press: Washington, DC. 2006.

Dias, J., Paiva, A., Vala, M., Aylett, R., Woods S., Zoll, C., and Hall, L. "Empathic Characters in Computer-Based Personal and Social Education." Pivec, Maja ed. *Affective and Emotional Aspects of Human-Computer Interaction: Game-Based and Innovative Learning Approaches*. IOS Press: Washington, DC. 2006.

Gadamer, H.G. *Truth and Method.* Weinsheimer and Marshall, trans. New York: Crossroad, 1989.

Gallagher, Shaun. *Hermeneutics and Education.* Albany: State University of New York Press, 1992.

Gallagher, Shaun. *How the Body Shapes the Mind.* Oxford University Press: New York, 2005.

Gallagher, S. and Hutto, D. (in press-2007). Understanding others through primary interaction and narrative practice. In: Zlatev, Racine, Sinha and Itkonen (eds). *The Shared Mind: Perspectives on Intersubjectivity.* Amsterdam: John Benjamins.

Gallese, V. "The 'Shared Manifold' Hypothesis: From Mirror Neurons to Empathy." *Journal of Consciousness Studies.* 8:33-50, 2001.

Goldman, A. "Empathy, Mind, and Morals." In *Mental Simulation*. Ed. M. Davies and T. Stone, 185-208. Oxford: Blackwell, 1995.

Grodal, Torben. "Stories for Eye, Ear, and Muscles: Video Games, Media, and Embodied Experiences." *The Video Game Theory Reader.* Mark J. Wolf and Bernard Perron, eds. Routledge: New York, 2003.

Johnson, Mark (1987) *The Body in the Mind: The Bodily Basis of Meaning, Imagination and Reason*. Chicago: Chicago University Press, 1987 and 1990.

Lahti, Martti. "As We Become Machines: Corporealized Pleasures in Video Games." *The Video Game Theory Reader.* Mark J. Wolf and Bernard Perron, eds. Routledge: New York, 2003.

McMahan, Alison. (2003). "Immersion, Engagement, and Presence: A Method for Analysing 3-D Video Games." *The Video Game Theory Reader.* Mark J. Wolf and Bernard Perron, eds. Routledge: New York.

Murray, Janet. *Hamlet on the Holodeck: The Future of Narrative in Cyberspace*. Cambridge: MIT Press, 1997.

Ryan, Marie-Laure. *Narrative as Virtual Reality: Immersion and Interactivity in Literature and Electronic Media.* Baltimore: Johns Hopkins University Press, 2001.

Sykes, Jonathan. (2006). "Affective Gaming: Advancing the Argument for Game-Based Learning." Pivec, Maja ed. *Affective and Emotional Aspects of Human-Computer Interaction: Game-Based and Innovative Learning Approaches*. IOS Press: Washington, DC.

White, H. (1981). "Narrative and History." In W.J.T. Mitchell, ed., *On Narrative.* Chicago: University of Chicago Press.

Wilhelmsson, Ulf. (2006). "What is a Game Ego? (or How the Embodied Mind Plays a Role in Computer Game Environments)." Pivec, Maja ed. *Affective and Emotional Aspects of Human-Computer Interaction: Game-Based and Innovative Learning Approaches.* IOS Press: Washington, DC.

Wittgenstein, Ludwig. *Philosophical Investigations*. Trans. G.E.M. Anscombe. Blackwell Publishing: Cambridge, 1972.

Normalizing Identity: The Role of Blogging Software in Creating Digital Identity

Kirsten Carol Uszkalo
kuszkalo@stfx.ca
St. Francis Xavier University, Canada

Darren James Harkness
webmaster@staticred.net
University of Alberta, Canada

We offer a new in-depth methodology for looking at how the use of blogging software delineates and normalizes the blogger's creation of posts and, by extension, the creation of her blog self. The simple choice of software is not simple at all and in fact has a great influence on the shape a blogger's identity will take through the interface, program design, and data structures imposed on her by the software. This primarily technical discussion of a topic seldom considered in studies which look at the cultural impact of blogging, will illuminate the inner workings of the medium and gives due credence to Marshall McLuhan's argument that "the 'message' of any medium or technology is the change of scale or pace or pattern that it introduces into human affairs".

Technology plays such an integral part in distinguishing the blog as a new medium, apart from that of written (paper-based) text; as a result, a study of the blog that does not investigate its infrastructure is incomplete. Critics of blogging scholarship point out the lack of technical discussion around the software used by bloggers as a weakness (Scheidt 2005, Lawley 2004). The criticism is valid; critics' attention has been focused on the output of the blogs, categorizing them as an extension of existing genres, whether that be of the diary (Rak 167) or the newspaper (Koman). This study serves as a response to the criticism, and aims to start the discussion by looking into the dark recesses of the software, databases, and code to illustrate just how influential infrastructure is in defining identity.

Programmers do not care about the content of any given blog. The people who develop Movable Type, Blogger, LiveJournal, and Wordpress *are* developing software which helps making blogging a much simpler process, and they do listen to customer requests for features. But the developer is not concerned whether your blog will be an online journal, a political commentary, or a collection of cat pictures – what she is concerned about is memory allocation, disk usage, and transaction speed. Every shortcut taken in the source code, every data type or archiving scheme not supported, every function written, and every decision made by the programmer to achieve these goals has an influence on the interface, and therefore on the content the blogger produces. Despite working at an indifferent distance, the developer heavily influences the blog – and by extension, blogger's identity – by the decisions she makes when she codes the software.

The way we structure language helps create meaning; likewise, the way in which it is stored also has meaning. To the programmer, language is nothing more than a set of bits and data types, which must be sorted into different containers. How the programmer deals with data affects how she creates the interface; if she has no data structure in place to handle a certain kind of information, she cannot request it from the user in the interface. The data structure is created through a process called normalization – breaking data down into its smallest logical parts. Developers normalize data in order to make it easier to use and reuse in a database: the title of your blog entry goes in one container; the body text goes into another, and so on. The structure of the data does not necessarily match the structure of its original context, however. Although a title and the body text are related to the same entry, no consideration is given by the developer as to whether one comes before the other, whether it should be displayed in a specific style, or if one has hierarchical importance over the other in the page. The data structure is dictated by the individual pieces of data themselves. The developer takes the data within each of these containers and stores it within a database. This may be a simple database, such as a CSV[1] or Berkeley DB[2] file, or it may reside within a more complex relational database such as MySQL or Microsoft SQL Server. Within the database exists a series of tables, and within each table resides a series of fields. A table holds a single record of data – a blog entry – and the table's fields hold properties of that data, such as the title or entry date. *Figure 1* illustrates an example of the above; a developer has created an *Entries* table with the fields *EntryID*[3]*, Title, Date, BodyText, ExtendedText, Keywords, Category,* and *Post Status*.

When is possible, such as with the *Category* and *Post Status* fields, the developer will actually replace a string (alphanumeric) value with a numeric pointer to the same data within another table in the database. For example, an author may create a set of categories for her blog (such as "Personal Life," "School," et cetera, which are stored in a separate database table named *Categories* and associated with a unique ID (*CategoryID*). When an entry is marked with the Personal category, the software queries the database to see what the *CategoryID* of the *Personal* category is in the *Categories* table, and places that in the *Category* field in an entry's record in the *Entries* table (see Figure 2). This sets up a series of relations within a database, and helps keep the database smaller; an integer takes far less space in the database than a string: 1 byte to store a single-digit integer, compared to 8 bytes for the string "Personal"; when you start working with hundreds of entries, this difference adds up quickly. It is also easier to maintain; if you want to rename the "Personal" category to "Stories from the woeful events of my unexaggerated life" for example, you would only have to update the entry once in the *Categories* table; because it is referenced by its *CategoryID* in each entry, it will automatically be updated in all records that reference it. By abstracting often-used data such as a category into separate database tables, data can be reused within the database, which in turn keeps the size of the database smaller. If we know we will be referring to a single category in multiple entries, it makes sense to create a table of possible categories and then point to their unique identifier within each individual entry.

Each field within a database table is configured to accept a specific format of information known as a data type. For example, the *Date* field in the *Entries* table above would be given a data type of DATETIME,[4] while the *Category* field would be given a data type of INT (to specify an integer value). The body text of an entry would be placed in a binary data type known as the BLOB, since this is a type of data whose size is variable from record to record. Normalization conditions data to its purpose, and ensures that the developer always knows what kind of data to expect when he or she retrieves it later. It also has the benefit of loosely validating the data by rejecting invalid data types. If an attempt to store a piece of INT data in the *Date* field is made, it will trigger an error, which prevents the data from being misused within an application.

The decisions made by the developer at this point, which involve configuring the tables and fields within the database, ultimately determine what will appear in the blog's interface. If tables and fields do not exist in the database to support categorization of an entry, for example, it is unlikely to appear in the interface since there is no facility to store the information (and by extension, not prompt the blogger to categorize her thoughts).

The interface gives the blogger certain affordances, something Robert St. Amant defines as "an ecological property of the relationship between an agent and the environment" (135).[5] Amant describes affordance as a function we can see that is intuitive: "we can often tell how to interact with an object or environmental feature simply by looking at it, with little or no thought involved" (135, 136) – for example, we instinctively know not only what a chair is for, but the best way to make use of it. St. Amant further breaks down the affordance into four separate affordance-related concepts: relationship, action, perception, and mental construct (136-7). He goes on to discuss how to incorporate the idea of affordance into developing a user interface, focusing on action and relationship. The last of these concepts, affordance as a mental construct, is most relevant to our discussion. St. Amant writes "these mental affordances are the internal encodings of symbols denoting relationships, rather than the external situations that evoke the symbols" (137). In the authoring of the blog, the affordance of developing identity cannot be pinned on a single HTML control or text box; it is the process as a whole. LiveJournal and DiaryLand, for example, have the affordance of keeping a personal journal, or online diary. Blogger has the affordance of developing identity in a broader way by not necessarily focusing it on an autobiographical activity. The interface leads the blogger into a mode of writing through the affordances it provides The infrastructure of the blog is its most fundamental paratextual element creating a mirror for the blogger to peer into, but it is the blogger that makes the decision to look.

Notes

1 Comma-Separated Values. Each record of data consists of a single, unbroken line within a text file, which contains a series of values – each separated by a comma or other delimiter. An example of a CSV file for our entry would look like the following: EntryID,Title,Da te,BodyText,ExtendedText, Keywords,Category, PostStatus

1,My Entry,12/15/2006,This is the entry,,personal,Personal,Published

2,My Other Entry,12/20/2006,Look – another entry,And some extended text,personal,Personal,Published

2 Berkeley DB is a file-based database structure, which offers some basic relational mechanisms, but is not as robust or performant as other database systems.

3 For compatibility with multiple database systems, spaces are generally discouraged in both table and field names.

4 For the purposes of this section, I will use MySQL Data types. Data types may vary slightly between different database applications.

5 As an interesting historical footnote, St. Amant wrote about the affordance in user interfaces around the time the first blog software packages were released in 1999.

References

Bolter, J.D. & Grusin, R. *Remediation: Understanding New Media.* Cambridge, MA: The MIT Press, 1999

Lawley, Liz. "Blog research issues" *Many 2 Many: A group weblog on social software.* June 24, 2004. Online. <http://many.corante.com/archives/2004/06/24/blog_research_issues.php> Accessed October 23, 2007.

Koman, Richard. "Are Blogs the New Journalism" *O'Reilly Digital Media.* January 8, 2005. Online. Accessed April 8, 2007. http://www.oreillynet.com/digitalmedia/blog/2005/01/are_blogs_the_new_journalism.html

McLuhan, Marshall. Understanding Media: The Extension of Man. Cambridge, Mass: MIT Press, 1994.

Rak, Julie (2005, Winter). "The Digital Queer: Weblogs and Internet Identity." Biography 28(1): 166-182.

Scheidt, Lois Ann. "Quals reading – Outlining in Blood" Professional-Lurker: Comments by an Academic in Cyberspace. Online. Accessed May 23, 2005. <http://www.professional-lurker.com/archives/000410.html>

Serfaty, Viviane. The Mirror and the Veil: An Overview of American Online Diaries and Blogs. Amsterdam & New York: Rodopi, 2004.

St. Amant, Robert. "Planning and User Interface Affordances" *Proceedings of the 4th international conference on Intelligent user interfaces.* 1998.

A Modest proposal. Analysis of Specific Needs with Reference to Collation in Electronic Editions

Ron Van den Branden

ron.vandenbranden@kantl.be

Centrum voor Teksteditie en Bronnenstudie (KANTL), Belgium

Text collation is a vital aspect of textual editing; its results feature prominently in scholarly editions, either in printed or electronic form. The Centre for Scholarly Editing and Document Studies (CTB) of the Flemish Royal Academy for Dutch Language and Literature has a strong research interest in electronic textual editing. Much of the research for the electronic editions of *De teleurgang van den Waterhoek* (2000) and *De trein der traagheid* (forthcoming) concerned appropriate ways of visualising textual traditions. Starting from the functionality of collation results in electronic editions, as illustrated in the latter edition, this paper will investigate the needs for a dedicated automatic text collation tool for text-critical research purposes.

The second section sets out with identifying different scenarios in the production model of electronic scholarly editions (based on Ott, 1992; Vanhoutte, 2006). Consequently, possible criteria for an automatic text collation tool in these production scenarios are identified, both on a collation-internal and generic software level. These criteria are analysed both from the broader perspective of automatic differencing algorithms and tools that have been developed for purposes of automatic version management in software development (Mouat, 2002; Komvoteas, 2003; Cobéna, 2003; Cobéna e.a., 2002 and [2004]; Peters, 2005; Trieloff, 2006), and from the specific perspective of text encoding and collation for academic purposes (Kegel and Van Elsacker, 2007; Robinson, 2007). Collation-internal criteria distinguish between three phases of the collation process: the automatic text comparison itself, representation of these comparisons and aspects of visualisation. Especially in the context of electronic scholarly editions, for which TEI XML is the de facto encoding standard, a degree of XML-awareness can be considered a minimal requirement for collation algorithms. A dedicated collation algorithm would be able to track changes on a structural and on word level. Moreover, since textual editing typically deals with complex text traditions, the ability to compare more than two versions of a text could be another requirement for a collation algorithm. Regarding representation of the collation results, an XML perspective is preferable as well. This allows for easier integration of the collation step with other steps in electronic editing. In a maximal scenario, a dedicated text collation tool would represent the collation results immediately in one or other TEI flavour for encoding textual variation. Visualisation of the collation results could be considered a criterion for a text collation tool, but seems less vital, however prominent it

features in the broader context of developing an electronic edition. From a generic software perspective, a dedicated tool would be open source, free, multi-platform, and embeddable in other applications. These criteria are summarised in a minimal and a maximal scenario. The plea for a tool that meets the criteria for a minimal scenario is illustrated in the third section of the paper.

The third section of the paper is a case study of another electronic edition in preparation at the CTB: the complete works of the 16th century Flemish poetess Anna Bijns, totalling around 300 poems that come in about 60 variant pairs or triplets. The specific circumstances of this project led to an investigation for collation solutions in parallel with the transcription and markup of the texts. After an evaluation of some interesting candidate tools for the collation step, the choice was made eventually to investigate how a generic XML-aware comparison tool could be put to use for multiple text collations in the context of textual editing. Besides formal procedures and a set of XSLT stylesheets for different processing stages of the collation results, this development process provided an interesting insight in the specific nature of 'textual' text collation, and the role of the editor. The description of the experimental approach that was taken will illustrate the criteria sketched out in the previous section, indicate what is possible already with quite generic tools, and point out the strengths and weaknesses of this approach.

Finally, the findings are summarised and a plea is made for a text collation tool that fills the current lacunae with regards to the current tools' capacities for the distinct steps of comparison, representation and visualisation.

References

Cobéna, Grégory (2003). *Change Management of semi-structured data on the Web.* Doctoral dissertation, Ecole Doctorale de l'Ecole Polytechnique. <ftp://ftp.inria.fr/INRIA/publication/Theses/TU-0789.pdf>

Cobéna, Grégory, Talel Abdessalem, Yassine Hinnach (2002). A comparative study for XML change detection. Verso report number 221. France: Institut National de Recherche en Informatique et en Automatique, July 2002. <ftp://ftp.inria.fr/INRIA/Projects/verso/VersoReport-221.pdf>

Cobéna, Grégory, Talel Abdessalem, Yassine Hinnach ([2004]). A comparative study of XML diff tools. Rocquencourt, [Unpublished update to Cobéna e.a. (2002), seemingly a report for the Gemo project at INRIA.] <http://www.deltaxml.com/dxml/90/version/default/part/AttachmentData/data/is2004.pdf>

Kegel, Peter, Bert Van Elsacker (2007). "'A collection, an enormous accumulation of movements and ideas". Research documentation for the digital edition of the Volledige Werken (Complete Works) of Willem Frederik Hermans'. In: Georg Braungart, Peter Gendolla, Fotis Jannidis (eds.), *Jahrbuch für Computerphilologie 8* (2006). Paderborn: mentis Verlag. p. 63-80. <http://computerphilologie.tu-darmstadt.de/jg06/kegelel.html>

Komvoteas, Kyriakos (2003). *XML Diff and Patch Tool.* Master's thesis. Edinburgh: Heriot Watt University. <http://treepatch.sourceforge.net/report.pdf>

Mouat, Adrian (2002). *XML Diff and Patch Utilities*. Master's thesis. Edinburgh: Heriot Watt University, School of Mathematical and Computer Sciences. <http://prdownloads.sourceforge.net/diffxml/dissertation.ps?download>

Ott, W. (1992). 'Computers and Textual Editing.' In Christopher S. Butler (ed.), *Computers and Written Texts*. Oxford and Cambridge, USA: Blackwell, p. 205-226.

Peters, Luuk (2005). Change Detection in XML Trees: a Survey. Twente Student Conference on IT, Enschede, Copyright University of Twente, Faculty of Electrical Engineering, Mathematics and Computer Science. <http://referaat.cs.utwente.nl/documents/2005_03_B-DATA_AND_APPLICATION_INTEGRATION/2005_03_B_Peters,L.J.-Change_detection_in_XML_trees_a_survey.pdf>

Robinson, Peter (2007). How CollateXML should work. Anastasia and Collate Blog, February 5, 2007. <http://www.sd-editions.com/blog/?p=12>

Trieloff, Lars (2006). *Design and Implementation of a Version Management System for XML documents.* Master's thesis. Potsdam: Hasso-Plattner-Institut für Softwaresystemtechnik. <http://www.goshaky.com/publications/master-thesis/XML-Version-Management.pdf>

Vanhoutte, Edward (2006). Prose Fiction and Modern Manuscripts: Limitations and Possibilities of Text-Encoding for Electronic Editions. In Lou Burnard, Katherine O'Brien O'Keeffe, John Unsworth (eds.), *Electronic Textual Editing.* New York: Modern Language Association of America, p. 161-180. <http://www.tei-c.org/Activities/ETE/Preview/vanhoutte.xml>

(Re)Writing the History of Humanities Computing

Edward Vanhoutte
edward.vanhoutte@kantl.be
Centrum voor Teksteditie en Bronnenstudie (KANTL), Belgium

Introduction

On several occasions, Willard McCarty has argued that if humanities computing wants to be fully *of* the humanities, it needs to become historically self-aware (McCarty, 2004, p. 161). Contrary to what the frequently repeated myth of humanities computing claims (de Tollenaere, 1976; Hockey, 1980, p. 15; 1998, p. 521; 2000, p. 5; 2004, p. 4), humanities computing, like many other interdisciplinary experiments, has no very well-known beginning.

A Common History

Several authors contradict each other when it comes to naming the founding father of digital humanities. At least four candidates compete with each other for that honour. Depending on whether one bases one's research on ideas transmitted orally or in writing, on academic publications expressing and documenting these ideas, or on academic writings publishing results, Father Roberto Busa s.j., the Reverend John W. Ellison, Andrew D. Booth, and Warren Weaver are the competitors. This is not surprising since, when the use of automated digital techniques were considered to process data in the humanities, two lines of research were activated: Machine Translation (MT) and Lexical Text Analysis (LTA). Booth and Weaver belonged to the MT line of research and were not humanists by training, whereas Busa and Ellison were LTA scholars and philologists. As Antonio Zampolli pointed out, the fact that MT was promoted mainly in 'hard science' departments and LTA was mainly developed in humanities departments did not enhance frequent contacts between them. (Zampolli, 1989, p. 182) Although real scholarly collaboration was scarce, contacts certainly existed between the two on the administrative level. Since, as Booth (1967, p. VIII) observes, pure linguistic problems such as problems of linguistic chronology and disputed authorship, have profited from work on MT, the transition from MT to Computational Linguistics, dictated by the infamous US National Academy of Sciences ALPAC report (ALPAC, 1966), was a logical, though highly criticized, step. Moreover, the early writings on MT mention the essential use of concordances, frequency lists, and lemmatization – according to Zampolli (1989) typical products of LTA – in translation methods, which shows that both Computational Linguistics and Humanities Computing share a common history. This early history of both has never been addressed together, but it seems necessary for a good understanding of what it is that textual studies do with the computer.

The History of Humanities Computing

Researching and writing the history of humanities computing is no less but neither no more problematic than researching and writing the history of computing, technology in general, or the history of recent thought. Beverley Southgate considers the history of thought as an all-embracing subject matter which can include the history of philosophy, of science, of religious, political, economic, or aesthetic ideas, 'and indeed the history of anything at all that has ever emerged from the human intellect.' (Southgate, 2003, p. 243) Attached to the all-embracing nature of what she then calls 'intellectual history/history of ideas' is the defiance from the constraints of disciplinary structures it creates with its practitioners. The history of humanities computing for instance must consider the history of the conventional schools of theory and practice of humanities disciplines, the general histories of computing and technology, the history of relevant fields in computing science and engineering, and the history of the application of computational techniques to the humanities.

The History of Recent Things

The history of recent things, however, poses some new and unique challenges for which, in Willard McCarty's vision, a different conception of historiography is needed. As an illustration for his point, McCarty focuses on the different qualities of imagination that characterize the research and writing of the classical historian on the one hand and the historian of recent things on the other and situates them in a temporal and a spatial dimension respectively. A successful classical historian must manage the skill to move into the mental world of the long dead, whereas the historian of the recent past, like the anthropologist, must master the movement away 'from the mental world we share with our subjects while remaining engaged with their work' (McCarty, 2004, p. 163). In a rash moment, this double awareness of the historian of the recent is fair game to triumphalism, the historian's worst fiend. The temptation to search for historical and quantitative, rather than qualitative, 'evidence' to prove the importance of their own field or discipline is innate in somewhat ambitious insiders attempting at writing the history of their own academic field. Also since, as Alun Munslow has reiterated, it is historians rather than the past that generates (writes, composes?) history (Munslow, 1999), McCarty's astute but non-exhaustive catalogue of the new and unique challenges presents itself as a warning for the historiographer of the new, and I endorse this list fully: 'volume, variety, and complexity of the evidence, and difficulty of access to it; biases and partisanship of living informants; unreliability of memory; distortions from the historian's personal engagement with his or her informants—the 'Heisenberg effect', as it is popularly known; the 'presentism' of science and its perceived need for legitimation through an official, triumphalist account; and so on.' (McCarty, 2004, p. 163)

All history is inevitably a history *for*, and can never be ideologically neutral, as Beverly Southgate has recently emphasized in a review of Marc Ferro's *The Use and Abuse of History, or, How the Past is Taught*. (Southgate, 2005) Therefore the question can never be 'Comment on raconte l'histoire aux enfants à travers le monde entier' as the original title of Ferro's book reads. Greg Dening would say that writing history is a performance in which historians should be 'focused on the theatre of what they do'. (Dening, 1996, p. 30) According to McCarty, the different conception of historiography, could profit from acknowledging ethnographic theory and ethnography as a contributory discipline since it entails a poetic of 'dilatation beyond the textable past and beyond the 'scientific' reduction of evidence in a correct and singular account.' (McCarty, 2004, p. 174)

Prolegomena

The current lack of theoretical framework that can define and study the history of humanities computing echoes what Michael Mahoney wrote with respect to the history of computing: 'The major problem is that we have lots of answers but very few questions, lots of stories but no history, lots of things to do but no sense of how to do them or in what order. Simply put, we don't yet know what the history of computing is really about.' (Mahoney, 1993)

Three possible statements can be deducted from this observation:

- We don't know what history is about;
- We don't know what humanities computing is about;
- We don't know what the history of humanities computing is about.

This paper aims at addressing these three basic questions by sketching out prolegomena to the history of humanities computing.

References

ALPAC (1966) *Languages and machines: computers in translation and linguistics. A report by the Automatic Language Processing Advisory Committee, Division of Behavioral Sciences, National Academy of Sciences, National Research Council.* Washington, D.C.: National Academy of Sciences, National Research Council (Publication 1416).

<http://www.nap.edu/books/ARC000005/html/>

Booth, A.D. (1967). Introduction. In Booth, A.D. *Machine Translation.* Amsterdam: North-Holland Publishing Company, p. VI-IX.

De Tollenaere, Felicien (1976). *Word-Indices and Word-Lists to the Gothic Bible and Minor Fragments*. Leiden: E. J. Brill.

Dening, Greg (1996). *Performances*. Chicago: University of Chicago Press.

Hockey, Susan (1980). *A Guide to Computer Applications in the Humanities*. London: Duckworth.

Hockey, Susan (1998). An Agenda for Electronic Text Technology in the Humanities. *Classical World*, 91: 521-42.

Hockey, Susan (2000). *Electronic Texts in the Humanities. Principles and Practice*. Oxford: Oxford University Press.

Hockey, Susan (2004). The History of Humanities Computing. In Schreibman, Susan, Siemens, Ray, and Unsworth, John (eds.), *A Companion to Digital Humanities*. Malden, MA/Oxford/Carlton, Victoria: Blackwell Publishing, p. 3-19.

Southgate, Beverley (2003). Intellectual history/history of ideas. In Berger, Stefan, Feldner, Heiko, and Passmore, Kevin (eds.), *Writing History. Theory & Practice.* London: Arnold, p. 243-260.

Southgate, Beverley (2005). Review of Marc Ferro, The Use and Abuse of History, or, How the Past is Taught. *Reviews in History*, 2005, 441.

<http://www.history.ac.uk/reviews/reapp/southgate.html>

Mahoney, Michael S. (1993). Issues in the History of Computing. Paper prepared for the Forum on History of Computing at the ACM/SIGPLAN Second History of Programming Languages Conference, Cambridge, MA, 20-23 April 1993.

<http://www.princeton.edu/~mike/articles/issues/issuesfr.htm>

McCarty, Willard (2004). As It Almost Was: Historiography of Recent Things. *Literary and Linguistic Computing*, 19/2: 161-180.

Munslow, Alun (1999). The Postmodern in History: A Response to Professor O'Brien. Discourse on postmodernism and history. Institute of Historical Research.

<http://www.history.ac.uk/discourse/alun.html>

Zampolli, Antonio (1989). Introduction to the Special Section on Machine Translation. *Literary and Linguistic Computing*, 4/3: 182-184.

Knowledge-Based Information Systems in Research of Regional History

Aleksey Varfolomeyev
avarf@psu.karelia.ru
Petrozavodsk State University, Russian Federation

Henrihs Soms
henrihs.soms@du.lv
Daugavpils University, Latvia

Aleksandrs Ivanovs
aleksandrs.ivanovs@du.lv
Daugavpils University, Latvia

Representation of data on regional history by means of Web sites is practised in many countries, because such Web sites should perform – and more often than not they really perform – three important functions. First, they further preservation of historical information related to a certain region. Second, Web sites on regional history promote and publicize historical and cultural heritage of a region, hence they can serve as a specific 'visiting-card' of a region. Third, they facilitate progress in regional historical studies by means of involving researchers in Web-community and providing them with primary data and, at the same time, with appropriate research tools that can be used for data retrieving and processing. However, implementation of the above-mentioned functions calls for different information technologies. For instance, preservation of historical information requires structuring of information related to a certain region, therefore databases technologies should be used to perform this function. In order to draft and update a 'visiting-card' of a region, a Web site can be designed on the basis of Content Management System (CMS). Meanwhile, coordination of research activities within the frameworks of a certain Web-community calls for implementation of modern technologies in order to create knowledge-based information systems.

In this paper, a concept of a knowledge-based information system for regional historical studies is framed and a pattern of data representation, aggregation, and structuring is discussed. This pattern can be described as a specialized semantic network hereafter defined as *Historical Semantic Network* (HSN).

Up to now, the most popular pattern of representation of regional historical information is a traditional Web site that consists of static HTML-pages. A typical example is the Web site *Latgales Dati* (created in 1994, see http://dau.lv/ld), which provides information about the history and cultural heritage of Eastern Latvia – *Latgale* (Soms 1995; Soms and Ivanovs 2002). On the Web site *Latgales Dati*, historical data are stored and processed according to the above-mentioned traditional pattern. The shortcomings of this pattern are quite obvious: it is impossible to revise the mode of processing and representation of data on the Web site; it is rather difficult to feed in additional information, to retrieve necessary data, and to share information.

In the HEML project, a special language of representation and linkage of information on historical events has been developed (http://www.heml.org). Advantages of this language are generally known: a modern approach to processing and representation of historical information on the basis of XML, a developed system of tools of visualization of information, etc. However, the processing of information is subjected to *descriptions of historical events*. It can be regarded as a drawback of this language, since such descriptions do not form a solid basis for historical data representation in databases. Actually, any description of a historical event is an interpretation of narratives and documentary records. For this reason, databases should embrace diverse *historical objects*, namely, documents and narratives, persons, historical monuments and buildings, institutions, geographical sites, and so on, and so forth. These objects can be described using different attributes simultaneously, which specify characteristic features of the objects. Sometimes, objects can be labelled with definite, unique attributes only. The detailed, exhaustive information about the objects can be represented by means of links between objects, thus creating a specific semantic network. Within this network, paramount importance is attached to special temporal objects – the chronological 'markers'. Connecting links between a number of historical objects, including special temporal objects, will form rather a solid basis for historical discourse upon historical events, processes, and phenomena.

The first characteristic feature of the semantic network, which is being created on the basis of the Web site *Latgales Dati,* is the principle role assigned to such historical objects as documentary records and narratives. Thus, the connecting links between the objects are constituted (actually, *reconstructed*) in accordance with the evidences of historical sources. In order to provide facilities for verification of the links, the semantic network should contain full texts of historical records as well as scanned raster images of them. On the one hand, texts and images can be directly connected with the objects of the semantic network by means of SVG technology; on the other hand, this connection can be established by means of mark-up schemes used in XML technology (Ivanovs and Varfolomeyev 2005).

The second characteristic feature of HSN is fuzziness of almost all aspects of information in the semantic network (Klir and Yuan 1995). The source of this fuzziness is uncertainty of expert evaluations and interpretations of historical data; moreover, evidences of historical records can be either fragmentary and contradictory, or doubtful and uncertain. Therefore, the level of validity of information is 'measured' by variables, which can be expressed by ordinary numbers (0–1). However, it seems that they should be expressed by vector quantities as pairs (*true, false*); the components *true* and *false* can assume values from 0 to 1. In this case, the pair (1,1) refers to very contradictory and unreliable information, meanwhile the pair (0,0) means

the absence of information. This approach was used in construction of four-valued logic (Belnap 1977). The fuzziness values defined for some objects can be propagated within a semantic network by means of logical reasoning (Hähnle 1993). This 'measurement' of validity of data is performed by users of HSN; the results of 'measurement' should be changeable and complementary, thus reflecting cooperation of researchers within the frameworks of Web-community.

Creating a semantic network, a number of principal operations should be performed: reconstruction, linkage, and representation. *Reconstruction* is either manual or automatic generation of interrelations between objects that are not directly reflected in historical records. In this stage, new historical objects can emerge; these objects are reconstructed in accordance with the system of their mutual relations. *Linkage* can be defined as detection of similar (in fuzzy sense) objects, which within a semantic network (or within a number of interrelated networks) form a definite unit. *Representation* is a number of operations of retrieving of data from the semantic network including representation of information by means of tables, graphs, timeline, geographic visualization, etc.

HSN as a basic pattern of representation of historical information in Internet provides researchers with proper tools, which can easily transform data retrieved from such Web sites. However, some problems should be solved in order to 'substitute' the traditional Web site *Latgales Dati* for HSN.

There are different approaches to representation of semantic networks in Internet. A generally accepted approach is division of network links into subjects, predicates, and objects followed by their representation by means of RDF. However, the RDF pattern (actually, directed graph in terms of mathematics) is not adequate to the hyper-graph pattern accepted in HSN, since in HSN one and the same link can connect different objects simultaneously. In order to represent hyper-graphs, different semantic networks – such as *WordNet* (Fellbaum 1998), *MultiNet* (Helbig 2006), topic maps (Dong and Li 2004) or even simple Wiki-texts interconnected by so-called 'categories' – should be used. Unfortunately, the above-mentioned semantic networks do not provide proper tools to perform the tasks set by HSN. For this purpose, the experience acquired by designers of universal knowledge processing systems (e.g. *SNePS*, see Shapiro 2007, or Cyc, see Panton 2006) should be taken into consideration.

One more important problem is linkage of different HSN, including *Latgales Dati*. Such linkage can be conducive to cooperation of researchers, since it ensures remote access to historical data and exchange of information and results of research work. It seems that Web-services technology can carry out this task. In this case, HSN should be registered in a certain UDDI-server. A user's query activates interaction of a Web-server with numerous related Web-servers, which are supplied with HSN-module. As a result, a semantic network is being generated on the basis of fragments of different networks relevant to the initial query.

The principles of creation of HSN substantiated above are used in designing a new version of the knowledge-based information system *Latgales Dati*. As software environment a specialized Web application created at Petrozavodsk State University is applied (http://mf.karelia.ru/hsn).

References

Belnap, N. A Useful Four-Valued Logic. In J.M. Dunn and G. Epstein, ed. *Modern Uses of Multiple-Valued Logic*. Dordreht, 1977, pp. 8-37.

Fellbaum, C., ed. *WordNet: An Electronic Lexical Database*. Cambridge, Mass., 1998.

Hähnle, R. *Automated Deduction in Multiple-Valued Logics*. Oxford, 1993.

Helbig, H. *Knowledge Representation and the Semantics of Natural Language*. Berlin; Heidelberg; New York, 2006.

Ivanovs, A., Varfolomeyev, A. Editing and Exploratory Analysis of Medieval Documents by Means of XML Technologies. In *Humanities, Computers and Cultural Heritage*. Amsterdam, 2005, pp. 155-60.

Klir, G.R., Yuan, B. *Fuzzy Sets and Fuzzy Logic: Theory and Applications*. Englewood Cliffs, NJ, 1995.

Dong, Y., Li, M. HyO-XTM: A Set of Hyper-Graph Operations on XML Topic Map toward Knowledge Management. *Future Generation Computer Systems*, 20, 1 (January 2004): 81-100.

Panton, K. et al. Common Sense Reasoning – From Cyc to Intelligent Assistant. In Yang Cai and Julio Abascal, eds. *Ambient Intelligence in Everyday Life*, LNAI 3864. Berlin; Heidelberg; New York, 2006, pp. 1-31.

Shapiro, S.C. et al. *SNePS 2.7 User's Manual*. Buffalo, NY, 2007 (http://www.cse.buffalo.edu/sneps/Manuals/manual27.pdf).

Soms, H. Systematisierung der Kulturgeschichtlichen Materialien Latgales mit Hilfe des Computers. In *The First Conference on Baltic Studies in Europe: History.* Riga, 1995, pp. 34-5.

Soms, H., Ivanovs, A. Historical Peculiarities of Eastern Latvia (Latgale): Their Origin and Investigation. *Humanities and Social Sciences.* Latvia, 2002, 3: 5-21.

Using Wmatrix to investigate the narrators and characters of Julian Barnes' Talking It Over

Brian David Walker
b.d.walker1@lancaster.ac.uk
Lancaster University, UK

Introduction

The focus of this paper is on Wmatrix (Rayson 2008), and how the output from this relatively new corpus tool is proving useful in connecting language patterns that occur in *Talking It Over* - a novel by Julian Barnes - with impressions of narrators and characters. My PhD research is concerned with: 1) the role of the narrator and author in characterisation in prose fiction, in particular how the narrator intervenes and guides our perception of character – something that, in my opinion, has not been satisfactorily dealt with in existing literature; and 2) how corpus tools can be usefully employed in this type of investigation. In this paper I will show that Wmatrix helps to highlight parts of the novel that are important insofar as theme and/or narratorial style and/or characterisation. I will go on to discuss difficulties I encountered during my investigation, and possible developments to Wmatrix that could be beneficial to future researchers.

Of course, computer-assisted approaches to the critical analyses of novels are no longer completely new, and *corpus stylistics* (as it is sometimes known) is a steadily growing area of interest. Indeed, corpus approaches have already been used to investigate character in prose and drama. McKenna and Antonia (1996), for example, compare the internal monologues of three characters (Molly, Stephen and Leopold) in Joyce's *Ulysses*, testing the significance on the most frequent words in word-lists for each character, and showing that common word usage can form "[…] distinct idioms that characterize the interior monologues […]" (McKenna and Antonia 1996:65). Also, more recently, Culpeper (2002) uses the 'KeyWord' facility in WordSmith Tools (Scott 1996) to demonstrate how the analysis of characters' keywords in Shakespeare's *Romeo and Juliet* can provide data to help establish lexical and grammatical patterns for a number of the main characters in the play.

While my research adopts similar approaches to the two examples above, it takes a small yet useful step beyond word-level analysis, focussing instead on semantic analysis and key-concepts. This is possible because Wmatrix can analyse a text at the semantic level (see below).

Wmatrix

Wmatrix (Rayson 2008) is a web-based environment that contains a number of tools for the analyses of texts. Plain-text versions of texts can be uploaded to the Wmatrix web-server where they are automatically processed in three different ways:

1) Word level – frequency lists of all the words in a text are compiled and can be presented either in order of frequency or alphabetically;

2) Grammatical level – using the Constituent Likelihood Automatic Word-tagging System (CLAWS – see Garside 1987, 1996; Leech, Garside and Bryant 1994; and Garside and Smith 1997) developed by the University Centre for Computer Corpus Research on Language (UCREL[i]) at Lancaster University, every word in a text is assigned a tag denoting the part-of-speech (POS) or grammatical category to which the word belongs. The words are then presented in list form either alphabetically (by POS tag) or by frequency (the most frequent POS tags at the top of the list);

3) Semantic level – every word is semantically tagged using the UCREL semantic analysis system (USAS[ii]) and then listed in order of frequency or alphabetically by semantic tag. USAS groups words together that are conceptually related. It assigns tags to each word using a hierarchical framework of categorization, which was originally based on MacArthur's (1981) *Longman Lexicon of Contemporary English*. Tags consist of an uppercase letter, which indicates the general discourse field, of which there are 21 – see Table 1.

Table 1 The 21 top level categories of the USAS tagset

A - general & abstract terms	B - the body & the individual	C - arts & crafts	E - emotion	F - food & farming
G - government & public	H - architecture, housing & the home	I - money & commerce (in industry)	K - entertainment	L - life & living things
M - movement, location, travel, transport	N - numbers & measurement	O - substances, materials, objects, equipment	P - education	Q - language & communication
S - social actions, states & processes	T - time	W - world & environment	X - psychological actions, states & processes	Y - science & technology
Z - names & grammar				

The uppercase letter is followed by a digit, which indicates the first sub-division of that field. The tag may also include a decimal point followed by a further digit to indicate finer sub-division of the field. For example, the major semantic field of GOVERNMENT AND THE PUBLIC DOMAIN is designated by the letter G. This major field has three subdivisions: GOVERNMENT, POLITICS AND ELECTIONS – tagged G1; CRIME, LAW AND ORDER – tagged G2; and WARFARE, DEFENCE AND THE ARMY – tagged G3. The first subdivision (G1 - GOVERNMENT, POLITICS AND ELECTIONS) is further divided into: GOVERNMENT ETC. – which is has the tag G1.1; and POLITICS – which is tagged G1.2.

From these initial lists, further analyses and comparisons are possible within the Wmatrix environment. However, the focus of this paper will be on my analysis of the USAS (semantic) output from Wmatrix, as this output proves to be the most interesting and the most useful interpretatively.

Data

Talking It Over, by Julian Barnes, is a story with a fairly familiar theme – a love triangle – that is told in a fairly unusual way – there are nine first person narrators. This offers interesting data with regard to the interaction between character and narrator, as the story is told from a number of different perspectives, with the different narrators often commenting on the same events from different angles. Of the nine narrators Oliver, Stuart and Gillian are the three main ones, not only because they are the three people involved in the love triangle, but also because they say substantially more than any of the other narrators in the novel. The six other narrators are people whose involvement in the story is more peripheral, but come into contact with one or more of the main narrators. Within the novel each contribution by different narrators is signalled by the narrator's name appearing in bold-type at the beginning of the narration. Some chapters consist of just one contribution from one narrator while others consist of several "narrations". The three main narrators make contributions of varying lengths throughout the novel.

Approach

The exploration of narrators described in this paper adopts a similar approach to that used by Culpeper (2002) in his analysis of characters in *Romeo and Juliet.* That is, I compare the words of one narrator with the words of all the other narrators combined, using Wmatrix in its capacity as a tool for text comparison. This comparison produced a number of lists ranked by statistical significance. The measure of statistical significance used by Wmatrix is log-likelihood (LL).

The process of comparison involved, to some extent, disassembling the novel, in order to extract, into separate text files, the narrated words of each of the three main narrators. The disassembly process had to take into account the fact that within the different narrations there were the words of other people and characters. That is to say, the narrations contained speech, writing and thought presentation (SW&TP). This needed to be accounted for in the analysis in order to be as precise as possible about descriptions and impressions of narrators based on the Wmatrix output. The best way to achieve this was to produce a tagged version of the novel, from which relevant portions could be extracted more exactly using computer-tools. These extracted portions could then be analysed using Wmatrix.

My paper will discuss the tagging process, the Wmatrix output and the subsequent analysis and show how key semantic concepts highlighted for each of the three main characters draw attention to important themes within the narrations, and to styles of narration which can then be linked to character traits. Included in my discussion will be issues concerning Wmatrix's lexicon and the type of automated semantic analysis Wmatrix carries out.

Conclusions

Interpretative conclusions

The list of semantic groups produced by Wmatrix for each of the three main narrators show a number of differences in narrator characteristics and narratorial styles. Stuart's list contains semantic groups that relate to his job. An investigation of highly significant categories in the list showed that Stuart's attitude relating to particular themes or concepts change during the novel. For example, Stuart's relationship with money alters, which is reflected in a change of attitude toward love. Stuart also becomes less worried about disappointing people and more concerned about not being disappointed himself.

An investigation of items in Gillian's list of semantic categories identified a more conversational style to her narration when compared to the rest of the narrators, as well as a determination to give an account of the story that was accurate and complete.

The top item in Oliver's list of categories contains the words that Wmatrix could not match. While on one hand this could be seen as a short-coming of Wmatrix, in the case of this study, the result gave a clear indication of the number of unusual, foreign and hyphenated words Oliver uses, and showed that he has very broad vocabulary as well as knowledge of a wide variety of topics. The list of unmatched words also highlighted Oliver's creativity with language, and that creativity is very much part of his style. He uses unusual, technical and poetic words in place of more commonplace synonyms and this could be seen as evidence towards Oliver being showy and flamboyant.

Wmatrix development

The results from this study, in particular those relating to Oliver's narration and the failure of Wmatrix to successfully categorise many of the words used in it, raise issues regarding the tool's lexicon and the semantic categories it uses. These issues will be discussed as potential avenues for the development of this useful corpus tool. For instance, the Wmatrix-lexicon is currently progressively updated and expanded, meaning that, in practice, there is no one static point from which comparisons of texts can be made. While there is a case for a lexicon that is as comprehensive as possible, the present way of managing the lexicon can cause difficulties for research projects that continue over an extended period of time. However, creating a 'standard' fixed lexicon is not without difficulties and raises questions about what counts as 'standard'. Even though Wmatrix allows users to define their own lexicon, a possible way forward might be to have multiple

lexicons, such as a scientific lexicon or a learner-English lexicon, which researchers could select depending on their research needs.

Notes

i For further information about UCREL see http://www.comp.lancs.ac.uk/ucrel/

ii For further information about USAS, see http://www.comp.lancs.ac.uk/ucrel/usas/

References

Culpeper, J. (2002) "Computers, language and characterisation: An Analysis of six characters in Romeo and Juliet." In: U. Melander-Marttala, C. Ostman and Merja Kytö (eds.), *Conversation in Life and in Literature: Papers from the ASLA Symposium, Association Suedoise de Linguistique Appliquee (ASLA)*, 15. Universitetstryckeriet: Uppsala, pp.11-30.

Garside, R. (1987). The CLAWS Word-tagging System. In: R. Garside, G. Leech and G. Sampson (eds), *The Computational Analysis of English: A Corpus-based Approach.* London: Longman.

Garside, R. (1996). The robust tagging of unrestricted text: the BNC experience. In J. Thomas and M. Short (eds) *Using corpora for language research: Studies in the Honour of Geoffrey Leech.* Longman, London, pp 167-180.

Garside, R., and Smith, N. (1997) A hybrid grammatical tagger: CLAWS4, in Garside, R., Leech, G., and McEnery, A. (eds.) *Corpus Annotation: Linguistic Information from Computer Text Corpora*. Longman, London, pp. 102-121.

Leech, G., Garside, R., and Bryant, M. (1994). CLAWS4: The tagging of the British National Corpus. In *Proceedings of the 15th International Conference on Computational Linguistics (COLING 94).* Kyoto, Japan, pp622-628.

McKenna, W. and Antonia, A. (1996) " 'A Few Simple Words' of Interior Monologue in Ulysses: Reconfiguring the Evidence". *In Literary and Linguistic Computing* 11(2) pp55-66

Rayson, P. (2008) *Wmatrix: a web-based corpus processing environment,* Computing Department, Lancaster University. http://www.comp.lancs.ac.uk/ucrel/wmatrix/

Rayson, P. (2003). Matrix: A statistical method and software tool for linguistic analysis through corpus comparison. *Ph.D. thesis*, Lancaster University.

Scott, M. (1996) *Wordsmith*. Oxford: Oxford University Press

The Chymistry of Isaac Newton and the Chymical Foundations of Digital Humanities

John A. Walsh
jawalsh@indiana.edu
Indiana University, USA

My paper will examine both a specific digital humanities project, The *Chymistry of Isaac Newton* <http://www.chymistry.org/>, and reflect more broadly on the field of digital humanities, by suggesting useful parallels between the disciplines of alchemy and digital humanities. The *Chymistry of Isaac Newton* project is an effort to digitize and edit the alchemical writings of Newton and to develop digital scholarly tools (reference, annotation, and visualization) for interacting with the collection.

Newton's "chymistry" has recently become a topic of widespread public interest.[1] It is no longer a secret that the pre-eminent scientist of the seventeenth century spent some thirty years working in his alchemical laboratory, or that he left a manuscript Nachlass of about 1,000,000 words devoted to alchemy. Newton's long involvement in chymistry figures prominently in NOVA's 2005 "Newton's Dark Secrets," and it occupies a major portion of the documentary's website at <http://www.pbs.org/wgbh/nova/newton/>. Even more attention was devoted to Newton's alchemy in the 2003 BBC production "Newton: The Dark Heretic." Newton's chymistry also is featured in recent popularizing studies such as Gleick's 2003 *Isaac Newton* and White's 1997 *Isaac Newton: the Last Sorcerer*. Despite the fascination that Newton's chymistry holds for the public, the subject has not received a corresponding degree of scrutiny from historians since the untimely passing of B.J.T. Dobbs in 1994 and Richard Westfall in 1996. Dobbs had made the issue of Newton's alchemy a cause célèbre in her influential historical monograph, *The Foundations of Newton's Alchemy* of 1975, and Westfall built upon her work in his own magisterial biography of Newton, Never at Rest , of 1980.

The relative lack of subsequent scholarly grappling with Newton's alchemy is regrettable for many reasons, particularly since Dobbs and Westfall raised strong claims about the relationship of his alchemical endeavors to his physics. In particular, they suggested that Newton's concept of force at a distance was strongly influenced by his work in alchemy, and that it was alchemy above all else that weaned Newton away from the Cartesian mechanical universe in favor of a world governed by dynamic interactions operating both at the macrolevel and the microlevel. Although Dobbs backed away from these earlier positions in her 1991 *Janus Faces of Genius*, she made the equally strong claim there that Newton's alchemy was primarily concerned with the operations of a semi-material ether that acted as God's special agent in the material world. Westfall too emphasized the putative connections between Newton's religion and his alchemy.

Interestingly, the historical speculations of Westfall, Dobbs, and their popularizers have relatively little to say about the relationship of Newton's alchemy to the alchemical tradition that he inherited. Perhaps because of the underdeveloped state of the historiography of alchemy in the 1970s, both Westfall and Dobbs portrayed Newton's seemingly obsessive interest in the subject as something exotic that required the help of extradisciplinary motives to explain it. Hence Dobbs looked beyond chymistry itself, invoking the aid of Jungian psychology in her *Foundations* (1975) and that of Newton's heterodox religiosity in her *Janus Faces* (1991). In neither case did she see chymistry as a field that might have attracted Newton on its own merits. Recent scholarship, however, has thrown a very different picture on alchemy in the seventeenth-century English speaking world. We now know that Newton's associate Robert Boyle was a devoted practitioner of the aurific art, and that the most influential chymical writer of the later seventeenth century, Eirenaeus Philalethes, was actually the Harvard-trained experimentalist George Starkey, who tutored Boyle in chymistry during the early 1650s. Boyle's "literary executor," the empiricist philosopher John Locke, was also deeply involved in chrysopoetic chymistry, and even Newton's great mathematical rival, Leibniz, had an abiding interest in the subject. In short, it was the norm rather than the exception for serious scientific thinkers of the seventeenth century to engage themselves with chymistry. We need no longer express amazement at the involvement of Newton in the aurific art, but should ask, rather, how he interacted with the practices and beliefs of his predecessors and peers in the burgeoning field of seventeenth-century chymistry. A first step in this endeavor, obviously, lies in sorting out Newton's chymical papers and making them available, along with appropriate digital scholarly tools, for systematic analysis.

The most recent phase of the project focuses on digital tools, including a digital reference work based on Newton's *Index Chemicus*, an index, with bibliographic references, to the field and literature of alchemy. Newton's *Index* includes 126 pages with 920 entries, including detailed glosses and bibliographic references. Our online reference work edition of Newton's *Index* will not be constrained by Newton's original structure and will extend functionality found in traditional print-based reference works. It will leverage information technologies such as searching and cross-linking to serve as an access point to the domain of seventeenth-century alchemy and a portal to the larger collection and to visualizations that graphically plot the relative occurrences of alchemical terms, bibliographic references, and other features of the collection. Other tools being developed include systems for user annotation of XML-encoded texts and facsimile page images.

My presentation will look at current progress, developments, and challenges of the *Chymistry of Isaac Newton*. With this project as context, I will venture into more theoretical areas and examine parallels between alchemy and digital humanities. For example, like digital humanities, alchemy was an inherently interdisciplinary field. Bruce T. Moran, in his *Distilling Knowlege* describes alchemy as an activity "responding to nature so as to make things happen without necessarily having the proven answer for why they happen" (10). This approach has a counterpart in the playful experimentation and affection for serendipitious discovery found in much digital humanities work. Moran also explains that the "primary procedure" of alchemy was distillation, the principle purpose of which was to "make the purest substance of all, something linked, it was thought, to the first stuff of creation" (11). Similarly, much of digital humanities work is a process of distillation in which visualizations or XML trees, for instance, are employed to reveal the "purest substance" of the text or data set. Codes and symbols and metadata were important parts of the alchemical discipline, as they are in digital humanities. Alchemy also had a precarious place in the curriculum. Moran, again, indicates that "although a sprinkling of interest may be found in the subject within the university, it was, as a manual art, always denied a part in the scholastic curriculum" (34). Digital Humanities is likewise often cordoned off from traditional humanities departments, and its practice and research is conducted in special research centers, newly created departments of digital humanities, or in schools of information science. The idea of the alchemist as artisan and the digital humanist as technician is another interesting parallel between the two fields. My paper will examine these and other similarities between the two fields and examine more closely the works of individual alchemists, scientists, and artists in this comparative context. Alchemy is an interdisciplinary field--like much digital humanities work--that combines empirical science, philosophy, spirituality, literature, myth and magic, tools and totems. I will argue that, as we escape the caricature of alchemy as a pseudo-science preoccupied with the occult, alchemy can serve as a useful model for future directions in digital humanities.

[1]As William R. Newman and Lawrence M. Principe have argued in several co-authored publications, it is anachronistic to distinguish "alchemy" from "chemistry" in the seventeenth century. The attempt to transmute metals was a normal pursuit carried out by most of those who were engaged in the varied realm of iatrochemistry, scientific metallurgy, and chemical technology. The fact that Newton, Boyle, Locke, and other celebrated natural philosophers were engaged in chrysopoeia is no aberration by seventeenth-century standards. Hence Newman and Principe have adopted the inclusive term "chymistry," an actor's category employed during the seventeenth century, to describe this overarching discipline. See William R. Newman and Lawrence M. Principe, "Alchemy vs. Chemistry: The Etymological Origins of a Historiographic Mistake," Early Science and Medicine 3(1998), pp. 32-65. Principe and Newman, "Some Problems with the Historiography of Alchemy," in William R. Newman and Anthony Grafton, Secrets of Nature: Astrology and Alchemy in Early Modern Europe (Cambridge, MA: MIT Press, 2001), pp. 385-431.

References

Dobbs, Betty Jo Teeter. *The Foundations of Newton's Alchemy or, "The Hunting of the Greene Lyon"*. Cambridge: Cambridge UP, 1975.

- --. *The Janus Faces of Genius: The Role of Alchemy in Newton's Thought*. Cambridge: Cambridge UP, 1991.

Gleick, James. *Isaac Newton*. New York: Pantheon, 2003.

Moran, Bruce T. *Distilling Knowledge: Alchemy, Chemistry, and the Scientific Revolution*. Cambridge: Harvard UP, 2005.

Newman, William R. *Promethean Ambitions: Alchemy and the Quest to Perfect Nature*. Chicago: U of Chicago P, 2004.

Newton, Isaac. *The Chymistry of Isaac Newton*. Ed. William R. Newman. 29 October 2007. Library Electronic Text Resource Service / Digital Library Program, Indiana U. 25 November 2007. <http://www.chymistry.org/>.

Westfall, Richard. Never at Rest: *A Biography of Isaac Newton*. Cambridge: Cambridge UP, 1980.

White, Michael. *Isaac Newton: The Last Sorcerer*. Reading: Addison-Wesley, 1997.

Document-Centric Framework for Navigating Texts Online, or, the Intersection of the Text Encoding Initiative and the Metadata Encoding and Transmission Standard

John A. Walsh
jawalsh@indiana.edu
Indiana University, USA

Michelle Dalmau
mdalmau@indiana.edu
Indiana University, USA

Electronic text projects range in complexity from simple collections of page images to bundles of page images and transcribed text from multiple versions or editions (Haarhoff, Porter). To facilitate navigation within texts and across texts, an increasing number of digital initiatives, such as the Princeton University Library Digital Collections (http://diglib.princeton.edu/) and the Oxford Digital Library (http://www.odl.ox.ac.uk/), and cultural heritage organizations, such as the Culturenet Cymru (http://www.culturenetcymru.com) and the Library of Congress (http://www.loc.gov/index.html), are relying on the complementary strengths of open standards such as the Text Encoding Initiative (TEI) and the Metadata Encoding and Transmission Standard (METS) for describing, representing, and delivering texts online.

The core purpose of METS is to record, in a machine-readable form, the metadata and structure of a digital object, including files that comprise or are related to the object itself. As such, the standard is a useful tool for managing and preserving digital library and humanities resources (Cantara; Semple). According to Morgan Cundiff, Senior Standards Specialist for the Library of Congress, the maintenance organization for the METS standard:

> METS is an XML schema designed for the purpose of creating XML document instances that express the hierarchical structure of digital library objects, the names and locations of the files that comprise those digital objects, and the associated descriptive and administrative metadata. (53)

Similarly, the TEI standard was developed to capture both semantic (e.g., metadata) and syntactic (e.g., structural characteristics) features of a document in machine-readable form that promotes interoperability, searchability and textual analysis. According to the Text Encoding Initiative web site, the TEI standard is defined thusly:

> The TEI *Guidelines for Electronic Text Encoding and Interchange* define and document a markup language for representing the structural, renditional, and conceptual features of texts. They focus (though not exclusively) on the encoding of documents in the humanities and social sciences, and in particular on the representation of primary source materials for research and analysis. These guidelines are expressed as a modular, extensible XML schema, accompanied by detailed documentation, and are published under an open-source license. ("TEI Guidelines")

METS, as its name suggests, is focused more exclusively on metadata. While digital objects—such as a text, image, or video—may be embedded within a METS document, METS does not provide guidelines, elements and attributes for representing the digital object itself; rather, the aim of METS is to describe metadata about a digital object and the relationships among an object's constituent parts. In sum, METS is data-centric; TEI is document-centric.

In April 2006, the Indiana University Digital Library Program released a beta version of METS Navigator (http://metsnavigator.sourceforge.net/), a METS-based, open source software solution for the discovery and display of multi-part digital objects. Using the information in the METS structural map elements, METS Navigator builds a hierarchical menu that allows users to navigate to specific sections of a document, such as title page, specific chapters, illustrations, etc. for a book. METS Navigator also allows simple navigation to the next, previous, first, and last page image or component parts of a digital object. METS Navigator can also make use of the descriptive metadata in the METS document to populate the interface with bibliographic and descriptive information about the digital object. METS Navigator was initially developed at Indiana University (IU) for the online display and navigation of brittle books digitized by the IU Libraries' E. Lingle Craig Preservation Laboratory. However, realizing the need for such a tool across a wide range of digital library projects and applications, we designed the system to be generalizable and configurable. To assist with the use of METS Navigator for new projects, a METS profile, also expressed in XML, is registered with the Library of Congress (http://www.loc.gov/standards/mets/profiles/00000014.html). The profile provides detailed documentation about the structure of the METS documents required by the METS Navigator application.

Cantara reported in her brief article, "The Text-Encoding Initiative: Part 2," that discussion of the relationship between the TEI and METS was a primary focus during the Fourth Annual TEI Consortium Members' Meeting held in 2004 at Johns Hopkins University (110). The relationship between the standards was also a topic of discussion during the 2007 Members' Meeting, and was specifically raised in Fotis Jannidis' plenary entitled "TEI in a Crystal Ball." Despite these discussions, the community is still lacking a well-documented workflow for the derivation of METS documents from authoritative TEI files. We believe the TEI, if properly structured, can be used as the "master" source of information from which a full METS document can be automatically generated, facilitating the display of text collections in METS Navigator. The TEI, much like METS, provides a rich vocabulary and framework for encoding descriptive and structural metadata for a variety of documents. The descriptive metadata typically found in the TEI Header may be used to populate the corresponding components (descriptive metadata section) of a METS document. The embedded structural metadata that describe divisions, sections, headings and page breaks in a TEI document may be used to generate the structural map section of a METS document. Unlike the TEI, the METS scheme has explicit mechanisms in place for expressing relationships between multiple representations of the digital content such as encoded text files and page images. By integrating the TEI-METS workflow into METS Navigator, scholars and digital library and humanities programs can more easily implement online text collections. Further, the intersection between TEI and METS documents can provide the foundation for enhanced end-user exploration of electronic texts.

The IU Digital Library Program as a result of enhancing the functionality and modularity of the METS Navigator software is also in the process of formalizing and integrating the TEI-METS workflow in support of online page turning. Our paper will trace the development of METS Navigator including the TEI-METS workflow, demonstrate the METS Navigator system using TEI-cum-METS documents, review METS Navigator configuration options, detail findings from recent user studies, and outline plans for current and future development of the METS Navigator.

References

Cantara, Linda. "Long-term Preservation of Digital Humanities Scholarship." OCLC Systems & Services 22.1 (2006): 38-42.

Cantara, Linda. "The Text-encoding Initiative: Part 2." *OCLC Systems & Services* 21.2 (2005): 110-113.

Cundiff, Morgan V. "An introduction to the Metadata Encoding and Transmission Standard (METS)." *Library Hi Tech* 22.1 (2004): 52-64.

Haarhoff, Leith. "Books from the Past: E-books Project at Culturenet Cymru." *Program: Electronic Library and Information Systems* 39.1 (2004): 50-61.

Porter, Dorothy. "Using METS to Document Metadata for Image-Based Electronic Editions." Joint International Conference of the Association for Computers and the Humanities and the Association for Literary and Linguistic Computing, June 2004. Göteborg: Centre for Humanities Computing (2004): 102-104. Abstract. 18 November 2007 <http://www.hum.gu.se/allcach2004/AP/html/prop120.html>.

Semple, Najla. "Develpoing a Digital Preservation Strategy at Edinburgh University Library." VINE: *The Journal of Information and Knowledge Management Systems* 34.1 (2004): 33-37.

"Standards, METS." *The Library of Congress*. 18 November 2007 <http://www.loc.gov/standards/mets/>.

"TEI Guidelines." *The Text Encoding Initiative*. 18 November 2007

< http://www.tei-c.org/Guidelines/>.

iTrench: A Study of the Use of Information Technology in Field Archaeology

Claire Warwick
c.warwick@ucl.ac.uk
University College London, UK

Melissa Terras
m.terras@ucl.ac.uk
University College London, UK

Claire Fisher
c.fischer@ucl.ac.uk
University College London, UK

Introduction

This paper presents the results of a study by the VERA project (**V**irtual **R**esearch **E**nvironment for **A**rchaeology: http://vera.rdg.ac.uk) which aimed to investigate how archaeologists use information technology (IT) in the context of a field excavation. This study was undertaken by researchers at School of Library, Archive and Information Studies, University College London, who collaborate on the project with the School of Systems Engineering, and the Department of Archaeology, University of Reading.

VERA is funded by the JISC Virtual Research Environments Programme, Phase 2, (http://www.jisc.ac.uk/whatwedo/programmes/vre2.aspx) and runs from April 2007 until March 2009. We aim to produce a fully operational virtual research environment for the archaeological community. Our work is based on a research excavation of part of the large Roman town at Silchester, which aims to trace the site's development from its origins before the Roman Conquest to its abandonment in the fifth century A.D (Clarke *et al* 2007). This complex urban site provides the material to populate the research environment, utilising the Integrated Archaeological Data Base (IADB: http://www.iadb.co.uk/specialist/it.htm), an online database system for managing recording, analysis, archiving and online publication of archaeological finds, contexts and plans. The dig allows us to: study the use of advanced IT in an archaeological context; investigate the tasks carried out within archaeological excavations; ascertain how and where technology can be used to facilitate information flow within a dig; and inform the designers of the IADB how it may be adapted to allow integrated use of the tools in the trench itself.

Research Context

Although archaeologists were quick to embrace IT to aid in research analysis and outputs (Laflin 1982, Ross *et al* 1990, Reilly and Rahtz 1992), and the use of IT is now central to

the manipulation and display of archaeological data (Lock and Brown 2000, McPherron and Dibble, 2002, Lock 2003) the use of IT to aid field archaeology is in its relative infancy due to the physical characteristics of archaeological sites, and the difficulties of using IT in the outdoor environment. Whilst the use of electronic surveying equipment (total stations, (Eiteljorg, 1994)) and digital cameras is now common place on archaeological sites there are relatively few archaeological organizations that use digital recording methods to replace the traditional paper records which rely on manual data input at a (sometimes much) later date. With ever increasing amounts of data being generated by excavations onsite databases are becoming increasingly necessary, for example at the excavations at Catalhoyuk in Turkey (http://www.catalhoyuk.com/database/catal/) and the site of Terminal 5 at Heathrow airport (http://www.framearch.co.uk/t5/), and some archaeologists have begun to use digital data input from total stations, PDAs, tablet PCs, digital cameras, digital callipers, digital pens and barcodes (McPherron and Dibble, 2003, Dibble et al., 2007). For many excavations, however, the use of IT is restricted to the analysis stage rather than field recording. The aim of the VERA project is to investigate the use of IT within the context of a field excavation and to ascertain if, and how, it may be appropriated to speed up the process of data recording, entry and access.

Method

We used a diary study to gather information about the work patterns of different archaeological roles and the way that they are supported by both digital and analogue technologies. The study was carried out by the UCL team, at the Silchester dig during the summer of 2007. (http://www.rdg.ac.uk/AcaDepts/la/silchester/publish/field/index.php) Researchers from Reading also carried out a study into the use of ICT hardware to support digging and data entry. A detailed record of the progress of both the dig and the study was kept on the VERA blog (http://vera.rdg.ac.uk/blog).

Diary studies enable researchers to understand how people usually work and can be used to identify areas that might be improved by the adoption of new working practices or technologies. (O'Hara et al. 1998). They have been used in the area of student use of IT, and to study the work of humanities scholars. (Rimmer et. al. Forthcoming) however, this is the first use of this method to study field archaeology that we are aware of.

During diary studies, participants are asked to keep a detailed record of their work over a short period of time. The participant records the activity that they are undertaking, what technologies they are using and any comments they have on problems or the progress of their work. This helps us to understand the patterns of behaviour that archaeologists exhibit, and how technology can support these behaviours.

We also obtained contextual data about participants using a simple questionnaire. This elicited information about the diary survey participants (role, team, status) and their experience of using the technology on site. A cross section of people representing different types of work and levels of experience were chosen. For example we included inexperienced and experienced excavators; members of the finds team, who process the discoveries made on site; those who produce plans of the site and visitor centre staff.

A defined area of the Silchester site was used to test the use of new technologies to support excavation. In this area archaeologists used digital pens and paper, (http://www.logitech.com/index.cfm/mice_pointers/digital_pen/devices/408&cl=us,en) digital cameras, and Nokia N800 PDAs (http://www.nokia.co.uk/A4305204). Diaries from this area were compared to those using traditional printed context sheets to record their work.

Findings

This paper will present the findings from the study, covering various issues such as the attitude of archaeologists to diary studies, previous and present experiences of using technology on research excavations, the effect of familiarity of technology on uptake and use, and resistance and concerns regarding the use of technology within an archaeological dig. We also evaluate specific technologies for this purpose, such as the Nokia N800 PDA, and Logitech Digital Pens, and ascertain how IT can fit into existing workflow models to aid archaeologists in tracking information alongside their very physical task.

Future work

This year's diary study supplied us with much interesting data about the past, current and potential use of IT in the trench. We will repeat the study next year to gain more detailed data. Participants will be asked to focus on a shorter period of time, one or two days, as opposed to five this year. Next year we will have a research assistant on site, allowing us to undertake interviews with participants, clarify entries, and build up a good working relationship with experts working on the excavation. Two periods of diary study will also be undertaken, allowing for analysis and refining methods between studies. This will also be juxtaposed with off-site user testing and analysis workshops of the IADB, to gain understanding of how archaeologists use technology both on and off site. We also plan to run an additional training session in the use of ICT hardware before the start of next year's dig, in addition to the usual archaeological training.

Conclusion

The aim of this study is to feed back evidence of use of IT to the team developing the virtual environment interface of the IADB. It is hoped that by ascertaining and understanding user needs, being able to track and trace information workflow throughout the dig, and gaining an explicit understanding of the tasks undertaken by archaeologists that more intuitive

technologies can be adopted, and adapted, to meet the needs of archaeologists on site, and improve data flow within digs themselves.

Acknowledgements

We would like to acknowledge the input of the other members of the VERA project team, Mark Barker, Matthew Grove (SSE, University of Reading) Mike Fulford, Amanda Clarke, Emma O'Riordan (Archaeology, University of Reading), and Mike Rains, (York Archaeological Trust). We would also like to thank all those who took part in the diary study.

References

Clarke, A., Fulford, M.G., Rains, M. and K. Tootell. (2007). Silchester Roman Town Insula IX: The Development of an Urban Property c. AD 40-50 - c. AD 250. *Internet Archaeology*, 21. http://intarch.ac.uk/journal/issue21/silchester_index.html

Dibble, H.L., Marean, C.W. and McPherron (2007) The use of barcodes in excavation projects: examples from Mosel Bay (South Africa) an Roc de Marsal (France). *The SAA Archaeological Record* 7: 33-38

Eiteljorg II, H. (1994). Using a Total Station. *Centre for the Study of Architecture Newsletter*, II (2) August 1994.

Laflin, S. (Ed). (1982). *Computer Applications in Archaeology*. University of Birmingham, Centre for Computing & Computer Science.

Lock, G. (2003). *Using Computers in Archaeology*. London: Routledge.

Lock, G. and Brown, K. (Eds). (2000). *On the Theory and Practice of Archaeological Computing.* Oxford University School of Archaeology

McPherron, S.P. and Dibble H.L. (2002) *Using computers in archaeology: a practical guide.* Boston, [MA]; London: McGraw-Hill/Mayfield. See also their website http://www.oldstoneage.com/rdm/Technology.htm

McPherron, S.P. and Dibble H.L. (2003) Using Computers in Adverse Field Conditions: Tales from the Egyptian Desert. *The SAA Archaeological Record* 3(5):28-32

O'Hara, K., Smith, F., Newman, W., & Sellen, A. (1998). Student readers' use of library documents: implications for library technologies. *Proceedings of the SIGCHI conference on Human factors in computing systems*. Los Angeles, CA. New York, NY, ACM Press/Addison-Wesley.

Reilly, P. and S. Rahtz (Eds.) (1992). *Archaeology and the Information Age: A global perspective*. London: Routledge, One World Archaeology 21.

Rimmer, J., Warwick, C., Blandford, A., Buchanan, G., Gow, J. An examination of the physical and the digital qualities of humanities research. *Information Processing and Management* (Forthcoming)

Ross, S. Moffett, J. and J. Henderson (Eds) (1990). *Computing for Archaeologists*. Oxford, Oxford University School of Archaeology.

LogiLogi: A Webplatform for Philosophers

Wybo Wiersma
wybo@logilogi.org
University of Groningen, The Netherlands

Bruno Sarlo
brunosarlo@gmail.com
Overbits, Uruguay

LogiLogi is a hypertext platform featuring a rating-system that tries to combine the virtues of good conversations and the written word. It is intended for all those ideas that you're unable to turn into a full sized journal paper, but that you deem too interesting to leave to the winds. It's central values are openness and quality of content, and to combine these values it models peer review and other valuable social processes surrounding academic writing (in line with Bruno Latour). Contrary to early websystems it does not make use of forumthreads (avoiding their many problems), but of tags and links that can also be added to articles by others than the original author. Regardless of our project, the web is still a very young medium, and bound to make a change for philosophy in the long run.

Introduction

The growth of the web has been rather invisible for philosophy so far, and while quite some philosophizing has been done about what the web could mean for the human condition, not much yet has been said about what it could mean for philosophy itself (ifb; Nel93; Lev97, mainly). An exception is some early enthusiasm for newsgroups and forums in the nineties, but that quickly died out when it became apparent that those were not suitable at all for in-depth philosophical conversations. The web as a medium however is more than these two examples of early web-systems, and in the meantime it has further matured with what some call Web 2.0, or social software (sites like MySpace, Del.icio.us and Wikipedia). Time for a second look. . .

LogiLogi Manta (Log), the new version of LogiLogi, is a webplatform that hopes to — be it informally and experimentally—allow philosophers and people who are interested in philosophy to use the possibilities that the internet has in stock for them too. It was started with a very small grant from the department of Philosophy of the Rijksuniversiteit Groningen. It is Free Software, has been under development for almost 2 years, and will be online by June 2008.

In the following paragraph we will explain what LogiLogi is, and in section 3 LogiLogi and the web as a new medium are embedded in the philosophical tradition. Optionally section 2 can be read only after you have become interested by reading 3.

A Webplatform for Philosophers

LogiLogi becomes an easy to use hypertext platform, also featuring a rating- and review-system which is a bit comparable to that found in journals. It tries to find the middle-road between the written word and a good conversation, and it's central values are openness and quality of content.

It makes commenting on texts, and more generally the linking of texts very easy. Most notably it also allows other people than the original author of an article to add outgoing links behind words, but it does not allow them to change the text itself, so the author's intellectual responsibility is guarded. Also important is that all conversations on the platform run via links (comparable to footnotes), not via forum-threads, avoiding their associated problems like fragmentation and shallowing of the discussion.

To maximize the advantages of hypertext, texts are kept short within LogiLogi, at maximum one to a few pages. They can be informal and experimental and they can be improved later on, in either of two ways: The text of the original document can be changed (earlier versions are then archived). Or secondly, links can be added inside the text, possibly only when some terms or concepts appear to be ambiguous, when questions arise, or when the text appears to arouse enough interest to make it worth of further elaboration.

Links in LogiLogi can refer to documents, to versions, and — by default — to tags (words that function as categories or concepts). Articles can be tagged with one or more of these tags. Multiple articles can have the same tag, and when a link is made to a tag or to a collection of tags, multiple articles can be in the set referred to. From this set the article with the highest rating is shown to the user.

In essence one can rate the articles of others by giving them a grade. The average of these grades forms the rating of the article. But this average is a weighted average. Voting-powers can vary. If an authors contributions are rated well, he receives more voting-power. Authors can thus gain 'status' and 'influence' through their work. This makes LogiLogi a peer-reviewed meritocracy, quite comparable to what we, according to Bruno Latours philosophy of science, encounter in the various structures surrounding journals (Lat87). Most notably this quality control by peer review, and it's accompanying social encouragement, was missing from earlier web-systems.

But the comparison goes further, and in a similar fashion to how new peergroups can emerge around new journals, in LogiLogi too new peergroups can be created by duplicating the just described rating-system. Contributions can be rated from the viewpoints of different peergroups, and therefore an article can have multiple ratings, authors won't have the same voting-power within each peergroup, and visitors can pick which peergroup to use as their filter. Thus except meritocratic, LogiLogi is also open to a diversity of schools and paradigms

in the sense of early Thomas Kuhn (Kuh96), especially as here creating new peergroups—unlike for journals—does not bring startup-costs.

Plato, Free Software and Postmodernism

The web is a relatively new medium, and new media are usually interpreted wrongly — in terms of old media. This is has been called the *horseless carriage syndrome* (McL01); according to which a car is a carriage without a horse, film records theater-plays, and—most recently—the web enables the downloading of journals. Even Plato was not exempt of this. In *Phaedrus* he stated that true philosophy is only possible verbally, and that writing was just an aid to memory. Regardless of this ironically enough his 'memory aid' unleashed a long philosophical tradition (dM05). New media take their time. And we should not forget that the web is still very young (1991). Also the web is especially relevant for philosophy in that it combines conversation and writing; the two classical media of philosophy.

And where previous mass-media like TV and radio were not suitable for philosophy, this was because they were *one to many*, and thus favored the *factory model of culture* (Ado91). The web on the other hand is *many to many*, and thereby enables something called *peer to peer* production (Ben06). An early example of this is Free Software: without much coordination tenthousands of volunteers have created software of the highest quality, like Linux and Firefox. Eric Raymond (Ray99) described this as a move from the *cathedral*- to the *bazaar*-model of software development. The *cathedral*-model has a single architect who is responsible for the grand design, while in the *bazaar*-model it evolves from collective contributions.

This *bazaar*-model is not unique for the web. It shares much with the academic tradition. The move from the book to the journal can be compared with a move in the direction of a *bazaar*-model. Other similarities are decentralized operation and peer-review. The only new thing of the Free Software example was it's use of the web which — through it's shorter turnaround times — is very suitable for *peer to peer* production.

Another development that LogiLogi follows closely is one within philosophy itself: Jean-Franois Lyotard in his *La Condition Postmoderne* proclaimed the end of great stories (Lyo79). Instead he saw a diversity of small stories, each competing with others in their own domains. Also Derrida spoke of the materiality of texts, where texts and intertextuality gave meaning instead of 'pure' ideas (Ber79; Nor87). The web in this sense is a radicalisation of postmodernism, allowing for even more and easier intertextuality.

And instead of trying to undo the proliferation of paradigms, as some logic-advocates tried, and still try, we think the *breakdown of language*—as in further segmentation—is here to stay, and even a good thing, because it reduces complexity in the sense of Niklas Luhmann (Blo97). Take human intelligence as fixed and you see that specialized (or 'curved' as in curved space) language allows for a more precise analysis. LogiLogi thus is explicitly modeled to allow for fine-grained specialization, and for a careful definition and discussion of terms *in context*.

Conclusion

To reiterate; LogiLogi will offer an easy to use hypertext-environment, and thanks to it's rating system a combination of quality and openness will be achieved: everyone can contribute, and even start new peergroups, but within peergroups quality is the determining factor. LogiLogi thus combines the informal, incremental and interactive qualities of good conversations, with conservation over time and space, as we traditionally know from the written word. LogiLogi is still very experimental.

Nevertheless what we can be sure about is that the web, as a medium that has proven to be very suitable for *peer to peer* production and that promises increased inter-textuality and differentiation of language, is bound to make a change for philosophy in the long run; with or without LogiLogi.

References

[Ado91] Theodor Adorno. Culture industry reconsidered. In Theodor Adorno, editor, *The Culture Industry: Selected Essays on Mass Culture*, pages 98–106. Routledge, London, 1991.

[Ben06] Yochai Benkler. *The Wealth of Networks*. Yale University Press, London, 2006.

[Ber79] Egide Berns. *Denken in Parijs: taal en Lacan, Foucault, Althusser, Derrida.* Samsom, Alpen aan den Rijn, 1979.

[Blo97] Christiaan Blom. *Complexiteit en Contingentie: een kritische inleiding tot de sociologie van Niklas Luhmann.* Kok Agora, Kampen, 1997.

[dM05] Jos de Mul. *Cyberspace Odyssee*. Klement, Kampen, 2005.

[ifb] http://www.futureofthebook.org. The Institute for the Future of the Book, MacArthur Foundation, University of Southern California.

[Kuh96] Thomas Kuhn. *The Structure of Scientific Revolutions*. University of Chicago Press, Chicago, 1996.

[Lat87] Bruno Latour. *Science in Action*. Open University Press, Cambridge, 1987.

[Lev97] Pierre Levy. *Collective Intelligence: Mankinds Emerging World in Cyberspace*. Plenum Press, NewYork, 1997.

[Log] http://foundation.logilogi.org. LogiLogi & The LogiLogi Foundation.

[Lyo79] Jean-Franois Lyotard. *La condition postmoderne: rapport sur le savoir.* Les ditions de Minuit, Paris, 1979.

[McL01] Marshall McLuhan. *Understanding Media:The Extensions of Man.* Routledge, London, 2001.

[Nel93] Ted Nelson. Literary Machines: *The report on, and of, Project Xanadu concerning word processing, electronic publishing, hypertekst, thinkertoys, tomorrow's intellectual. . . including knowledge, education and freedom.* Mindful Press, Sausalito, California, 1993.

[Nor87] Christopher Norris. *Derrida*. Fontana Press, London, 1987.

[Ray99] Eric S. Raymond. *The cathedral & the bazaar: musings on Linux and open source by an accidental revolutionary.* O'Reilly, Beijing, 1999.

Clinical Applications of Computer-assisted Textual Analysis: a Tei Dream?

Marco Zanasi
marco.zanasi@uniroma2.it
Università di Roma Tor Vergata, Italy

Daniele Silvi
d.silvi@email.it, silvi@lettere.uniroma2.it
Università di Roma Tor Vergata, Italy

Sergio Pizziconi
s.p@email.it
Università per Stranieri di Siena, Italy

Giulia Musolino
arcangelo.abbatemarco@fastwebnet.it
Università di Roma Tor Vergata, Italy

This research is the result of the collaboration of the Departments of Literature and Medicine of the University of Rome "Tor Vergata".

Marco Zanasi's idea of coding and studying dream reports of his patients at Tor Vergata University Hospital Psychiatric Unit started a team research on a challenging hypothesis: Is there any correlation between linguistic realization of dream reports and the psychopathology from which the dreamer is suffering? So far, analyses of dream reports have focused mainly on actants, settings, and descriptors of emotional condition during the oneiric activity.

The goal of "Dream Coding" is to collect, transcribe, catalogue, file, study, and comment on dreams reported by selected patients at Tor Vergata University Hospital in Rome. In the group of observed patients, different psychopathological statuses are represented.

The aim of the project is to build up a code which would enable a different reading of the reports that patients make of their dreams. A code that in the future could also provide new tools for initial and on-going diagnoses of the psychopathology affecting the patients.

Then we want to verify a set of linguistic features that can be significantly correlated to the type of psychopathology on a statistical basis. Numerous aspects, both theoretical and methodological, are discussed in this paper, ranging from the nature of the variation of language to be investigated to the tag set to be used in corpus preparation for computer analysis.

The first issue that the analysis has to solve is the speech transcription and how to obtain an accurate transcoded text from the oral form. After having considered several working hypothesis we have decided to use a "servile transcription" of

the speech: by recording the narrators' voice and transferring it in an electronic format.

At an early stage, we realized that some important elements of the speech could get lost in this process. This is why we have decided to reproduce the interruptions and hesitations. Interruptions have been classified on the basis of their length.

Starting from this consideration a form (enclosed) has been created in order to catalogue the dreams and other data, such as age, sex, address, birthday, job, and so on.

Furthermore the form contains information related to the transcriptor so as to detect the "noise" added by them. The noise can be partially isolated by noticing the continuity of some elements.

Obtained data is stored on the computer as a word file (.rtf format) to be analyzed later on. The working process has been divided into four steps, namely:

1. first step: in this step we collect the data of the patients that already have a diagnosis (che hanno già una diagnosi?)

2. second step: in this step we analyze the dreams collected in the previous phase and point out (individuare?) recurrent lexical structures (ricorrenti?)

3. third step: in this step we collect data from all other patients in order to analyze them and propose a kind of diagnosis

4. fourth step: in this step we make a comparison between the results of the first group of patients and the second one, and try to create a "vademecum" of the particularities noticed from a linguistic point of view. We call this classification "textual diagnosis", to be compared with the "medical diagnosis", in order to verify pendants and incongruities.

Although our research is at a preliminary stage, the statistical results on text analysis obtained so far allow us to make a few considerations. The first noticeable aspect is that differences between dreams of psychopathologic patients and controls are observed, which suggests the existence of specific features most likely correlated to the illness. We also observe a scale of narration complexity progressively decreasing from psychotic to bipolar patients and to controls. Since the patients recruited are in a remission phase, we may hypothesize that the variations observed are expression of the underlying pathology and thus not due to delusions, hallucinations, hypomania, mania and depression.

For such an accurate analysis the research group has started marking the texts of the dreams reports. The marking will allow the computer-aided analyses on semantics, syntax, rhetorical organization, metaphorical references and text architecture. Under the notation of "text architecture" we have included items such as coherence and cohesion and the anaphoric chains. The complete, quantitative results will be soon available, but it is possible to show at least the set of variables for which the texts will be surveyed, in order to account for the lexical variety we registered.

The syntactic and semantic opposition of NEW and GIVEN information is the basis according to which both formal and content items will be annotated. Therefore, as an example of the connection between formal and content items, the use of definite or indefinite articles will be linked to the introduction of the following new narrative elements:

- a new character is introduced in the plot, be it a human being or an animal;
- a shift to a character already in the plot;
- a new object appears in the scene;
- a shift to an object already in the plot;
- a shift of location without a verb group expressing the transfer.

The difficulty of recognizing any shift of time of the story without the appropriate time adverb or expression in the discourse, made it not relevant to register this variation since it would be necessarily correlated with the correct formal element.

The non-coherent and often non-cohesive structure of many patients' dreams shows to be relevant to determine what the ratios indicate as a richer lexicon. The consistent, abrupt introduction of NEW pieces of information, especially when correlated with the formal features, is likely to be one of the most significant differences between the control persons and patients, as well as between persons of the first two large psychopathologic categories analyzed here, bipolar disorders and psychotic disorders.

Our research is an ongoing project that aims to amplify samples and evaluate the possible influences of pharmacological treatments on the oniric characteristics being analyzed.

We have elaborated a measurement system for interruptions that we called "Scale of Silence"; it provides several special mark-ups for the interruptions during the speech. The scale (that is in a testing phase) has three types of tags according to the length of the interruption (0-2 sec; 2-4 sec; 4 or more). Special tags are also used for freudian slips and self-corrections the narrator uses.

The Chinese version of JGAAP supports Chinese word segmentation first then followed by a feature selection process at word level, as preparation for a later analytic phase. After getting a set of ordered feature vectors, we then use different analytical methods to produce authorship judgements. Unfortunately, the errors introduced by the segmentation method(s) will almost certainly influence the final outcome, creating a need for testing.

Almost all methods for Chinese word segmentation developed so far are either structural (Wang et al., 1991) and statistical-based (Lua, 1990).A structural algorithm resolves segmentation ambiguities by examining the structural relationship between words, while a statistical algorithm usually compares word frequency and character co-occurrence probability to detect word boundaries. The difficulties in this study are the ambiguity resolution and novel word detection (personal names, company names, and so on). We use a combination of Maximum Matching and conditional probabilities to minimize this error.

Maximum matching (Liu et al., 1994) is one of the most popular structural segmentation algorithms, the process from the beginning of the text to the end is called Forward Maximal Matching (FMM), the other way is called Backward Maximal Matching (BMM). A large lexicon that contains all the possible words in Chinese is usually used to find segmentation candidates for input sentences. Here we need a lexicon that not only has general words but also contains as many personal names, company names, and organization names as possible for detecting new words.

Before we scan the text we apply certain rules to divide the input sentence into small linguistic blocks, such as separating the document by English letters, numbers, and punctuation, giving us small pieces of character strings. The segmentation then starts from both directions of these small character strings. The major resource of our segmentation system is this large lexicon. We compare these linguistic blocks with the words in the lexicon to find the possible word strings. If a match is found, one word is segmented successfully. We do this for both directions, if the result is same then this segmentation is accomplished. If not, we take the one that has fewer words. If the number of words is the same, we take the result of BMM as our result. As an example : Suppose ABCDEFGH is a character string, and our lexicon contains the entries A, AB, ABC, but not ABCD. For FMM, we start from the beginning of the string (A) If A is found in the lexicon, we then look for AB in the lexicon. If AB is also found, we look for ABC and so on, till the string is not found. For example, ABCD is not found in the Lexicon, so we consider ABC as a word, then we start from character D unti the end of this character string. BMM is just the opposite direction, starting with H, then GH, then FGH, and so forth.

Suppose the segmentation we get from FMM is

(a) A \ B \ CD \ EFG \ H

and the segmentation from BMM is

(b) A \ B \ C \ DE \ FG \ H

We will take result (a), since it has fewer words. But if what we get from BMM is

(c) AB \ C \ DE \ FG \ H

We will take result (c), since the numbers of words is same in both method.

After the segmentation step we take the advantage of JGAAP's features and add different event sets according to the characteristics of Chinese, then apply statistical analysis to determine the final results. It is not clear at this writing, for example, if the larger character set of Chinese will make character-based methods more effective in Chinese then they are in other languages written with the Latin alphabet (like English). It is also not clear whether the segmentation process will produce the same type of set of useful "function words" that are so useful in English authorship attribution. The JGAAP structure (Juola et al, 2006;Juola et al., submitted), however, will make it easy to test our system using a variety of different methods and analysis algorithms.

In order to test the performance on Chinese of our software, we are in the process of constructing a Chinese test corpus. We will select three popular novelists and ten novels from each one, eight novels from each author will be used as training data, the other two will be used as testing data. We will also test on the blogs which will be selected from internet. The testing procedure will be the same as with the novels.

This research demonstrates, first, the JGAAP structure can easily be adapted to the problems of non-Latin scripts and not English languages, and second, provides somes cues to the best practices of authorship attribution in Chinese. It can hopefully be extended to the development of other non-Latin systems for authorship attribution.

References:

Patrick Juola, (in press). *Authorship Attribution*. Delft:NOW Publishing.

Patrick Juola, John Noecker, and Mike Ryan. (submitted). "JGAAP3.0: Authorship Attribution for the Rest of Us." Submitted to DH2008.

Patrick Juola, John Sofko, and Patrick Brennan. (2006). "A Prototype for Authorship Attribution Studies." *Literary and Linguistic Computing* 21:169-178

Yuan Liu, Qiang Tan, and Kun Xu Shen. (1994). "The Word Segmentation Rules and Automatic Word Segmentation Methods for Chinese Information Processing" (in Chinese). Qing Hua University Press and Guang Xi Science and Technology Press, page 36.

K.T. Lua. (1990). From Character to Word. An Application of Information Theory. *Computer Processing of Chinese and Oriental Languages,* Vol. 4, No. 4, pages 304--313, March.

Liang-Jyh Wang, Tzusheng Pei, Wei-Chuan Li, and Lih-Ching R. Huang. (1991). "A Parsing Method for Identifying Words in Mandarin Chinese Sentences." In *Processings of 12th International Joint Conference on Artificial Intelligence*, pages 1018--1023, Darling Harbour, Sydney, Australia, 24-30 August.

Automatic Link-Detection in Encoded Archival Descriptions

Junte Zhang
j.zhang@uva.nl
University of Amsterdam, The Netherlands

Khairun Nisa Fachry
k.n.fachry@uva.nl
University of Amsterdam, The Netherlands

Jaap Kamps
j.kamps@uva.nl
University of Amsterdam, The Netherlands

In this paper we investigate how currently emerging link detection methods can help enrich encoded archival descriptions. We discuss link detection methods in general, and evaluate the identification of names both within, and across, archival descriptions. Our initial experiments suggest that we can automatically detect occurrences of person names with high accuracy, both within (F-score of 0.9620) and across (F-score of 1) archival descriptions. This allows us to create (pseudo) encoded archival context descriptions that provide novel means of navigation, improving access to the vast amounts of archival data.

Introduction

Archival finding aids are complex multi-level descriptions of the paper trails of corporations, persons and families. Currently, finding aids are increasingly encoded in XML using the standard Encoded Archival Descriptions. Archives can cover hundreds of meters of material, resulting in long and detailed EAD documents. We use a dataset of 2,886 EAD documents from the International Institute of Social History (IISG) and 3,119 documents from the Archives Hub, containing documents with more than 100,000 words. Navigating in such archival finding aids becomes non-trivial, and it is easy to loose overview of the hierarchical structure. Hence, this may lead to the loss of important contextual information for interpreting the records.

Archival context may be preserved through the use of authority records capturing information about the record creators (corporations, persons, or families) and the context of record creation. By separating the record creator's descriptions from the records or resources descriptions themselves, we can create "links" from all occurrences of the creators to this context. The resulting descriptions of record creators can be encoded in XML using the emerging Encoded Archival Context (EAC) standard. Currently, EAC has only been applied experimentally. One of the main barriers to adoption is that it requires substantial effort to adopt EAC. The information for the creator's authority record is usually available in some

form (for example, EAD descriptions usually have a detailed field <bioghist> about the archive's creator). However, linking such a context description to occurrences of the creator in the archival descriptions requires more structure than that is available in legacy data.

Our main aim is to investigate if and how automatic link detection methods could help improve archival access. Automatic link detection studies the discovery of relations between various documents. Such methods have been employed to detect "missing" links on the Web and recently in the online encyclopedia Wikipedia. Are link detection methods sufficiently effective to be fruitfully applied to archival descriptions? To answer this question, we will experiment on the detection of archival creators within and across finding aids. Based on our findings, we will further discuss how detected links can be used to provide crucial contextual information for the interpretation of records, and to improve navigation within and across finding aids.

Link Detection Methods

Links generated by humans are abundant on the World Wide Web, and knowledge repositories like the online encyclopedia Wikipedia. There are two kinds of links: incoming and outgoing links. Substrings of text nodes are identified as *anchors* and become clickable. Incoming links come from text nodes of target files (destination node) and point to a source file (origin node), while an outgoing link goes from text node in the source document (origin node) to a target file (destination node). Two assumptions are made: a link from document *A* to document *B* is a recommendation of *B* by *A*, and documents linked to each other are related.

To automatically detect whether two nodes are connected, it is necessary to search the archives for some string that both share. Usually it is only one specific and extract string. A general approach to automatic link detection is first to detect the *global similarity* between documents. After the relevant set of documents has been collected, the *local similarity* can be detected by comparing text segments with other text segments in those files. In structured documents like archival finding aids, these text segments are often marked up as logical units, whether it be the title <*titleproper*>, the wrapper element <*c12*> deep down in the finding aid, or the element <*persname*> that identifies some personal names. These units are identified and retrieved in XML Element retrieval. The identification of relevant anchors is a key problem, as these are used in the system's retrieval models to point to (parts of) related finding aids.

Experiment: Name Detection

A specific name detection trial with the archive of Joop den Uyl (1919-1987), former Labor Party prime minister of the Netherlands, is done as a test to deal with this problem. This archive consists of 29,184 tokens (with removal of the XML markup and punctuation), of which 4,979 are unique, and where a token is a sequence of non-space characters. We collect a list of the name variants that we expect to encounter: "J.M. Den Uyl", "Joop M. Den Uyl", "Johannes Marten den Uyl", "Den Uyl", etc. We construct a regular expression to fetch the name variants. The results are depicted in illustration 1, which shows the local view of the Joop den Uyl archive in our *Retrieving EADs More Effectively* (README) system.

Illustration 1: Links detected in EAD

The quality of the name detection trial is evaluated with explicit feedback, which means manually checking the detected links for (1) correctness, (2) error, and (3) whether any links were missing. This was done both within finding aids, and across finding aids:

- First, the quality is checked within finding aids, by locating occurrences of creator Joop den Uyl in his archive. For detecting name occurrences within an archive, our simple method has a precision of (114/120 =) 0.9500, a recall of (114/117 =) 0.9744, resulting in an F-score of 0.9620. Some interesting missing links used name variants where the prefix "den" is put behind the last name "Uyl" -- a typical Dutch practice. Incorrect links mostly are family members occurring the archive, e.g., "Saskia den Uyl", "E.J. den Uyl-van Vessem", and also "Familie Den Uyl". Since these names occur relatively infrequent, few errors are made. The matching algorithm could easily be refined based on these false positives.

Table 1: Archive "Den Uyl"

	Link	No link
Name	114	3
No name	6	-

- Second, the same procedure to detect proper names of Joop den Uyl is applied across finding aids with the related archive of "Partij van de Arbeid Tweede-Kamer Fractie (1955-1988)" (Dutch MPs from the Labor Party). For detecting name occurrences across archives, we obtain a perfect precision, recall, and thus F-score of 1.

Table 2: Archive "PvdA"

	Link	No link
Name	16	0
No name	0	-

Concluding Discussion

In this paper we investigated how currently emerging link detection methods can help enrich encoded archival descriptions. We discussed link detection methods in general, and evaluated the identification of names both within, and across, archival descriptions. Our initial experiments suggest that we can automatically detect occurrences of person names, both within (F-score of 0.9620) and across (F-score of 1) archival descriptions. This allows us to create (pseudo) encoded archival context (EAC) descriptions that provide novel means of navigation and improve access to archival finding aids. The results of our experiments were promising, and can also be expanded to names of organizations, events, topics, etc. We expect those to be more difficult than personal name detection.

There are more uses for detecting cross-links in finding aids besides creating extra contextual information. Detecting missing links is useful for improving the retrieval of separate finding aids, for example, an archival finding aid with many detected incoming links may have a higher relevance. Links can also offer a search-by-example approach, like given one finding aids, find all related finding aids. A step further is to use the cross-links in the categorization of archival data. Concretely for historians and other users, who rely on numerous lengthy archival documents, new insights can be gained by detecting missing cross-links.

Acknowledgments

This research is supported by the Netherlands Organization for Scientific Research (NWO) grant # 639.072.601.

References

Agosti, M., Crestani, F., and Melucci, M. 1997. On the use of information retrieval techniques for the automatic construction of hypertext. *Information Processing and Management* 33, 2 (1997), 133-144.

Allan, J. 1997. Building hypertext using information retrieval. *Information Processing and Management* 33, 2 (1997), 145-159.

EAC, 2004. Encoded Archival Context. http://www.iath.virginia.edu/eac/

EAD, 2002. Encoded Archival Description. http://www.loc.gov/ead/

Fissaha Adafre, S. and De Rijke, M. 2005. Discovering missing links in Wikipedia. In *Proceedings of the 3rd international Workshop on Link Discovery.* LinkKDD '05. ACM Press, 90-97.

Huang, W. C., Trotman, A., and Geva, S. 2007. Collaborative Knowledge Management: Evaluation of Automated Link Discovery in the Wikipedia. In *Proceedings of the SIGIR 2007 Workshop on Focused Retrieval*, 2007.

INEX LTW, 2007. INEX Link The Wiki Track, 2007. http://inex.is.informatik.uni-duisburg.de/2007/linkwiki.html

ISAAR (CFP), 2004. *International Standard Archival Authority Record for Corporate bodies, Persons and Families.* International Council on Archives, Ottawa, second edition, 2004.

ISAD(G), 1999. *General International Standard Archival Description.* International Council on Archives, Ottawa, second edition, 1999.

Jenkins, N., 2007. Can We Link It. http://en.wikipedia.org/wiki/User:Nickj/Can_We_Link_It

A Chinese Version of an Authorship Attribution Analysis Program

Mengjia Zhao
zhaom@duq.edu
Duquesne University, USA

Patrick Juola
juola@mathcs.duq.edu
Duquesne University, USA

Authorship Attribution (Juola, in press) is a form of text analysis to determine the author of a given text. Authorship Attribution in Chinese (AAC) is that task when the text is written in Chinese. It can be considered as a typical classification problem, where a set of documents with known authorship are used for training and the aim is to automatically determine the corresponding author of an anonymous text. Beyond simple questions of the identity of individual authors, the underlying technology may also apply to gender, age, social class, education and nationality.

JGAAP (Java Graphical Authorship Attribution Program) (Juola et al., 2006) is a program aimed at automatically determining a document's author by using corpus linguistics and statistical analysis. It performs different types of analysis and gives the best results to the user while hiding the detailed methods and techniques involved. It can therefore be used by non-experts. We extend JGAAP to cover the special issues involved in Chinese attribution and present the results of some experiments involving novels and blogs.

Why can't we just use the existing JGAAP software? Chinese introduces its own problems, caused by the different structures of English and Chinese. As with English, Chinese texts are composed of sentences, sentences are composed of words, and words are composed of characters. However, the Chinese character set is approximately 50,000 individually meaningful characters, compared with fifty or so meaningless symbols for English. Furthermore, in Chinese texts words are not delimited by spaces as they are in English. As with English, the word is the basic meaningful unit in Chinese, but the meaning of a word may differ from the meaning of the characters compose this word.

Analysis at the character level is thus fundamentally different between the two languages. Studies in English show that analysis at the word level is likely to be a better way to understand the style and linguistic features of a document, but it is not clear whether this will apply to Chinese as well. So before we can analyze word-level features (for comparison) we need to segment sentences at word-level not by characters. Therefore the first step for any Chinese information processing system is the automatically detection of word boundaries and segmentation.

The Chinese version of JGAAP supports Chinese word segmentation first then followed by a feature selection process at word level, as preparation for a later analytic phase. After getting a set of ordered feature vectors, we then use different analytical methods to produce authorship judgements. Unfortunately, the errors introduced by the segmentation method(s) will almost certainly influence the final outcome, creating a need for testing.

Almost all methods for Chinese word segmentation developed so far are either structural (Wang et al., 1991) and statistical-based (Lua, 1990). A structural algorithm resolves segmentation ambiguities by examining the structural relationship between words, while a statistical algorithm usually compares word frequency and character co-occurrence probability to detect word boundaries. The difficulties in this study are the ambiguity resolution and novel word detection (personal names, company names, and so on). We use a combination of Maximum Matching and conditional probabilities to minimize this error.

Maximum matching (Liu et al., 1994) is one of the most popular structural segmentation algorithms, the process from the beginning of the text to the end is called Forward Maximal Matching (FMM), the other way is called Backward Maximal Matching (BMM). A large lexicon that contains all the possible words in Chinese is usually used to find segmentation candidates for input sentences. Here we need a lexicon that not only has general words but also contains as many personal names, company names, and organization names as possible for detecting new words.

Before we scan the text we apply certain rules to divide the input sentence into small linguistic blocks, such as separating the document by English letters, numbers, and punctuation, giving us small pieces of character strings. The segmentation then starts from both directions of these small character strings. The major resource of our segmentation system is this large lexicon. We compare these linguistic blocks with the words in the lexicon to find the possible word strings. If a match is found, one word is segmented successfully. We do this for both directions, if the result is same then this segmentation is accomplished. If not, we take the one that has fewer words. If the number of words is the same, we take the result of BMM as our result. As an example : Suppose ABCDEFGH is a character string, and our lexicon contains the entries A, AB, ABC, but not ABCD. For FMM, we start from the beginning of the string (A) If A is found in the lexicon, we then look for AB in the lexicon. If AB is also found, we look for ABC and so on, till the string is not found. For example, ABCD is not found in the Lexicon, so we consider ABC as a word, then we start from character D unti the end of this character string. BMM is just the opposite direction, starting with H, then GH, then FGH, and so forth.

Suppose the segmentation we get from FMM is

(a) A \ B \ CD \ EFG \ H

and the segmentation from BMM is

(b) A \ B \ C \ DE \ FG \ H

We will take result (a), since it has fewer words. But if what we get from BMM is

(c) AB \ C \ DE \ FG \ H

We will take result (c), since the numbers of words is same in both method.

After the segmentation step we take the advantage of JGAAP's features and add different event sets according to the characteristics of Chinese, then apply statistical analysis to determine the final results. It is not clear at this writing, for example, if the larger character set of Chinese will make character-based methods more effective in Chinese then they are in other languages written with the Latin alphabet (like English). It is also not clear whether the segmentation process will produce the same type of set of useful "function words" that are so useful in English authorship attribution. The JGAAP structure (Juola et al, 2006;Juola et al., submitted), however, will make it easy to test our system using a variety of different methods and analysis algorithms.

In order to test the performance on Chinese of our software, we are in the process of constructing a Chinese test corpus. We will select three popular novelists and ten novels from each one, eight novels from each author will be used as training data, the other two will be used as testing data. We will also test on the blogs which will be selected from internet. The testing procedure will be the same as with the novels.

This research demonstrates, first, the JGAAP structure can easily be adapted to the problems of non-Latin scripts and not English languages, and second, provides somes cues to the best practices of authorship attribution in Chinese. It can hopefully be extended to the development of other non-Latin systems for authorship attribution.

References:

Patrick Juola, (in press). *Authorship Attribution*. Delft:NOW Publishing.

Patrick Juola, John Noecker, and Mike Ryan. (submitted). "JGAAP3.0: Authorship Attribution for the Rest of Us." Submitted to DH2008.

Patrick Juola, John Sofko, and Patrick Brennan. (2006). "A Prototype for Authorship Attribution Studies." *Literary and Linguistic Computing* 21:169-178

Yuan Liu, Qiang Tan, and Kun Xu Shen. (1994). "The Word Segmentation Rules and Automatic Word Segmentation Methods for Chinese Information Processing" (in Chinese). Qing Hua University Press and Guang Xi Science and Technology Press, page 36.

K.T. Lua. (1990). From Character to Word. An Application of Information Theory. *Computer Processing of Chinese and Oriental Languages,* Vol. 4, No. 4, pages 304--313, March.

Liang-Jyh Wang, Tzusheng Pei, Wei-Chuan Li, and Lih-Ching R. Huang. (1991). "A Parsing Method for Identifying Words in Mandarin Chinese Sentences." In *Processings of 12th International Joint Conference on Artificial Intelligence*, pages 1018--1023, Darling Harbour, Sydney, Australia, 24-30 August

Posters

The German Hamlets: An Advanced Text Technological Application

Benjamin Birkenhake
University of Bielefeld, Germany

Andreas Witt
andreas.witt@uni-tuebingen.de
University of Tübingen, Germany

The highly complex editorial history of Shakespeare's Hamlet and its many translations to German forms a interesting corpus to show some of the major advantages current texttechnologies. In the following we want to present the corpus of four English editions an several German translations of the play we have gathered together and annotated and crosslinked in different ways.

Although Shakespeare's Hamlet is obviously not a unique hypertext, it is an interesting object to test advanced hypertext- and text technologies. There is no original edition of Hamlet, which was authorized by Shakespeare during his lifetime. We only have different print editions, which all have a different status concerning their quality, overall length, content and story-line. The most important among these are the so called first folio, the first quattro and the second quattro edition of Hamlet. During the centuries editors tried to combine these early editions to the best edition possible. The famous Arden-editons as well as the in the internet widespread Moby-editon are such compositions.

A comparable but a bit more complex situation exists within in the field of german translations of the play. The earliest translation is by Christoph Martin Wieland from about 1766. After this at least 18 translations have been published which are accompanied by countless translations of theatre directors, which are mostly not documented. The corpus contains 8 digitalized Translations. 2 further translation are already scanned but not yet digitalized, because they are printed in fraktur - a old german typeface - which can not be recognized by common OCR-programs yet. The remaining 10 Translations are available in print, but not yet digitalized, too. Of the 8 digitalized translations we chose 4 for further text technological use.

What makes the corpus so interesting is the fact, that almost every translator used several of the early english editions as a basis for a new translation. This leads to a situation in which almost every german or english edition of Shakespeare's Hamlet is a composition of several sources. The relation the editions have with their sources and with each other form a wide network, which could be presented in a hypertext system.

Another interesting aspect of Shakespeare's Hamlet is the outstanding position the play has within the western culture for centuries. Hamlet is the single most researched piece of literature, has been analyzed from various perspectives and is a part of western common education. This leads to the request, that a digital environment should represent the variety of perspectives on the play. This lead us to a corpus of Hamlet editions in which each text may exist in multiple forms.

Basis for the XML-annotations are text files, which are transformed to XML using regular expressions. The basic XML-format is TEI 4 drama base tag set. TEI 4 is a major open source concept of the Text Encoding Initiative. The drama base tag set offers almost all tags needed for a general, formal annotation of a play. In order to provide an easy to annotate mechanism we added some attributes to represent the translation- or origin-relation between lines, paragraphs or speeches within the editions on the one hand and the sources on the other hand.

The TEI-annotated documents are used for further annotations an presentation. The TEI-documents were automatically enriched with further markup, using an open source auto-tagger. This auto-tagger annotates single words, including the part of speech and the principle form. The TEI-documents are also the basis for the XHTML-presentation. As the TEIstructure contains all information necessary for a graphical presentation, these documents are transformed to XHTML, which is used to present the corpus. This transformation is made with several XSLT-Stylesheets. In the same way XSLFO is used to generate PDF-versions of each edition.

edition	Txt	TEI	XHTML	STTS	Narration
1st Folio	x	x	x		
1st Quattro	x	x	x		
2nd Quattro	x	x	x		
Moby	x	x	x		
Wieland	x	x	x		x
Schlegel	x	x	x	x	x
Fontane	x	x	x	x	x
Hauptmann	x	x	x	x	

Table 1: The different layers of the mayor editions within the corpus

In many cases translators have re-arranged the flow of stanzas or the course of action. Therefore it is useful to provide an alternative linking mechanism, which does not only focus on the language and the formal structure, but also on the plot. To provide this reference the narrative information is annotated in another layer. This allows to find the same event in different translations of the play. The narrative annotation layer basically consists of events, which can be seen as the smallest elements of the plot.

Obviously, events may start within one line and end several lines or even speeches later. Since the narrative structure is overlapping with the TEI, both are stored in separate

annotations. Scenes can provide a meaningful unit for basic parts of the plot. Thus the formal and the narrative annotation are semantically aligned - in addition to their reference on identical textual data. This relation can be exploited by creating links between the concept of a scene and the concept of specific actions. The respective linking mechanism is located on a meta level: it operates on the schemas themselves and not on their instances. The references are generated mechanically on the meta level, linking different perspectives together. Readers can explore the relations between events and scenes. The procedure could also be used to create a recommendation system as e.g. proposed by Macedo et al. (2003): the annotation integrates the knowledge of experts on narrative structures in the play Hamlet and provides this information to the reader. This leads to a multi rooted tree, each tree represents one level of information, i.e. textual structure and linguistic, philological or narrative information. This allows for creating a network of multiple perspectives on one text being linked to one another. As a result, hypertext is no longer based on links between nodes, but offers a reference mechanism between perspectives.

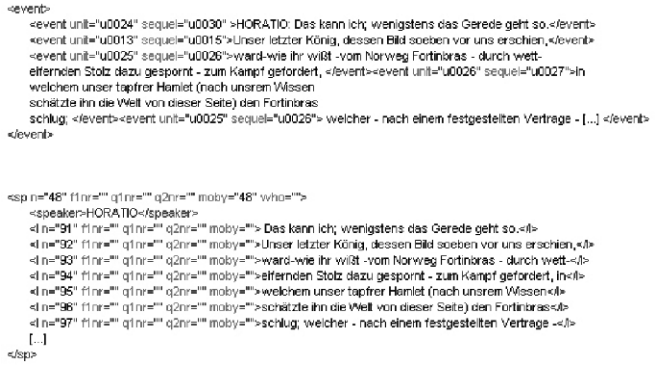


Figure 1: A multi rooted tree above a single textual data

As a first result of these multiple annotations, we got a corpus that is based on XML-technology and available via the web. As a second result we developed methods to cope with multiple annotated documents, which is a task, that has to be performed more often with the growing popularity of XML-technologies. Especially the integration of the narration annotation layer has to bee seen as a example for further parallel annotations. In detail these methods described above lead to an environment, which offers different types of user different perspectives on a single, textual object or a corpus. Some of these benefits will be presented in the following:

1. The common TEI-annotation allows a structural linking-mechanism between the editions. This allows a user to jump from the first scene in the second act of one edition to the same scene in another edition.

2. Alternatively this annotation can be used to present the user a part of the play in on or more editions of his choice for direct comparison.

3. The narration annotation layer allows several ways to explore a single text or compare some texts with each other. In the first case, the annotation of events and actions provides a way of comparing different editions esp. translations.

4. Using SVG - an XML-based format for graphics - the narrative structure of each translation could be visualized, ignoring the textual basis. This gives an »overview« of plot of the current edition.

5. The introduced concept of cross annotation linking allows us to offer the user automatically generated links from one annotation to another.

With this set of different linking-concepts we offer users a new freedom to explore the corpus in a way that fits to their needs. We ensured, that every layer of information offers a way to access information of another layer in a different perspective. We assume that this method can be transferred to any kind of multiple annotation.

References

Macedo, A. A., Truong, K.N. and Camacho-Guerrero, J. A. (2003). Automatically Sharing Web Experiences through a Hyperdocument Recommender System. In: Proceedings of the 14th conference on Hypertext and Hypermedia (Hypertext03), (Nottingham, UK, August, 26-30, 2003). Online: http://www.ht03.org/papers/pdfs/6.pdf [Available, Last checked: 6/16/2005]

Fine Rolls in Print and on the Web: A Reader Study

Arianna Ciula
arianna.ciula@kcl.ac.uk
King's College London, UK

Tamara Lopez
tamara.lopez@kcl.ac.uk
King's College London, UK

Introduction

A collaboration between the National Archives in the UK, the History and Centre for Computing in the Humanities departments at King's College London, the Henry III Fine Rolls project (http://www.frh3.org.uk) has produced both a digital and a print edition (the latter in collaboration with publisher Boydell & Brewer) of the primary sources known as the Fine Rolls. This dual undertaking has raised questions about the different presentational formats of the two resources and presented challenges for the historians and digital humanities researchers involved in the project, and, to a certain extent, for the publisher too. These challenges, and the adopted solutions for the two types of published resource present a novel starting point from which to examine how the artefacts of digital humanities are used.

This poster will report on the progress to-date of an ongoing exploratory study that examines the information-seeking behavior of historians, with the aim of developing a clearer picture of how research is conducted using sources like the Calendar of Fine Rolls. As with other digital humanities efforts (Buchanan, Cunningham, Blandford, Rimmer, & Warwick; 2005), this study focuses on the ways in which interactions occur using physical and virtual environments, and in particular on the ways in which the components of hybrid scholarly editions are used in combination to answer research questions. A secondary pragmatic component of the study seeks to adapt methodologies from the fields of information seeking research and human-computer interaction to the evaluation of digital humanities and in particular, to the Fine Rolls of Henry III project resources.

Reading the Fine Rolls of Henry III

The two publication formats of the Fine Rolls of Henry III are drawn from the same data substrate. Given the nature of the materials, *reading* is expected to be the primary activity performed using both. A stated design goal for the project is that the two publications will form a rich body of materials with which to conduct historical research. Our first research goal is to establish the context for work done using historical sources like the Fine Rolls: to establish the kinds of research questions asked, to articulate the methods followed in answering these questions, and to develop profiles of the researchers who perform the work.

Given the heterogeneity of the materials and the different focus of each medium, the question arises whether the materials do in fact form a single body of work, and how the experience of using them comprises a continuous experience. It suggests that working with the texts will of necessity also involve periods of *information seeking*: moments encountered while reading that give rise to questions which the material at hand cannot answer and the subsequent process embarked upon in order to answer them. We hypothesize that to fill these information gaps, readers of the Fine Rolls will seek particular text in the alternative medium to find answers. To answer a question about a translation found in a printed volume, for example, we suggest that the researcher will seek the image of the membrane in the web edition in order to consult the original language of the text.

Having established the impetus for using the two formats together, one questions the effectiveness of particular features of each medium in creating *bridges* between the formats. One implicit design goal of the project has been to optimize the movement between states of reading (within a medium) and moments of seeking (between media). Our final research goal, therefore is to discover the ways in which design choices taken in each facilitate or hinder movement between the website and the books, thereby enriching or diminishing the utility of the body of work as a whole.

Methodology

As suggested above, our analytical framework for evaluating the experience of reading the Fine Rolls of Henry III is drawn from the field of information seeking behavior research, and in particular by the conceptual framework developed in Brenda Dervin's *Sense-Making Theory* (Dervin, 1983 as summarized by Wilson; Wilson, 1999).

Our methodological approach will similarly be guided by data collection methods established in this field. Structured in three phases, our study will first identify a representative sample of scholars from leading institutions who perform research using sources like the Fine Rolls. In the first phase of data collection, we intend to use established field and (where necessary) online questionnaires to establish the kinds of research questions asked of materials like the Fine Rolls, and to articulate the methods followed in answering these questions. Given the hybrid nature of the Fine Rolls edition, we will seek to elicit comments on the use of secondary, primary and surrogate (web, microfilm and microfiche) formats.

A phase of analysis will follow in which we develop profiles of the researchers who perform this kind of work to identify the information seeking features that exist within materials related to a range of different tools, such as indexes, search engines, front matter materials, printed catalogues and finding aids, and the perceived strengths and weaknesses of doing work with existing support materials. As the final part of this stage, we will identify a series of open-ended research questions that can

be answered using the Fine Rolls materials. These questions will be formulated to encourage use of a range of formats and informational features of the components that comprise the Fine Rolls resource.

Given our focus on a single scholarly edition and our corresponding pragmatic need to develop an understanding of the effectiveness and utility of same, the third phase of our work will utilize established usability testing techniques to evaluate project resources. Drawing upon the profiles developed in the first phase of data collection, a sample of representative users of the Fine Rolls materials will be selected. In a series of guided, task-based sessions, participants will be asked to answer the research questions formulated during analysis. Data collection, expected to include a combination of direct observation and interview will be used to quantitatively identify the information features of both editions that are used to answer research questions. To generate a more qualitative assessment of the effectiveness of the features, users will be encouraged to "think-aloud" (Preece, Rogers, & Sharp.; 2002, p. 365) about the process and their observations will be recorded.

Sample size permitting, sessions are to be held to evaluate use of the book only, the web materials only, and both web and book.

Conclusions

With this research, we hope to elicit specific information about design improvements that can be made to support information seeking activities that span the Fine Rolls digital and print materials, and to articulate general heuristics that can be used in the design of other hybrid digital humanities publications. With our focus on a single scholarly edition, we also contribute to work begun elsewhere (Buchanan, et. al.; 2005) to explore how established methods for evaluating the use of digital materials can be adapted and applied to the work of humanities researchers. Finally, we contribute to understanding of the evolution of the scholarly edition as a resource that extends beyond the self-contained print edition, and of the deepening interdependence between humanities research activities in digital and traditional environments.

Notes

[1] The first volume was published in September 2007 (Dryburgh et al. 2007).

[2] Our understanding here will both draw upon and contribute to the dialog regarding use of electronic sources by humanities scholars begun elsewhere, see e.g. Bates, Wiberley, Buchanan, et. al.

Bibliography

Bates, M.; Wilde, D. & Siegfried, S. (1995), 'Research practices of humanities scholars in an online environment: The Getty online searching project report no. 3', *Library and Information Science Research* 17(1), 5--40.

Buchanan, G.; Cunningham, S.; Blandford, A.; Rimmer, J. & Warwick, C. (2005), 'Information seeking by humanities scholars', *Research and Advanced Technology for Digital Libraries, 9th European Conference, ECDL*, 18--23.

Dervin, B. (1983), 'An overview of sense-making research: Concepts, methods, and results to date', *International Communication Association Annual Meeting*, Dallas, TX.

Dryburgh, P. & Hartland, B. eds. Ciula A. and José Miguel Vieira tech. Eds. (2007) *Calendar of the Fine Rolls of the Reign of Henry III [1216-1248]*, vol. I: 1216-1224, Woodbridge: Boydell & Brewer.

Jones, A. ed. (2006) *Summit on Digital Tools for the Humanities: Report on Summit Accomplishments*. Charlottesville, VA: University of Virginia.

Preece, J.; Rogers, Y. & Sharp, H. (2002), *Interaction design: beyond human-computer interaction*, Wiley.

Siemens, R.; Toms, E.; Sinclair, S.; Rockwell, G. & Siemens, L. 'The Humanities Scholar in the Twenty-first Century: How Research is Done and What Support is Needed.' *ALLC/ACH 2004 Conference Abstracts*. Göteborg: Göteborg University, 2004. <http://www.hum.gu.se/allcach2004/AP/html/prop139.html>

Wiberley, S. & Jones, W. (2000), 'Time and technology: A decade-long look at humanists' use of electronic information technology', *College & Research Libraries* 61(5), 421--431.

Wiberley, S. & Jones, W. (1989), 'Patterns of information seeking in the humanities', *College and Research Libraries* 50(6), 638--645.

Wilson, T. (1999), 'Models in information behaviour research', *Journal of Documentation 55(3), 249–270*.

PhiloMine: An Integrated Environment for Humanities Text Mining

Charles Cooney
cmcooney@diderot.uchicago.edu,
University of Chicago, USA

Russell Horton
russ@diderot.uchicago.edu
University of Chicago, USA

Mark Olsen
mark@barkov.uchicago.edu
University of Chicago, USA

Glenn Roe
glenn@diderot.uchicago.edu
University of Chicago, USA

Robert Voyer
rlvoyer@diderot.uchicago.edu
University of Chicago, USA

PhiloMine [http://philologic.uchicago.edu/philomine/] is a set of data mining extensions to the PhiloLogic [http://philologic.uchicago.edu/] full-text search and retrieval engine, providing middleware between PhiloLogic and a variety of data mining packages that allows text mining experiments to be run on documents loaded into a PhiloLogic database. We would like to present a poster describing and demonstrating how PhiloMine works.

The text mining process under PhiloMine has three main components -- coprus selection, feature selection and algorithm selection. Experimental corpora can be constructed from the documents in the PhiloLogic database using standard bibliographic metadata criteria such as date of publication or author gender, as well as by attributes of sub-document level objects such as divs and paragraphs. This makes it possible, for example, to compare poetry line groups in male-authored texts from 1800 - 1850 with those in female-authored texts from that period or any other. The PhiloMine user then selects the features to use for the experiment, choosing some or all feature sets including surface forms, lemmas, part-of-speech tags, and bigrams and trigrams of surface forms and lemmas. Once the corpus and feature sets are selected, the machine learning algorithm and implementation is chosen. PhiloMine can talk to a range of freely available data mining packages such as WEKA, Ken William's Perl modules, the CLUTO clustering engine and more. Once the learning process has executed, the results are redirected back to the browser and formatted to provide links to PhiloLogic display of the documents involved and queries for the individual words in each corpus. PhiloMine provides an environment for the construction, execution and analysis of text mining experiments by bridging the gap between the source documents, the data structures that form the input to the learning process, and the generated models and classifications that are its output.

Corpus Selection

Under PhiloMine, the text mining corpus is created by selecting documents and sub-document level objects from a particular PhiloLogic database. PhiloLogic can load documents in a number of commonly used formats such as TEI, RTF, DocBook and plain text. From all documents in a particular PhiloLogic collection, the text mining corpus is selected using bibliographic metadata to choose particular documents, and sub-document object selectors to choose objects such as divs and line groups. The two levels of criteria are merged, so that the experimenter may easily create, for example, a corpus of all divs of type "letter" appearing within documents by female authors published in Paris in the 19th century. For a supervised mining run, the PhiloMine user must enter at least two sets of such criteria, and a corpus is created which contains multiple sub-corpora, one for each of the classes. For an unsupervised mining run, such as clustering, one corpus is created based on one set of criteria.

The PhiloMine user is also able to specify the granularity of text object size which is presented to the learning algorithm, the scope of the "instance" in machine learning terminology. Under PhiloMine, an instance may be either an entire document, a div or a paragraph. A single document consisting of a thousand paragraphs may be presented to a machine learner as a thousand distinct text objects, a thousand vectors of feature data, and in that case the learner will classify or cluster on the paragraph level. Similarly, even if the user has chosen to use sub-document level criteria such as div type, the selected text objects can be combined to reconstitute a document-level object. Thus the PhiloMine experimenter can set corpus criteria at the document, div and/or paragraph level and independently decide which level of text object to use as instances.

Several filters are available to ensure that selected text objects suit the experimental design. PhiloMine users may set minimum and maximum feature counts per instance. They may also balance instances across classes, which is useful to keep your machine learner honest when dealing with classifiers that will exploit differential baseline class frequencies. Finally, instance class labels may be shuffled for a random falsification run, to make sure that your accuracy is not a result of an over-fitting classifier.

Feature Selection

In machine learning generally, features are attributes of an instance that take on certain values. The text mining process often involves shredding the documents in the corpus into a bag-of-words (BOW) representation, wherein each unique word, or type, is a feature, and the number of tokens, or

occurrences of each type in a given document, is the value of that feature for that document. This data structure can be envisioned as a matrix, or spreadhsheet, with each row corresponding to a text object, or instance, and each column representing a type, or feature, with an extra column for class label if supervised learning is being undertaken. PhiloMine generates a BOW matrix for the user-selected corpus which serves as the input to the machine learner.

Because PhiloLogic creates extensive indices of document content as part of its loading process, its internal data structures already contain counts of words for each document in a given database. PhiloMine extends PhiloLogic so that is available for divs and paragraphs. In addition to the surface forms of words, PhiloMine will also generate vectors for lemmas or parts-of-speech tags, provided by TreeTagger, and bigrams and trigrams of surface forms or lemmas. The user may select one or more of these feature sets for inclusion in a given run.

One practical concern in machine learning is the dimensionality of the input matrix. Various algorithms scale in different ways, but in general adding a new instance or feature will increase the time needed to generate a classificatory model or clustering solution, sometimes exponentially so. For this reason, it can be very helpful to limit a priori the number of features in the matrix before presenting it to the machine learner, and PhiloMine provides the capability to filter out features based on a number of criteria. For each featureset, the user may limit the features to use by the number of instances in which the feature occurs, eliminating common or uncommon features. Additionally, include lists and/or exclude lists may be submitted, and only features on the include list and no features on the exclude list are retained. Finally, features may be filtered by their value on a per-instance basis, so that all features that occur more or less times than the user desires may be removed from a given instance, while remaining present in other instances.

Algorithm and Implementation Selection

PhiloMine can wrap data for, and parse results from, a variety of machine learning implementations. Native Perl functions currently include Ken William's naive Bayesian and decision tree classifiers, a vector space implementation and a differential relative rate statistics generator. The WEKA toolkit provides numerous implementations and currently PhiloMine works with the information gain, naive Bayes, SMO support vector machine, multilayer perceptron, and J48 decision tree WEKA components. PhiloMine also can talk to the compiled SVMLight support vector machine and CLUTO clustering engine. Relevant parameters for each function may also be set on the PhiloMine form.

When the user selects an implementation, the feature vectors for each instance are converted from PhiloMine's internal representation into the format expected by that package, generally a sparse vector format, such as the sparse ARFF format used by WEKA. The mining run is initiated either by forking a command to the system shell or by the appropriate Perl method call. Results of the run are displayed in the browser, typically including a list of text objects instances with classification results from the model and a list of features used by the classifier. Each instance is hyperlinked to the PhiloLogic display for that text object, so that the user can easily view that document, div or paragraph. Similarly, the user can push a query to PhiloLogic to search for any word used as a feature, either in the entire corpus or in any of the classed sub-corpora. If results show, for instance, that a support vector machine has heavily weighted the word "power" as indicative of a certain class of documents, the experimenter can quickly get a report of all occurrences of "power" in that class of documents, any other class or all classes.

This ability to easily move between the text mining results and the context of the source documents is meant to mitigate some of the alienating effects of text mining, where documents become anonymous instances and words are reduced to serialized features. For industrial applications of text mining, the accuracy of a certain classifier may be the only criterion for success, but for the more introspective needs of the digital humanist, the mining results must be examined and interpreted to further the understanding of the original texts. PhiloMine allows researchers to frame experiments in familiar terms by selecting corpora with standard bibliographic criteria, and then relate the results of the experiment back to the source texts in an integrated environment. This allows for rapid experimental design, execution, refinement and interpretation, while retaining the close association with the text that is the hallmark of humanistic study.

The Music Information Retrieval Evaluation eXchange (MIREX): Community-Led Formal Evaluations

J. Stephen Downie
jdownie@uiuc.edu
University of Illinois, USA

Andreas F. Ehmann
aehmann@uiuc.edu
University of Illinois, USA

Jin Ha Lee
jinlee l @uiuc.edu
University of Illinois, USA

Introduction

This paper provides a general overview of the infrastructure, challenges, evaluation results, and future goals of the Music Information Retrieval Evaluation eXchange (MIREX). MIREX [1] represents a community-based formal evaluation framework for the evaluation of algorithms and techniques related to music information retrieval (MIR), music digital libraries (MDL) and computational musicology (CM). MIREX is coordinated and managed by the International Music Information Retrieval Systems Evaluation Laboratory (IMIRSEL) at the University of Illinois at Urbana-Champaign. To date, since its inception in 2005, three annual MIREX evaluations have been performed covering a wide variety of MIR/MDL/CM tasks. The task definitions and evaluation methods for each annual MIREX are largely determined by community discussion through various communication channels with dedicated Wikis [4] playing a special role. Section 2 presents and defines the tasks associated with each of the past three MIREX evaluations.

In many respects, MIREX shares similarities to the Text Retrieval Conference (TREC) [5] in its overall approach to handling the evaluation of algorithms designed by the research community. Both MIREX and TREC are predicated upon the standardization of data collections; the standardization of the tasks and queries to be performed on the collections; and the standardization of evaluation methods used on the results generated by the tasks/queries [1]. However, associated with MIREX, there exist a unique set of challenges that cause MIREX to deviate from many of the methodologies associated with TREC. Section 3 covers some of these challenges, and the resultant solutions that comprise the overall framework and methodology of how MIREX evaluations are executed. Since MIREX is an ever-evolving entity, Section 4 will present key advances made between MIREX 2005 and MIREX 2006, as well as future goals for MIREX 2007 and beyond.

MIREX 2005, 2006, and 2007 tasks

The tasks associated with MIREX 2005, 2006 and 2007 are shown in Table 1.

Table 1. Task lists for MIREX 2005, 2006, and 2007 (with number of runs evaluated for each)

TASK	2005	2006	2007
Audio Artist Identification	7		7
Audio Beat Tracking		5	
Audio Classical Composer Identification			7
Audio Cover Song Identification		8	8
Audio Drum Detection	8		
Audio Genre Classification	15		7
Audio Key Finding	7		
Audio Melody Extraction	10	10 (2 subtasks)	
Audio Mood Classification			9
Audio Music Similarity and Retrieval		6	12
Audio Onset Detection	9	13	17
Audio Tempo Extraction	13	7	
Multiple F0 Estimation			16
Multiple F0 Note Detection			11
Query-by-Singing/Humming		23 (2 subtasks)	20 (2 subtasks)
Score Following	2		
Symbolic Genre Classification	5		
Symbolic Key Finding	5		
Symbolic Melodic Similarity	7	18 (3 subtasks)	3

A more detailed description of the tasks, as well as the formal evaluation results can be found on each year's associated Wiki pages [4]. The tasks cover a wide range of techniques associated with MIR/MDL/CM research, and also vary in scope. Tasks such as "Audio Onset Detection" (i.e., marking the exact time locations of all musical events in a piece of audio) and "Audio Melody Extraction" (i.e., tracking the pitch/fundamental frequency of the predominant melody in a piece of music) can be considered low-level tasks, in that they are primarily concerned with extracting musical descriptors from a single piece of musical audio. The motivation for evaluating such low-level tasks is that higher level MIR/MDL/CM systems such as "Audio Music Similarity and Retrieval" (i.e., retrieving similar pieces of music in a collection to a specified query song) or "Audio Cover Song Identification" (i.e., finding all variations of a given musical piece in a collection) can be built using many of the techniques involved in the low-level audio description tasks.

Table 2 provides a summary of participation in the past three MIREX evaluations. To date, 300 algorithm runs have been performed and evaluated.

Table 2. Summary data for MIREX 2005, 2006, and 2007

COUNTS	2005	2006	2007
Number of Task (and Subtask) "Sets"	10	13	12
Number of Teams	41	46	40
Number of Individuals	82	50	73
Number of Countries	19	14	15
Number of Runs	86	92	122

Challenges and Methodology

Although largely inspired by TREC, MIREX differs significantly from TREC in that the datasets for each task are not freely distributed to the participants. One primary reason for the lack of freely available datasets is the current state of copyright enforcement of musical intellectual property preventing the free distribution of many of the collections used in the MIREX evaluations. In addition, MIREX relies heavily on "donated" data and ground-truth. For tasks which require extremely labor-intensive hand-annotation to generate ground-truth — most notably low-level description tasks such as "Audio Onset Detection" — there is an overall reluctance of contributors to make their data and annotations freely available. As a result, it is nearly impossible to generate a representative dataset that encompasses all possible varieties, instrumentations, etc. of music. As such, there exists the potential of "tuning" or "overfitting" to a specific dataset at the expense of generalizablity of the algorithm to all varieties and types of music.

Due to the inability to freely distribute data, MIREX has adopted a model whereby all the evaluation data are housed in one central location (at IMIRSEL). Participants in MIREX then submit their algorithms to IMIRSEL to be run against the data collections. This model poses a unique set of challenges for the IMIRSEL team in managing and executing each annual MIREX. Firstly, data must be gathered and managed from various sources. For some tasks, differing formats for both the data and ground truth exist, as well as the potential for corrupted or incorrectly annotated ground-truth necessitating testing of the integrity of the data itself. The music collections used for MIREX tasks have already surpassed one terabyte and are continuously growing. In addition, many algorithms generate a large amount of intermediate data in their execution which must also be managed. In some cases, the intermediate data are larger in size than the actual music they describe and represent.

Moreover, IMIRSEL is responsible for supporting a wide variety of programming languages (e.g., MATLAB, Java, C/C++, PERL, Python, etc.) across different platforms (e.g., Windows, *NIX, MacOS, etc.). Despite guidelines dictating file input/output formats, coding conventions, linking methods, error handling schemes, etc., the largest amount of effort expended by IMIRSEL is in compiling, debugging, and verifying the output format and validity of submitted algorithms. Collectively, submissions to MIREX represent hundreds of hours of CPU computation time and person-hours in managing, setting up, and performing their execution.

Advances and Future Goals

One of the most significant advances made after MIREX 2005 was the incorporation of musical similarity evaluation tasks contingent upon subjective, human evaluation (MIREX 2006 and 2007). The addition of human evaluations of music similarity systems was born out of a community desire to reflect real-world applications and needs, and culminated in the "Audio Music Similarity and Retrieval" and "Symbolic Melodic Similarity" tasks. Both similarity tasks involve retrieving the top-N relevant or "similar" musical pieces in a collection using a specific musical piece as a query. A web interface with embedded audio players called the Evalutron 6000 was designed to allow evaluators to judge the similarity of a query "seed" with a retrieved "candidate" on both a broad scale (i.e., Not Similar, Somewhat Similar, Very Similar) and a fine, continuous, 10-point scale [3].

Another significant advance made manifest in MIREX 2006, and then repeated for MIREX 2007, was the application of formal statistical significance testing of returned results. These tests were applied in order to test whether performance differences between systems were truly significant. Because of its non-parametric nature, Friedman's ANOVA test was used on a variety of tasks to compare system performances. In general, these tests have shown that there are clusters of top performing techniques but these top-ranked techniques are not performing significantly better than their other top-ranked peers.

For future MIREX evaluations, IMIRSEL is presently developing a web service system that intends to resolve some of the key challenges associated with the execution of submitted algorithms by placing many of the responsibilities in the participant's hands. The web service, called MIREX DIY, represents a "black box" architecture, whereby a participant submits their algorithm/code through the web service, remotely begins its execution, and receives real-time feedback regarding the execution state of their algorithm. Execution failures can be monitored by the participant, and fixed if necessary. Upon successful execution and completion of the algorithm, performance results are returned to the participant. Eventually, such a system would allow submission and evaluation of algorithms year-round. Feel free to explore the MIREX DIY demo at http://cluster3.lis.uiuc.edu:8080/mirexdiydemo.

Acknowledgments

MIREX has received considerable financial support from both the Andrew W. Mellon Foundation and the National Science Foundation (NSF) under grant numbers NSF IIS-0340597 and NSF IIS- 0327371.

References

[1] Downie, J. S. 2006. The Music Information Retrieval Evaluation eXchange (MIREX). *D-Lib Magazine, 12,* 12 (December 2006: http://www.dlib.org/dlib/december06/downie/12downie.html.

[2] Downie, J. S., West, K., Ehmann, A., and Vincent, E. 2005. The 2005 Music Information Retrieval Evaluation eXchange (MIREX 2005): Preliminary overview. In *Proceedings of the 6th International Conference on Music Information Retireval (ISMIR 2005)*, Queen Mary, UK, 2005, pp. 320-323.

[3] Gruzd, A. A., Downie, J. S., Jones, M. C., and Lee, J. H. 2007. Evalutron 6000: Collecting music relevance judgments. *ACM IEEE Joint Conference on Digital Libraries 2007*, p. 507.

[4] MIREX Wiki: http://music-ir.org/mirexwiki/.

[5] TREC: http://trec.nist.gov/.

Online Collaborative Research with REKn and PReE

Michael Elkink
melkink@uvic.ca
University of Victoria, Canada

Ray Siemens
siemens@uvic.ca
University of Victoria, Canada

Karin Armstrong
karindar@uvic.ca
University of Victoria, Canada

The advent of large-scale primary resources in the humanities such as EEBO and EEBO-TCP, and similarly large-scale availability of the full-texts secondary materials through electronic publication services and amalgamators, suggests new ways in which the scholar and student are able to interact with the materials that comprise the focus of their professional engagement. The Renaissance English Knowledgebase (REKn) explores one such prospect. REKn attempts to capture and represent essential materials contributing to an understanding of those aspects of early modern life which are of interest to the literary scholar - via a combination of digital representations of literary and artistic works of the Renaissance plus those of our own time reflecting our understanding of earlier works. REKn contains some 13,000 primary sources at present, plus secondary materials such as published scholarly articles and books chapters reflecting our understanding of these earlier works (some 100,000). These materials are accessed through a Professional Reading Environment (PReE), supported by a database system that facilitates their navigation and dynamic interaction, also providing access to inquiry-oriented analytical tools beyond simple search functions. The effect is that of providing an expert reading environment for those in our field, one that encourages close, comprehensive reading at the same time as it provides, conveniently, the building blocks of broad-based research inquiry. We are currently moving beyond the stage of proof-of-concept with these projects.

Our current research aim with these projects is to foster social networking functionality in our professional reading environment. For DH2008 we propose a poster/demo that details the current status of both the knowledgebase (REKn) and the reading environment (PREe) in relation to social networking, the direction each will take in the future, and a demonstration of the functional technologies we employ and our current implementation.

Rather than leveraging the power of an individual computer to perform complex computation on a personalized data set - which is the way most academics appear to work (see, for example, Siemens, et al., 2004), and is an approach exemplified

in our community by Bradley (2007) and the TAPoR research team (TAPoR 2007; Rockwell 2006) - our work complements that approach by attempting to harness the power of the social connectivity provided by current Web 2.0 practices to connect researchers and experts, authors and reviewers, computers and humans, all with the express goal of bringing a higher level of discourse, analysis and structure to the documents brought together in the REKn corpus (for one influential adaptation, slightly outside our domain, see Ellison 2007). We are considering more than just a simple full-text search as a starting point for research discovery, even sophisticated forms of search (as per Schreibman, et al.); rather, we are envisioning ways existing social technologies can be used in concert with search processes to facilitate the process of professional reading (as per Warwick, et al., 2005). A further goal of our current work is to integrate more readily a generalized subset of analytical tools, derived from the TAPoR project and others, so that other large corpora similar to REKn can benefit from the computational and user-generated connections among material facilitated by our system; the beginnings of this system will be demonstrated as well (as per Elkink, et al., 2007).

References

Bradley, John. "Making a Contribution: Modularity, Integration and Collaboration Between Tools in Pliny". Paper presented at Digital Humanities 2007 June, 2007.

Ellison, Nicole B., Charles Steinfield and Cliff Lampe. "The Benefits of Facebook 'Friends:' Social Capital and College Students' use of Online Social Network Sites." *Journal of Computer-Mediated Communication* 12 (2007): 1143-1168.

Elkink, Michael, Ray Siemens, Karin Armstrong. "Building One To Thrown Away, Toward The One We'll Keep: Next Steps for the Renaissance English Knowledgebase (REKn) and the Professional Reading Environment (PReE)." Presented at the Chicago Colloquium on Digital Humanities and Computer Science. October 2007.

Rockwell, Geoffrey. "TAPoR: Building a Portal for Text Analysis." pp. 285-300 in Siemens and Moorman.

Schreibman, S., Sueguen, G., & Roper, J. "Cross-collection searching: A pandora's box or the holy grail." *Literary and Linguistic Computing*. In press.

Siemens, Ray, and David Moorman, eds. *Mind Technologies: Humanities Computing and the Canadian Academic Community*. Calgary: U Calgary P, 2006.

Siemens, Ray, Elaine Toms, Stéfan Sinclair, Geoffrey Rockwell, and Lynne Siemens. "The Humanities Scholar in the Twenty-first Century: How Research is Done and What Support is Needed." ALLC/ACH 2004 Conference Abstracts. Göteborg: Göteborg U, 2004.

TAPoR: Text Analysis Portal for Research. 2007. http://portal.tapor.ca/.

9. Warwick, C., Blandford, A., Buchanan, G., J. Rimmer, J. (2005) "User Centred Interactive Search in the Humanities." *Proceedings of the Joint Conference on Digital Libraries*. Denver, Colorado, June 7-11, 2005. ACM Publications. pp. 279-81.

A Computerized Corpus of Karelian Dialects: Design and Implementation

Dmitri Evmenov
dmitri.evmenov@gmail.com
St. Petersburg State University, Russian Federation

In my presentation, I intend to cover in detail the main project I am working on at the moment, namely designing and implementing the computerized corpus of Karelian language dialects.

During the decades of scientific study of Karelian language, initiated in mid-19th century by Finnish scholars and largely expanded by Russian/Soviet linguists later on, a large volume of material, most remarkably dialectal speech samples, was amassed. Those data are however to a large extent essentially inaccessible to research due to lack of representative and accessible solutions allowing for representation of that rich material. Therefore in my research project I aim at developing and building an annotated computerized corpus of Karelian dialects as well as developing recommendations regarding the corpus' further expansion.

During the first stage of implementation a moderately sized "pilot corpus" is to be built so that different strategies and tools for annotation could be developed and tested with its help. The pilot corpus is to be expanded later on by feeding in other available source materials.

The pilot corpus shall contain dialectal speech samples belonging to one dialectal group, the Olonets Karelian (Livvi), mostly because there's more extant dialectal speech samples recorded for Olonets Karelian than for other groups, namely Karelian Proper and Ludik dialects. Also, albeit certainly endangered due to numerous reasons, Olonets Karelian yet shows less signs of attrition and interference with neighbouring languages (Veps, Finnish, and Russian) than the above mentioned two different dialectal groups.

The representativeness of the pilot corpus is to be achieved, above all, by proportional inclusion of dialectal speech samples from all language varieties found in the areal where Karelian language is spoken. In order to better account for dialectal variation in case of Karelian language, it appears reasonable to include into corpus dialectal material from each administrative division unit (volost), the volume being 100 000 symbols per one such unit).

It is intended to employ demographic criteria alongside with geographic ones during material selection. In grouping the informants in terms of their age, it appears reasonable to follow the division into "elder" (born in 1910-1920s), "middle" (born in 1930-1940s) and "younger" (born in 1950-1960s) groups. As for gender representativeness, equal representation of male and female informant's speech in the corpus appears impossible, at least for elder groups. The informant's education level, place of studies and career biography are all to be taken into consideration as well.

It is necessary also to include the information that the informant provides about her linguistic competence (which language she considers her native, how many language and to which extent she knows and can use) and performance (the domains where she uses Karelian language, in terms of Joshua Fishman's domain theory).

In the beginning of pilot corpus' implementation it is intended to use the already published samples of Karelian dialectal speech, while at later stages other published and specially transcribed materials are to be added, mostly those now stored in the archives of the Institute of Language, Literature and History of Karelian Research Center of Russian Academy of Sciences (Petrozavodsk, Russia).

Every block of dialectal material included into the corpus is to be accompanied by metadata, including the following:

- informant data (gender, age, place of birth, duration of stay in the locality where the record was made, duration and circumstances of stay away from Karelian language areal, native language according to informant's own judgment, informant's judgment regarding her mastery of Karelian, Russian and other languages, language choice habits for various usage domains)

- data on the situation of speech sample recording (researcher and informant dialogue, recording of the informant's spontaneous monological speech, recording of a dialogue where two or more informants are participating); it appears reasonable to develop a taxonomy of situations in order to encode it later on in corpus

- the theme of the conversation recorded and transcribed; in this case it also appears reasonable to develop and enhance an appropriate taxonomy to be employed for data encoding at a later stage of corpus expansion.

The detailed way of representing the data in the corpus ("the orthography") is to follow the universally accepted Standard Fenno-Ugric Transcription (Suomalais-Ugrilainen Tarkekirjoitus), although there are certain challenges, mainly stemming from not so easily encodable diacritic signs combinations, that require their own solutions to be developed; implementation details will surely depend on a chosen technology and/or software platform to be chosen for use. It should be mentioned though, that the borders of prosodic phrases normally marked in transcribed speech samples will be saved in the corpus as well and used later on for purposes of syntactic annotation.

For morphological annotation, a united intra-dialectal morphological tag set is to be developed; pilot corpus will be annotated manually, while later stages might be annotated with the help of existing parsing and tagging software, inasmuch as it is applicable for our purposes.

Other design and implementation details, now being actively worked upon, are also to be included into the presentation.

Digitally Deducing Information about the Early American History with Semantic Web Techniques

Ismail Fahmi
i.fahmi@rug.nl
University of Groningen, The Netherlands

Peter Scholing
pscholing@gmail.com
University of Groningen, The Netherlands

Junte Zhang
j.zhang@uva.nl
University of Amsterdam, The Netherlands

Abstract: We describe the Semantic Web for History (SWHi) System. The metadata about the Early American History is used. We apply Information Retrieval and Semantic Web technologies. Furthermore, we use Information Visualization techniques. Our resulting information system improves information access to our library's finding aids.

Introduction

The Early American Imprints Series are a microfiche collection of all known existing books, pamphlets and periodical publications printed in the United States from 1639-1800, and gives insights in many aspects of life in 17th and 18th century America, and are based on Charles Evans' American Bibliography. Its metadata consist of 36,305 records, which are elaborately described (title, author, publication date, etc) with numerous values, and have been compiled by librarians in the format MARC21, which we encoded in XML.

The Semantic Web for History (SWHi) project at the Digital Library department of the University Library Groningen has worked with the MARC21 metadata of this dataset. Our project aims to integrate, combine, and deduce information from this dataset to assist general users or historians in exploring American history by using new technology offered by the Semantic Web. Concretely, we developed a semantic search application for historical data, especially from a Digital Library point of view.

Semantic Web technologies seem ideally suited to improve and widen the services digital libraries offer to their users. Digital libraries rely heavily on information retrieval technology. The Semantic Web might be used to introduce meaningful and explicit relations between documents, based on their content, thereby allowing services that introduce forms of semantic browsing supplementing, or possibly replacing keyword-based searches. This is also a theme we address in SWHi project.

Semantic Web and Information Retrieval

Digital libraries increase their amount of metadata each day, but much of these metadata is not used as a finding aid. We have used these metadata to create a Semantic Web compatible "historical" ontology. Ontologies are the backbone of the Semantic Web, and also play a pivotal role in the SWHi application. One of the ideas of the Semantic Web is that existing resources can be reused. This was also one of the key ideas of our project.

Our ontology is constructed by using the existing ontology PROTON as its skeleton, and is enriched with other schemas. We reuse existing topical taxonomies created by librarians and other experts. We also extracted topic hierarchies from the metadata. To describe objects with more semantics and create relationships between objects, we also enriched the ontology with the vocabularies Dublin Core and Friend of a Friend (FOAF). For example, we link instances with time and space, but also topics and persons. This is useful for discovering historical social networks, exploring gazetteers, and clustering instances together by topic for faceted search, etc. We aligned the library metadata with schemas and vocabularies. Information is extracted from the metadata to populate the ontology. All of this is eventually processed and encoded in XML with RDF/OWL. In addition to the PROTON's basic ontology modules, we get 152 new classes from this mapping. And in total, we get 112,300 ontology instances from the metadata.

We also combine Information Retrieval technology with Semantic Web techniques. We use the open-source search engine SOLR to index ontology instances, parse user input queries, and eventually retrieve matching ontology instances from the search results. It supports faceted search and has been been designed for easy deployment.

During the indexing stage of the data, we apply inference from the ontology as a propagation for the importance of the different metadata records and fields. Using the ontology, instances with higher relevance can have higher position in the order. For example, a person who is known by many people and created many documents would get a higher score. We use Sesame for storage and retrieve using the RDF query language SPARQL.

User Access

On top of our storage and retrieval components, we developed novel techniques for our semantic search application to visualize information and offer users to browse for that information in an interactive manner. We let users search for information semantically, which means that information is linked together with certain relations. Results are clustered together based on such relations which allows faceted search by categories, see figure 1. We provide context to relevant nuggets of information by enabling users to traverse related RDF graphs. We visualize interconnected results using network graphs with the TouchGraph tool.

Illustration 1: Result list for query "saratoga".

We picked up this idea from the 'berrypicking' model: a user searches, picks a berry (a result), stores it in his basket, view the relations between the berries in the basket, and the search iteration continues. The purpose is to find new information from the collected berries (results).

As we are dealing with historical data, chronological timeline views are also presented using SIMILE's Timeline, which lets users browse for information by time. Besides time, we also offer users to search by location using Google Maps. Geographical entities, mostly locations in the US, are aligned with Google Maps.

In summary, we have four modes of visualization which gives users multiple views of the results: plain list, faceted search, timeline, and map. We have found that translating such an interface to an on-line environment offers interesting new ways to allow for pattern discovery and serendipitous information seeking. Adding information visualization tools like interactive and descriptive maps and time-lines to the electronic finding aid's interface could further improve its potential to augment cognition, and hence improve information access.

Conclusions

We presented the Semantic Web for History (SWHi) system, which deals with historical data in the form of library finding aids. We employed Semantic Web and Information Retrieval technologies to obtain the goal of improving user access to historical material.

Bibliography

Grigoris Antoniou, Frank van Harmelen. *A Semantic Web Primer.* The MIT Press, 2004.

Berners-Lee, T., Hendler, J., and Lassila, O. The semantic web. A new form of web content that is meaningful to computers will unleash a revolution of new possibilities. *The Scientific American,* 2001.

Ismail Fahmi, Junte Zhang, Henk Ellermann, and Gosse Bouma. "SWHi System Description: A Case Study in Information Retrieval, Inference, and Visualization in the Semantic Web." *The Semantic Web: Research and Applications*, volume 4519 of Lecture Notes in Computer Science, pages 769-778. Springer, 2007.

Dieter Fensel, Wolfgang Wahlster, Henry Lieberman, James Hendler. *Spinning the Semantic Web: Bringing the World Wide Web to Its Full Potential.* MIT Press, 2002.

Lucene SOLR. http://lucene.apache.org/solr/

Semantic Web for History (SWHi). http://evans.ub.rug.nl/swhi

SIMILE Timeline. http://simile.mit.edu/timeline/

Junte Zhang, Ismail Fahmi, Henk Ellermann, and Gosse Bouma. "Mapping Metadata for SWHi: Aligning Schemas with Library Metadata for a Historical Ontology." *Web Information Systems Engineering – WISE 2007 Workshops*, volume 4832 of Lecture Notes in Computer Science, pages 102-114. Springer, 2007.

TauRo - A search and advanced management system for XML documents

Alida Isolani
isolani@signum.sns.it
Scuola Normale Superiore -Signum, Italy

Paolo Ferragina
ferragina@di.unipi.it
Scuola Normale Superiore -Signum, Italy

Dianella Lombardini
dianella@signum.sns.it
Scuola Normale Superiore -Signum, Italy

Tommaso Schiavinotto
Scuola Normale Superiore -Signum, Italy

With the advent of Web 2.0 we have seen a radical change in the use of Internet, which is no longer seen as a tool from which to draw information produced by others, but also as a means to collaborate and to share ideas and contents (examples are Wikipedia, YouTube, Flickr, MySpace, LinkedIn, etc.). It is with this in mind that **TauRo**[1] was designed – a user-friendly tool with which the user can *create, manage, share, and search digital collections of XML documents* via Web. TauRo is indeed a collaborative system through which the user who has access to Internet and to a browser can publish and share their own XML documents in a simple and efficient manner, thus creating personal and/or shared thematic collections. Furthermore, TauRo provides extremely advanced search mechanisms, thanks to the use of a search engine for XML documents, **TauRo-core** – an *open source* software that offers implemented search and analysis functions to meet the current need of the humanity texts encoding.

TauRo-core: the search engine

The success of XML as an online data exchange format on the one hand, and the sucess of the search engines and Google on the other, offer a stimulating technological opportunity to use great quantities of data on standard PCs or on other portable devices such as smart-phones and PDA. The TauRo-core search engine is an innovative, modular, and sophisticated software tool which offers the user compressed storing and efficient analysis/research of arbitrary patterns in large collections of XML documents that are available both on a single PC and distributed among several PCs which are dislocated in various areas of the network. The flexibility of TauRo-core's modular architecture along with the use of advanced compression techniques for the storing of documents and for the memorization of indexes, makes it suitable to be used in the various scenarios illustrated in Figure 1.

Figure 1 – Application scenario for TauRo-core: client-server (a), distributed (b), P2P (c).

In particular, the use of the system in centralized modality – that is, in which both the documents and the engine are located in the same server – is already operative and suitable tested via implementation of the system on the Web (TauRo). We are currently working on the structure of *Web services* – matching the distributed mode – in order to supply *collection creation, submission of documents, search, and consultation* services.

Experiments have also been run to make it possible to consult collections via PDA or smart-phone: via a specific interface the user can make a query and consult the documents efficiently by using the **Nokia Tablet 770**.

This way, we can also evaluate the behavior of the software with reduced computational and storing resources.

Compared to the search engines available on the international scene, TauRo-core offers added search and analysis functions to meet the current needs of the humanity texts encoding. Indeed, these may be marked in such a way as to make their analysis difficult on behalf of the standard search engines designed for non-structured documents (i.e. Web search engines), or for highly-structured documents (i.e. database), or for semi-structured documents (i.e. search engines for XML), but in which these are no assumptions on the semantics of the mark-up itself.

TauRo-core, instead, allows the user to index XML texts for which the opportune tag categories have been defined. These tags are denominated **smart-tag**[2], and they are associated with specific management/search guidelines. In order to appreciate the flexibility of the smart-tag concept, we have illustrated the classification here below:

• **jump-tag**: the tags of this group indicate a temporary change in context – as in the case of a tag that indicates a note – and this way the tag content is distinct from the actual text and the search takes place while distinguishing the two semantic planes.

• **soft-tag**: these tags involve a change of context, if the starting or ending element of the tag is present within a character string which is not separated by a space, the string forms a single word.

• **split-tag**: the tags to which a meaning similar to the word separator is assigned, fall within this category. Therefore, the context does not change and the words are in effect considered as separate.

Furthermore, TauRo-core offers its own query language, called **TRQL**, which is powerful enough to allow the user to carry out complex text analysis that take into account the above classification and the relationship between content and structure of the documents. TRQL operates on document collections, manages the smart-tag and implements the main full-text search functions requested in the specifics of the W3C[3] for XQuery.

This flexibility allows TauRo-core to also be used in contexts that are different from the strictly literary one; for example, the collection of documents of the public administration, biological archives, manuals, legislative collections, news, etc. The literary context however remains the most complex and thus constitutes, due to its peculiar lack of uniformity, a more stimulating and meaningful trial.

TauRo: the system on the Web

TauRo is a collaborative system that allows any Web user, after free registration, to create and share collections of XML documents, and to exploit the potential of TauRo-core to run full-text searches for regular expressions, by similarity, and searches within the XML document structure. These searches can be run on a single collection at a time or on various collections simultaneously, independently from their nature. Aided screenshots, we show here below some characteristics of the system.

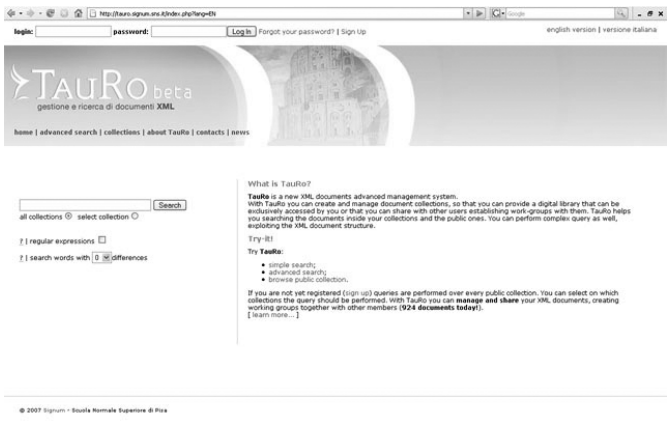


Figure 2 – TauRo home page

The collections

Each registered user can upload on TauRo their own collection of XML documents. Once uploaded, the collection will be indexed by TauRo-core and made available to the next search operations. The user may, at any time, modify their collection by adding or deleting documents, by moving documents from one collection to the other, and share documents between various collections, or modify the status of a collection that can be:

• **private:** accessible and searchable only by the owner;

• **public:** searchable by all the users after registering and modifiable only by the owner;

• **invite-only:** this collections can be subscribed only after invitation by one of the collection administrators. However, the user has the possibility to ask for the invitation.

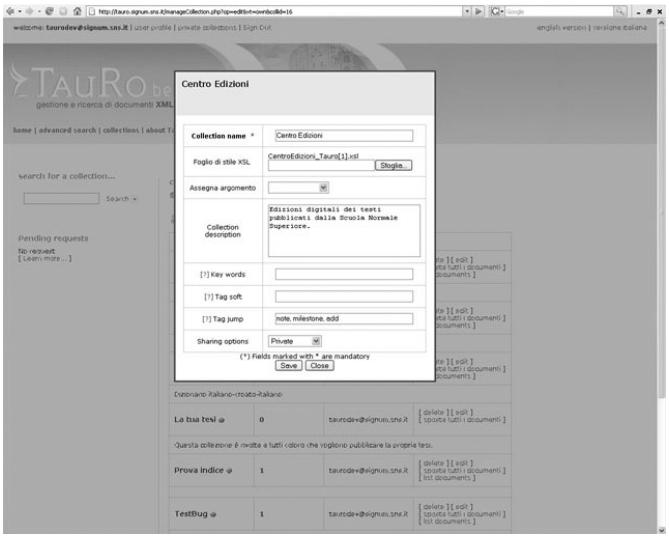


Figure 3 – Collection edit form.

During the uploading or modification of a collection, the user can select some parameter settings, such as the smart-tag and page interruption tag specifics, for the purpose of exploiting to the fulltest the search engine's potential. A further personalization option offered by TauRo consists in associating each collection with its own XSL stylesheet[4] aimed at visualizing in a pleasant way the results of the searches run on them.

The documents

The system provides the user with a group of functions that can upload, classify, and remove XML documents from the user's collections. During the upload, the system will try to run an automatic acknowledgment of the DTD and of the entity files used in the XML document by comparing the public name with those previously saved. If the acknowledgment fails, the user is give the option of entering the information. Every document can be freely downloaded by anyone or one of the *Creative Commons*[5] licenses that safeguard the owner from improper use can be selected.

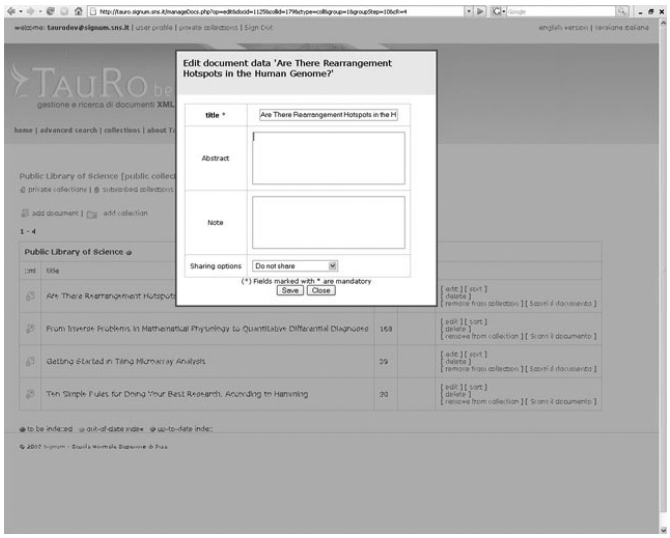


Figure 4 – In the foreground is the document edit form, and in the background the list of documents.

Search

TauRo offers two different search modes. The first is designed to search for words within some documents *(basic search)*, the second allows the user to also construct structural type of queries, namely on elements – tags and attributes – of the mark-up via a graphic interface *(advanced search)*. In both cases the queries are translated into a syntax that can be interpreted by TauRo-core and sent to it. The search result is the list of documents of the collection which verify the query, set alongside the distribution of the results within the documents themselves. By selecting a document, the user accesses a list of contextualized occurrences, namely those entered in a text fragment which contains them, with the possibility of directly accessing the text as a whole.

In both cases the search can be exact, by prefix, suffix, standard expression or by difference. The user can specify several words, and, in this case, they will appear next to the document. A basic search can also be run on several collections simultaneously.

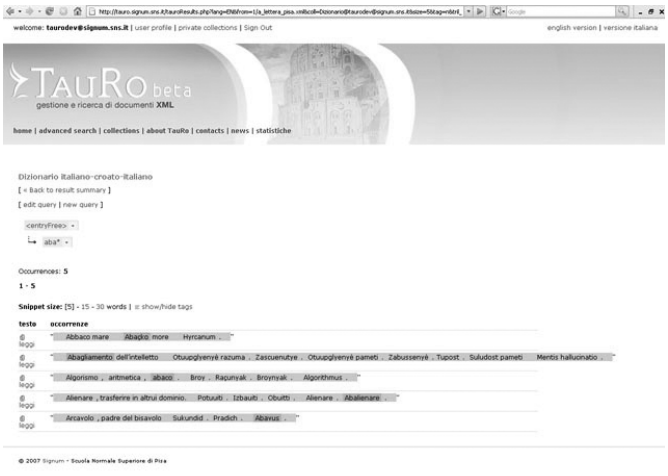


Figure 5 – Search results

The search results still consist in a list of collection documents that verify the query. By selecting a document, the user accesses the list of occurrences (Figure 5) that represents the starting point for accessing the text.

The project has been designed to allow any user to try and exploit the potential of the search engine by using their PC, and without having to install complex software. Thanks to the potentials of TauRo, indeed any user in any part of the world may create and manage via Web their own XML text collection.

Notes

1 http://tauro.signum.sns.it

2 L. Lini, D. Lombardini, M. Paoli, D. Colazzo, C. Sartiani, XTReSy: A Text Retrieval System for XML Documents. In *Augmenting Comprehension: Digital Tools for the History of Ideas*, ed by H. Short, G. Pancaldi, D. Buzzetti, luglio 2001. Office for Humanities Communication Publications 2001.

3 http://www.w3.org/TR/xquery-full-text-requirements/ Specifics of the language and interrogation requisits, XQuery.

4 eXtensible Stylesheet Language (XSL), a language for expressing stylesheets.

5 http://www.creativecommons.it/

The Orationes Project

Anthony Johnson
anthony.johnson@oulu.fi
University of Oulu, Finland

Lisa Lena Opas-Hänninen
lisa.lena.opas-hanninen@oulu.fi
University of Oulu, Finland

Jyri Vaahtera
jyri.vaahtera@utu.fi
Turku University, Finland

The *Orationes* Project is an interdisciplinary initiative intended to bring an important unpublished manuscript into the scholarly arena. The manuscript, preserved as Lit. MS E41 in the archive of Canterbury Cathedral, was assembled by George Lovejoy, the 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. Some of these celebrate the Restoration of Charles II to power or reconstruct the famous 'Oak-apple' adventure by which he escaped his enemies, during the Civil War, by hiding in an Oak tree. Some re-enact the Gunpowder Plot, which nearly destroyed Charles II's grandfather, James I. And others engage with a wide range of topical issues, from discussions of religious toleration, or the teaching of classics and grammar in the Restoration, to a dramatized dialogue between Ben Jonson and Richard Lovelace and an alchemical allegory on the politics of state.

In the shorter term, the aim of the project has been to produce a pilot study. To this end we have begun by transcribing the main texts from the corpus and will produce a series of critical studies of them. But due to the extensive size of the manuscript, in the longer term, this forms part of a planned *Digital Orationes Project* shared by the English Department at Oulu University, the Department of Classical Cultures and Languages at Turku University, and the Department of Religious Studies within the University of Kent at Canterbury. With the aid of Graduate research, these institutions will collaborate in the creation of a digital archive which will make these exciting Early Modern texts available (in parallel translation) to a wider audience. The present poster represents our first collaborative endeavour within this digital enterprise.

Our aims are the following: a) to digitize this collection, providing images of the manuscript and transcriptions of the texts within it. We will also b) provide translations in Modern English of the full texts in the manuscript; many of the texts characteristically shift language from Latin (or Greek) in mid-flow and thus warrant a Modern English translation. We intend

c) to interlink the manuscript images with the transcriptions and translations. Finally, we intend d) to make the corpus freely available on line for scholarly purposes. The texts will be marked up in XML and we will follow the TEI Guidelines. We intend to make this digital archive searchable and hope to make use of tools for the analysis of multimodal data created at the Faculty of Engineering, University of Oulu. As part of the pilot study mentioned above, we will begin by digitising those texts that have been transcribed. This will give us a good opportunity to produce a model by which to tackle the rest of the texts, since we can link the manuscript, the transcription, a translation and a critical study of the text. It will also give us an opportunity to explore what kinds of functionality we could or should include in the manipulation of the texts. We also intend to make use of the experiences of the Scots project in the use of the database model to handle the creation process of the digital text.

We therefore feel that DH2008 would offer us a good opportunity to discuss the project and its practices with others who have already carried out similar projects. We will also examine the best practices available online within related areas – as witnessed, for instance, by the Emily Dickinson, Lincoln and Perseus projects (http://www.emilydickinson.org/ ; http://quod.lib.umich.edu/l/lincoln/; http://www.perseus.tufts.edu/) – in order to finesse our search techniques and improve access routes to our manuscript materials.

Within a Finnish context the *Orationes* project aims to respond to Jussi Nuorteva (the director of the National Archive Service in Finland), who has criticised the Finnish Universities for their reticence in producing open access digital databases and archives for the use of other researchers. Hence, we plan to open up the *Digital Orationes Project* after 2009 on the same model as the *Diplomatarium Fennicum* project – http://193.184.161.234/DF/index.htm – (Nuorteva's own), which makes the Finnish medieval sources available for researchers in a digitized form. In this way our project will also be forward-looking: helping to position the activities of Oulu's English Department and Turku's Classics Department more firmly within the new domain of scholarship which has been opened up by digital archiving.

JGAAP3.0 -- Authorship Attribution for the Rest of Us

Patrick Juola
juola@mathcs.duq.edu
Duquesne University, USA

John Noecker
noeckerj@duq.edu
Duquesne University, USA

Mike Ryan
ryanm1299@duq.edu
Duquesne University, USA

Mengjia Zhao
zhaom@duq.edu
Duquesne University, USA

Authorship Attribution (Juola, in press) can be defined as the inference of the author or her characteristics by examining documents produced by that person. It is of course fundamental to the humanities; the more we know about a person's writings, the more we know about the person and vice versa. It is also a very difficult task. Recent advances in corpus linguistics have shown that it is possible to do this task automatically by computational statistics.

Unfortunately, the statistics necessary for performing this task can be onerous and mathematically formidable. For example, a commonly used analysis method, Principle Component Analysis (PCA), requires the calculation of "the eigenvectors of the covariance matrix with the largest eigenvalues," a phrase not easily distinguishable from Star Trek technobabble. In previous work (Juola, 2004; Juola et al., 2006) we have proposed a model and a software system to hide much of the details from the non-specialists, while specifically being modularly adaptable to incorporate new methods and technical improvements. This system uses a three-phase framework to canonicize documents, create an event set, and then apply inferential statistics to determine the most likely author. Because of the modular nature of the system, it is relatively easy to add new components.

We now report (and demonstrate) the recent improvements. Version 3.0 of the JGAAP (Java Graphical Authorship Attribution Program) system incorporates over fifty different methods with a GUI allowing easy user selection of the appropriate ones to use. Included are some of the more popular and/or well-performing methods such as Burrows' function word PCA (Burrows, 1989), Burrows' Delta (2003; Hoover 2004a, 2004b)), Juola's cross-entropy (2004), and Linear Discriminant Analysis (Baayen, 2002). The user is also capable of mixing and matching components to produce new methods of analysis; for example, applying PCA (following Burrows), but to an entirely different set of words, such as all the adjectives in a document as opposed to all the function words.

With the current user interface, each of these phases is independently selectable by the user via a set of tabbed radio buttons. The user first defines the document set of interest, then selects any necessary canonicization and pre-processing, such as case neutralization and/or stripping HTML markup from the documents. The user then selects a particular event set, such as characters, words, character or word N-grams, the K most common words/characters in the document, part of speech tags, or even simple word/sentence lengths. Finally, the user selects an analysis method such as PCA, LDA, histogram distance using a variety of metrics, or cross-entropy.

More importantly, the JGAAP framework can hide this complexity from the user; users can select "standard" analysis methods (such as "PCA on function words") from a set of menus, without needing to concern themselves with the operational details and parameters. Most importantly of all, the framework remains modular and easily modified; adding new modules, event models, and analytic methods can be done in minutes by Java programmers of only moderate skill. We will demonstrate this by adding new capacity on the fly.

Perhaps most importantly, we submit that the software has achieved a level of functionality and stability sufficient to make it useful to interested non-specialists. Like the Delta spreadsheet (Hoover, 2005), JGAAP provides general support for authorship attribution. It goes beyond the Delta spreadsheet in the variety of methods it provides. It has also been tested (using the University of Madison NMI Build-and-Test suite) and operates successfully on a very wide range of platforms. By incorporating many cooperational methods, it also encourages the use of multiple methods, a technique (often called "mixture of experts") that has been shown to be more accurate than reliance on any single technique (Juola, 2008).

Of course, the software is not complete and we hope to demonstrate some of its weaknesses as well. The user interface is not as clear or intuitive as we hope eventually to achieve, and we invite suggestions and comments for improvement. As the name suggested, the software is written in Java, and while Java programs are not as slow as is sometimes believed, the program is nevertheless not speed-optimized and can take a long time to perform its analysis. Analysis of large documents (novels or multiple novels) can exhaust the computer's memory. Finally, no authorship attribution program can be a complete survey of the proposed literature, and we invite suggestions about additional methods to incorporate.

Despite these weaknesses, we nevertheless feel that the new version of JGAAP is a useful and reliable tool, that the community at large can benefit from its use, and that the development of this tool can similarly benefit from community feedback.

References

Baayen, Harald et al. (2002). "An experiment in authorship attribution." *Proceedings of JADT 2002*.

Burrows, John F. (1989). "`An Ocean where each Kind...' : Statistical Analysis and Some Major Determinants of Literary Style." *Computers and the Humanities*, 23:309-21

Burrows, John F. (2002). "Delta : A Measure of Stylistic Difference and a Guide to Likely Authorship." *Literary and Linguistic Computing* 17:267-87

Hoover, David L. (2004a). "Testing Burrows's Delta." *Literary and Linguistic Computing*, 19:453-75.

Hoover, David L. (2004b). "Delta Prime?" *Literary and Linguistic Computing,* 19:477-95.

Hoover, David L. (2005) "The Delta Spreadsheet." *ACH/ALLC 2005 Conference Abstracts*. Victoria: University of Victoria Humanities Computing and Media Centre p. 85-86.

Juola, Patrick. (2004). "On Composership Attribution." *ALLC/ ACH 2004 Conference Abstracts*. Gothenburg: University of Gothenburg.

Juola, Patrick. (2008). "Authorship Attribution : What Mixture-of-Experts Says We Don't Yet Know." Presented at American Association of Corpus Linguistics 2008.

Juola, Patrick. (in press). *Authorship Attribution*. Delft:NOW Publishing.

Juola, Patrick, John Sofko, and Patrick Brennan. (2006). "A Prototype for Authorship Attribution Studies." *Literary and Linguistic Computing* 21:169-78

Translation Studies and XML: Biblical Translations in Byzantine Judaism, a Case Study

Julia G. Krivoruchko
jgk25@cam.ac.uk
Cambridge University, UK

Eleonora Litta Modignani Picozzi
eleonora.litta@kcl.ac.uk
King's College London, UK

Elena Pierazzo
elena.pierazzo@kcl.ac.uk
King's College London, UK

All definitions of translation describe this process as involving two or more texts running in parallel that are considered to be, in a sense, equivalent to each other. When producing a translation, a source text is divided into syntactic units and each of them is then translated. The translation can be either literal, i.e. it mirrors the structure of the original text very closely, or free, i.e. it ignores the original structure and translates freely. Because languages diverge greatly in their syntax, the structure of a language A can not be fully mapped on a language B, since the outcome in B may be incomprehensible. Besides, cultures differ greatly as to the degree of freedom/literalness tolerated in translation.

In dealing with translations compiled in antiquity and the Middle Ages, we are in a sense trying to discover how a specific culture understood the source text, the grammar of the source language and the usability of the translation product.

Even though a number of XML-based projects involving translated texts have been to date proposed to the attention of the community,[1] a model able to describe the precise relationship between source and target texts is still required.

Such issues have been dealt with at the Centre for Computing in the Humanities (King's College, London) in relation to a research project involving Biblical translations. The analysis process resulted in an encoding model that could solve problems specifically linked to this project. Yet the model has the potential to be generalized and adapted to work in other translation-based projects.

The Biblical translations are by far the most studied in the world. The text of the Hebrew Bible used in Hellenistic period was written down in a purely consonantal script without vowels, which left a large margin for differing interpretations. In addition to this, the Hebrew manuscript tradition was far from being homogenous. As a result, a number of translations of the Hebrew Bible into Greek emerged, some of them differing substantially from each other.

Until recently, the assumption was that Jews abandoned their use of Greek Biblical translations, since these were adopted by the Church. In particular, they were supposed to ignore the Septuagint, which was recognized as a canonical and authoritative text by Eastern Churches. However, the manuscripts found in Cairo Genizah have shaken this assumption. During the 20th century, new Biblical translations made by Jews into Greek during Ottoman and Modern period were discovered.[2]

The Greek Bible in Byzantine Judaism (GBBJ) Project[3] aims to gather textual evidence for the use of Greek Bible translations by Jews in the Middle Ages and to produce a corpus of such translations. Unfortunately, the majority of GBBJ evidence consists not of continuous texts, but of anonymous glossaries or single glosses mentioned by medieval commentators.[4] The functioning of continuous texts at our disposal is also unclear.

Further challenges arise from the peculiarities of the writing system used by the Byzantine translators. Since approximately the 7th-8th century AD, Jews stopped using the Greek alphabet and switched instead back to the Hebrew one. In order to unambiguously represent the Greek phonetics, the Hebrew alphabet was often supplied with vowel signs and special diacritics. Some manuscripts contain neither or use them inconsistently. In order to decode the writing an IPA reconstruction is therefore essential. However, Hebrew writing occasionally results in better reflecting the current medieval pronunciation of the Greek language. For what the linguistic structure is concerned, while in general Greek Jewish Biblical translations use the grammar and lexicon of the mainstream Greek, in some cases the translators invent lexical items and employ unusual forms and constructions, trying to calque the maximal number of grammatical features from one language into another. Few of the resulting forms are difficult or even impossible to understand without knowing the Hebrew source. To trace the features transferred and the consistency of transferring, the tagging of features is necessary. Therefore, lemmatization and POS-tagging both of the source and the target texts constitute an essential component for the research project.

The two main outcomes of the project will be a printed and a digital edition; the latter will allow users to query and browse the corpus displayed in parallel verses from the source and the target texts. The target will be readable in two alphabets: Hebrew and transliterated Greek.

In designing the encoding model, we have tried to follow TEI standards as much as possible, e.g. elements for the description of metadata, editorial features and transcription of primary fonts have been devised according to P5 guidelines since the beginning. Yet for what the textual architecture is concerned, TEI P5 does not include a model that would fit the project's needs, hence we have chosen to start working on the encoding model on the basis of a custom made DTD rather than a TEI compliant schema.

As a result, the tailored encoding model is simpler to apply for the encoders. However, all the elements have been mapped on TEI P5 for interchange and processing purposes.

The encoding model works on three different layers:

a) Structural layer. In any translation study, at least two texts need to be paired: the source and the target. The GBBJ project focuses on the target text, and therefore it constitutes our main document. This text was described as consisting of segments, within biblical verses, defined on palaeographical grounds. Since the Greek texts we are dealing with are written in Hebrew characters, they have to be transliterated into Greek characters and normalized within what we called the <lexicalTranscription>. The translational units are then compared to their presumed sources (the Masorethic Text, in our case) in <coupledPairs> and finally matched with the translations from other Biblical traditions (Septuagint and Hexaplaric versions), called <equivalents>. The latter are further connected to their respective apparatuses. There is a possibility to compare between the version of Hebrew lexeme as it appears in the manuscript and the Masoretic Text or its apparatus.

```
<GBBJ>
<metadata></metadata>
<text>
  <verse MT="Koh 2:19" LXX="Koh 2:19">
   
     <transcription language="Greek"
script="Hebrew">
        <seg>יִקְשׁוֹנִיי</seg>
        <IPA>jinoski</IPA>
     
```

to the research project, the approach and the conceptual model used here may be reused and adapted within the digital humanities community.

Notes

1 For example: The Emblem Project Utrecht (http://emblems.let.uu.nl/index.html, accessed 23/11/07), the English-Norwegian Parallel Corpus (http://www.hf.uio.no/ilos/forskning/forskningsprosjekter/enpc/, accessed 23/11/07).

2 For general discussion see Fernández Marcos (1998). Outline of the problems and suggestions for future research: RASHI 1040-1990.

3 A three year AHRC project based at University of Cambridge (Faculty of Divinity) and King's College, London (Centre for Computing in the Humanities); see project website at http://www.gbbj.org (accessed 7/3/2008).

4 For glossaries see De Lange (1996); Rueger (1959); Tchernetska, Olszowy-Schlanger, et al. (2007). For a continuous text see Hesseling, (1901).

References

Burnard, Lou and Bauman, Syd (2007), *TEI P5: Guidelines for Electronic Text Encoding and Interchange*, at http://www.tei-c.org/release/doc/tei-p5-doc/en/html/index.html (accessed 23/11/07).

De Lange, N. R. M. (1993). The Jews of Byzantium and the Greek Bible. Outline of the problems and suggestions for future research. *RASHI 1040-1990. Hommage à Ephraïm E. Urbach*. ed. G. Sed-Rajna. Paris, Éditions du Cerf: 203-10.

ID. (1996). *Greek Jewish Texts from the Cairo Genizah*, Tübingen: Mohr-Siebeck.

Fernández Marcos, Natalio. (1998). *Introducción a las versiones griegas de la Biblia*. Madrid: Consejo Superior de Investigaciones Científicas, ch.;

H. P. (1959). "Vier Aquila-Glossen in einem hebraischen Proverben-Fragment aus der Kairo-Geniza." *Zeitschrift für die Neutestamentlische Wissenschaft* 50: 275-277;

Hesseling, D. S. (1901). "Le livre de Jonas." *Byzantinische Zeitschrift* 10: 208-217.

RASHI 1040-1990. *Hommage à Ephraïm E. Urbach*. ed. G. Sed-Rajna. Paris, Éditions du Cerf: 203-10.

Tchernetska, N., Olszowy-Schlanger J., et al. (2007). "An Early Hebrew-Greek Biblical Glossary from the Cairo Genizah." *Revue des Études Juives* 166(1-2): 91-128.

Multi-Dimensional Markup: N-way relations as a generalisation over possible relations between annotation layers

Harald Lüngen
luengen@uni-giessen.de
Justus-Liebig-Universität, Germany

Andreas Witt
andreas.witt@uni-tuebingen.de
University of Tübingen, Germany

Text-technological background

Multi-dimensional markup is a topic often discussed. The main reasonwhy it is researched is the fact that the most important markup languages today make the implicit assumption that for a document, only a single hierarchy of markup elements needs to be represented. Within the field of Digital Humanities, however, more and more analyses of a text are expressed by means of annotations, and as a consequence, segments of a text are marked up by tags relating to different levels of description. Often, a text is explicitly or implicitly marked up several times. When using the TEI P5 as an annotation scheme one might use markup from different TEI modules concurrently as 'msdescription' for manuscript description, 'textcrit' for Text Criticism, and 'analysis' for (linguistic) analysis and interpretation, because the Guidelines state that "TEI schema may be constructed using any combination of modules" (TEI P5 Guidelines).

Abstracting away from limitations of specific markup languages, textual regions annotated according to different levels of descriptions can stand in various relationsships to each other. Durusau & O'Donnell (2002) list 13 possible relationsships between two elements A and B used to concurrently annotate a text span. Their list comprises the cases 'No overlap' (independence), 'Element A shares end point with start point of element B or the other way round', 'Classic overlap', 'Elements share start point', 'Elements share end point' and 'Element share both their start points and end points'. The latter case is known under the label 'Identity'. The possible relationships between A and B can also be partitioned differently, e.g. into Identity, Region A before region B, or the other way round. Witt (2004) has alternatively grouped the relations into three 'meta-relations' called 'Identity', 'Inclusion', and 'Overlap'. Meta-relations are generalisations over all the 13 basic relations inventorised by Durusau & O'Donnell. The reason for introducing meta-relations is to reduce the number of relations to be analysed to those cases that are most typically needed when querying annotations of multiply annotated documents. The query tool described in Witt et

al. (2005) provides 7 two-way query predicates for the 13 basic relations from Durusau & O'Donnell (where e.g. the two relations overlap(A,B) and overlap(B,A) are handled by one query predicate) and specialised predicates for the three meta-relations.

As argued above, often n-way relationships between elements from three or more annotation layers need to be queried. When the detailed accounts of cases of relations between two elements described above are extended to cases where three or more layers are analysed, the number of possible relationships is subject to a combinatorial explosion and rises into several hundreds and thousands. Only in the case of identity(A,B), additional 13 cases of three-way relationships can be distinguished; for all remaining cases of two-way relationships, considerably more three-way cases need to be distinguished. It seems impossible to invent names, let alone to formulate and implement queries for each one of them. Still, for a user it would be desirable to have a set of query predicates for n-way relations available, lest (s)he needs to repeatedly combine queries for two-way relationships, which often can be done only with the help of a fully-fledged programming language.

Application: Analysing n-way relations in text parsing

One text-technological application where relations between elements on more than two elements need to be analysed, is discourse parsing of argumentative texts. In a bottom-up operating discourse parser such as the one developed for German research articles in the SemDok project (Lüngen et al. 2006), it is checked successively whether a discourse relation holds between two known adjacent discourse segments such that they can be combined to form a larger segment. Often this depends on the presence of a lexical discourse marker, such as the adverb 'lediglich' ('only'), in the second segment. But with 'lediglich' as with numerous other markers, there is the additional condition that it has to occur in the so-called vorfeld (first topological field of a German sentence according to the syntax of German, cf. Hinrichs & Kübler 2006), of the first sentence of the second discourse segment. Thus, a combination of information from at least three different information levels (discourse segments, syntax, and discourse markers) needs to be checked, i.e. whether the following situation holds:

```
L1: <ds>.........................
.........................</ds>
L2: <s><vorfeld>.................
....</vorfeld>.............</s>
L3: <dm>lediglich</dm>
```

This situation corresponds to a meta-relation of three-way inclusion: <ds> from Layer 1 must include a <vorfeld> from Layer 2, which in turn must include a <dm> from Layer 3.

Querying n-way relations between elements of multiple annotations

We have identified a set of n-way meta-relations that are typically needed in text-technological applications for multiply annotated documents, namely N-way independence, N-way identity, N-way inclusion, and N-way overlap, (where independence, identity, and inclusion hold between the elements from all n layers, and overlap holds between at least one pair among the n elements). The proposed poster presentation illustrates further examples from text-technological applications such as discourse analysis and corpus linguistic studies, where querying n-way relations between elements is required. It explains our set of query predicates that have been implemented in Prolog for n-way meta-relations, and how they are applied to the examples. Furthermore it presents an evaluation of their usability and computational performance.

References

Durusau, Patrick and Matthew Brook O'Donnell (2002). *Concurrent Markup for XML Documents*. XML Europe 2002.

Hinrichs, Erhard and Sandra Kübler (2006). What Linguists Always Wanted to Know About German and Did not Know How to Estimate. In Mickael Suominen, Antti Arppe, Anu Airola, Orvokki Heinämäki, Matti Miestamo, Urho Määttä, Jussi Niemi, Kari K. Pitkänen and Kaius Sinnemäki (eds.): *A Man of Measure : Festschrift in Honour of Fred Karlsson on his 60th Birthday.* The Linguistic Association of Finland, Special Supplement to SKY Journal of Linguistics 19. Turku, Finland.

Lüngen, Harald, Henning Lobin, Maja Bärenfänger, Mirco Hilbert and Csilla Puskas (2006). Text parsing of a complex genre. In Bob Martens and Milena Dobreva (eds.): *Proceedings of the Conference on Electronic Publishing (ELPUB 2006)*. Bansko, Bulgaria.

Witt, Andreas (2004). Multiple hierarchies: New aspects of an old solution. In *Proceedings of Extreme Markup Languages.* Montreal, Canada.

Witt, Andreas, Harald Lüngen, Daniela Goecke and Felix Sasaki (2005). Unification of XML Documents with Concurrent Markup. *Literary and Linguistic Computing* 20(1), S. 103-116. Oxford, UK.

Bibliotheca Morimundi. Virtual reconstitution of the manuscript library of the abbey of Morimondo

Ernesto Mainoldi
ernesto.mainoldi@libero.it
University of Salerno, Italy

Bibliotheca Morimundi has been financed by CNR (Consiglio Nazionale delle Ricerche, Rome) in the context of the general national project L'identità culturale come fattore di integrazione (http://www.cnr.it/sitocnr/IlCNR/Attivita/PromozioneRicerca/Identitaculturale.html), CNR, Rome, 2005-2006.

http://www.cnr.it/sitocnr/IlCNR/Attivita/PromozioneRicerca/IdentitaCulturale_file/IdentitaculturaleGraduatoriaPG.html

Presentation of the plan of search and its object

The aim of the project *Bibliotheca Morimundi* is to recompose in a database of digital images the written heritage that belonged to the library of the Cistercian abbey of Morimondo (Milan), dispersed after the suppression of the monastery in 1796. The collection was composed by medieval manuscripts, renaissance incunabula and diplomas. According to medieval catalogues and to ex libris about 80 extant manuscripts (dating from the middle of 12th century to the 15th century), 3 incunabula and hundred of medieval and post-medieval diplomas belonged to the abbey of Morimondo. Morimondo manuscripts are generally characterised by a remarkable thematic range and relevant artistic preciousness: Bibles and liturgical books decorated with fine miniatures, books with musical notation (among which is the most ancient known antiphonary of Cistercian order), patristic texts that show an interest for authors from the Eastern Christianity, and finally juridical, scientific and theological codices that are witnesses of the participation of the Morimondo monastery to the European renewal of the twelfth and thirteenth century. Some of the diplomas are of high historical importance, as papal and imperial documents, and are now conserved at State Archive of Milan.

Since manuscripts from Morimondo are now conserved in ca. twenty libraries dislocated all around the world, a comparative and detailed study of all these sources has not yet been possible: the digital resource will allow a global study of this collection, based on the comparison of handwritten texts (manuscripts and diplomas) produced in the very same scriptorium. On the basis of these features the *Bibliotheca Morimundi* project will be a digital library that differentiates itself from other digital libraries, which build on an extant collection (see the Sankt Gallen or Köln projects: http://www.cesg.unifr.ch/it/index.htm; http://www.ceec.uni-koeln.de/). In the case of Morimondo the digital library takes the place of the no more extant library.

The virtual recollection of the manuscripts once belonged to Morimondo abbey in a digital database should allow a systematic study of the history and the culture developed in this monastic site, a singular oasis in the Ticino's valley, at 40 km from Milan, where religious life and textual production where cultivated together. Throughout the possibility to compare these manuscripts, of the majority of which were written in Morimondo, scholars would have the opportunity to detect palaeographical, codicological, liturgical and textual features of the monastic culture in Morimondo.

The realization of the database *Bibliotheca Morimundi* not only will reconstitute the ancient unity of the library of the abbey, but is also intended as an interactive instrument of research, by providing a wealth of resources including digital images of the manuscripts, incunabula and diplomas ,texts transcriptions, indices, and bibliography, organized as an hypertext structure capable to facilitate the research and to integrate progressing results and dissemination of the studies (such as printed papers but also notes made by users).

The Database

The database is accessed as a web site structure in html language. The consultation of the digitals images, due to copyright reasons, is allowed – at least in this first phase of the realization – only off line, on local computer terminals hosted by the institutions which support the project (Fondazione Franceschini of Florence and Fondazione Rusca of Como). Catalogues and textual resources will be published on line. When a library owner of manuscripts from Morimondo didn't allow a digital reproduction a digitized microfilm is inserted in the database. The pictures, provided as 8 bit colour JPEG at the resolution of 300 dpi, are readable in PDF format. The metadata will be codified in XML language following the MAG standard, http://www.iccu.sbn.it/genera.jsp?id=267, which scheme consent to mark either the digital images or the OCR files of secondary bibliography and other textual resources (such transcription, notes progressively added by users etc.).

Status of the work

At present about 30 of 78 mss. have been photographed. The most important collection of Morimondo manuscripts (summing 18 mss.), owned by the Biblioteca del Seminario Vescovile di Como, has been fully photographed in digital colour pictures and constituted the current base of the archive of digital photos. Others single codices kept in monastic library (such as Tamié in France or Horsham in England) have been digitally photographed. Digital images amount, up to now, at the number of 7000 ca. Others main collections, such as those kept in the Bibliothèque Nationale de France in Paris

or in the British Library in London, have been acquired as b/w microfilms and converted in digital images with a slide scanner. In a draft version the database is already consultable at the library of the Fondazione Rusca of Como, where also the mss. of the Biblioteca del Seminario are conserved. Some scholars have already begun their studies by using the data (images and texts) and metadata (codicological data about the manuscript source, digital image informations) already inserted in the database.

Developments of the search

A fundamental direction of development of the «Bibliotheca Morimundi» project will the multidisciplinary integration of differentiated scientific competences in the study of the sources included in the database. Already a team of scholars with academic roles has been contacted or involved in the project. The integration of digitals images and textual instruments that will progressively facilitates the study of the Morimondo primary sources will also constitute a field of application for developers of computational resources in the humanities. Another development will be the incentive toward the interrelation of several type of cultural institutional needs, such the needs of research and the needs of historical preservation and divulgation. The project has already made progress on the basis of such a cooperation between an institution involved in active scientific research (Fondazione Franceschini – SISMEL, one of the most active institution, at international level, in the study of the Middle Ages) and a library (Biblioteca del Seminario vescovile di Como, keeper of 18 manuscripts from Morimondo). Other conventions for scientific cooperation between institutions are currently in progress.

Bibliography

Correlated Projects

Codices Electronici Ecclesiae Coloniensis (CEEC)

http://www.ceec.uni-koeln.de/

Codices Electronici Sangallenses (CESG) – Virtual Library

http://www.cesg.unifr.ch/it/index.htm

Digital Image Archive of Medieval Music

http://www.diamm.ac.uk/

Manuscriptorium

http://www.manuscriptorium.com/Site/ENG/default_eng.asp

Monasterium Projekt (MOM)

www.monasterium.net

Papers

Michele Ansani, Edizione digitale di fonti diplomatiche: esperienze, modelli testuali, priorità, «Reti medievali VII- 2006 / 2 - luglio-dicembre», http://www.dssg.unifi.it/_RM/rivista/forum/Ansani.htm

Arianna Ciula, L'applicazione del software SPI ai codici senesi in Poesía medieval (Historia literaria y transmisión de textos) cur. Vitalino Valcárcel Martínez - Carlos Pérez González, Burgos, Fundación Instituto Castellano y Leonés de la Lengua 2005 pp. 483 (Beltenebros 12), pp.?305-25

Emiliano Degl'Innocenti, Il Progetto di digitalizzazione dei Plutei della Biblioteca Medicea Laurenziana di Firenze «DigItalia». Rivista del digitale nei beni culturali, Roma = DigItalia 1 (2007), pp. 103-14

Digitalisierte Vergangenheit. Datenbanken und Multimedia von der Antike bis zur frühen Neuzeit, cur. Christoph Schäfer - Florian Krüpe, Wiesbaden, Harrassowitz 2005 pp. XI-147 tavv. (Philippika. Marburger altertumskundliche Abhandlungen 5)

Investigating word co-occurrence selection with extracted sub-networks of the Gospels

Maki Miyake
mmiyake@lang.osaka-u.ac.jp
Osaka University, Japan

Graph representation and the techniques of graph theory and network analysis offer very effective ways of detecting and investigating the intricate patterns of connectivity that exist within large-scale linguistic knowledge resources. In this presentation, we apply a graph-clustering technique for data processing that utilizes a clustering-coefficient threshold in creating sub-networks for the Gospels of the New Testament. Specially, this study discusses some graph clustering results from the perspectives of optimal clustering and data sizes with a view to constructing an optimal semantic network by employing the hierarchical graph clustering algorithm of Recurrent Markov Clustering (RMCL). The corpus used in this study is a Greek version of the Gospels (Nestle-Aland, 1979) for which two semantic networks are created based on network features. The network structures are investigated from the perspective of constructing appropriate semantic networks that capture the relationships between words and concepts.

In the natural language processing of texts, word pairs are usually computed by the windowing method (Takayama, Flournoy, Kaufmann, & Peters, 1998). The windowing method provides a relatively simple representation of similarity levels that is suitable for clustering. The technique involves moving a certain sized window over a text to extract all fixed-sized word grams (Vechthomova, Roberston, & Jones, 2003). Word pairings are then made by combining all extracted words. In the present study, co-occurrence data is computed with two window sizes that reflect syntactic and semantic considerations. The first size (WS1) is set at 1 for the nearest co-occurrence, while the second size (WS10) is set to 10 to collect words and no stemming process was taken. Accordingly, 8,361 word occurrences were identified. An adjacency matrix for the graphs was created from the word co-occurrence data for a particular range of the texts. Using the clustering coefficient as a threshold, the network was reduced into 18 sub-networks which were created at 0.1 increments of the clustering coefficient value (from 0.1 to 0.9).

In Figure 1, the degree distributions for the two networks for all occurrence nodes are plotted along log-log coordinates. The window size of 1 shows the best fit to a power law distribution (r=1.7). The average degree value of 15.5 (2%) for the complete semantic network of 8,361 nodes clearly indicates that the network exhibits a pattern of sparse connectivity; in other words, that it possesses the characteristics of a scale-free network according to Barabasi and Albert (1999). In contrast, the window size of 10 has a bell-shaped distribution and its average degree value of 106.4 (13%) indicates that the network exhibits a much greater level of connectivity compared to the window size of 1.

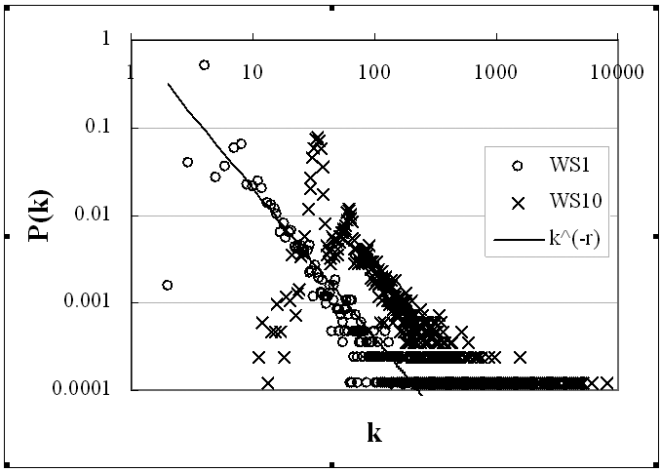


Figure.1 Degree distributions

The second network feature is the clustering coefficient which represents the inter-connective strength between neighboring nodes in a graph. Following Watts and Strogatz (1998), the clustering coefficient of the node (n) can be defined as: (number of links among n's neighbors)/(N(n)*(N)-1)/2), where N(n) denotes the set of n's neighbors. The coefficient assumes values between 0 and 1.

Moreover, Ravasz and Barabasi (2003) advocate a similar notion of clustering coefficient dependence on node degree, based on a hierarchical model of scaling laws. The results of scaling C(k) with k for the two networks are presented in Figure 2. As the two networks conform well to a power law, we can conclude that they both possess intrinsic hierarchies.

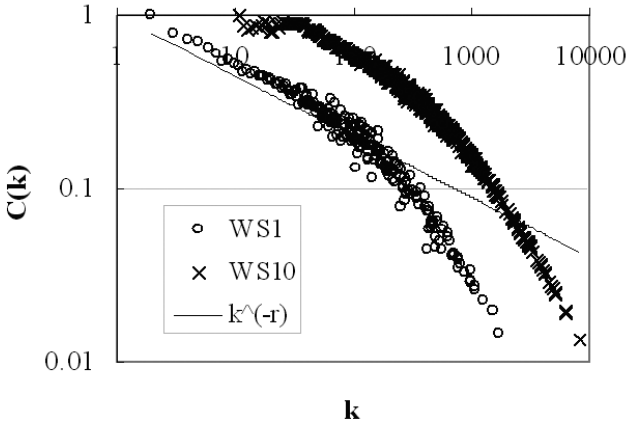


Figure.2 Clustering coefficient distributions

The graph clustering method applied in this study is Recurrent Markov Clustering (RMCL), recently proposed by Jung, Miyake,

and Akama (2006), as an improvement to Van Dongen's (2000) Markov Clustering (MCL) which is widely recognized as an effective method for the detection of patterns and clusters within large and sparsely connected data structures (Dorow, Widdows, Ling, Eckmann, Sergi & Moses (2005), Steyvers & Tenenbaum (2005)). The first step in the MCL consists of sustaining a random walk across a graph. The recurrent process is essentially achieved by incorporating feedback data about the states of overlapping clusters prior to the final MCL output stage. The reverse tracing procedure is a key feature of RMCL making it possible to generate a virtual adjacency matrix for non-overlapping clusters based on the convergent state yielded by the MCL process. The resultant condensed matrix represents a simpler graph that can highlight the conceptual structures that underlie similar words.

Figure 3 presents cluster sizes as a function of the clustering coefficient for the two window sizes of 1 and 10. The terms for the input data (WS1-data, WS10-data) refer to the two initial adjacency matrices. Focusing on the trend in cluster sizes for the window size of 10 (WS10), the fact that cluster sizes remain relatively constant regardless of size of the data indicates that as the window size increases, clusters are less dependent on the clustering coefficient.

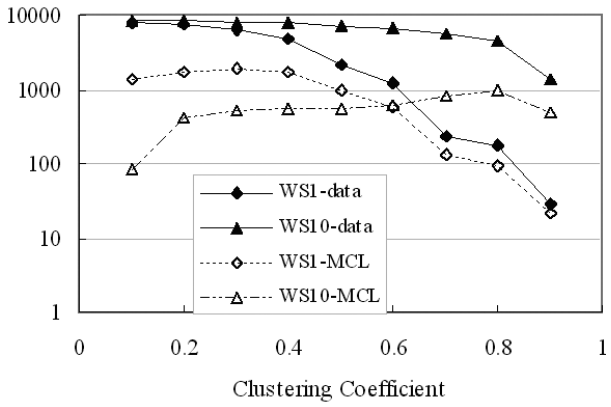


Figure.3 MCL results

In order to optimize the co-occurrence data, we employ Newman and Girvan's (2004) notion of modularity, which is particularly useful in assessing the quality of divisions within a network. Modularity Q indicates differences in edge distributions between a graph of meaningful partitions and a random graph under the same vertices conditions (numbers and sum of their degrees). Accordingly, Miyake and Joyce (2007) have proposed the combination of the RMCL clustering algorithm and this modularity index in order to optimize the inflation parameter within the clustering stages of the RMCL process.

Moreover, to consider the recall rates of nodes, the F measure is also employed to optimize the selection of the most appropriate co-occurrence data, because the precision rate, P, always depends on a trade-off relationship between modularity Q and the recall rate R.

Figure 4 plots modularity Q and the F measure as a function of clustering coefficients for the two window sizes of 1 and 10. The results indicate that there are no peaks in the plot of Q values. This finding suggests that as the value of r decreases, the value of Q increases. In the case of WS1, there are differences in the peaks for the two measures of Q and F, for while Q peaks at 0.6, the value of F peaks at 0.4. In this study, we regard the peak in the F measure as a recall rate according to the size of the data. As the results for WS10 show no peaks in the Q value, we also take the peak value for F which is 0.7. In this way, the number of node is about 5,000 for the two selected sub-networks, which are almost the same size as the networks.

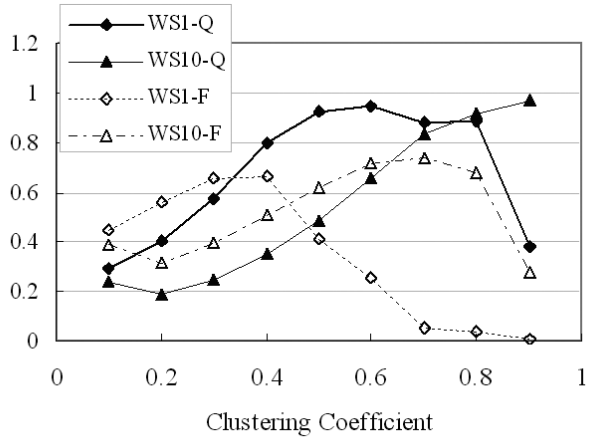


Figure 4.

In conclusion, two network representations of the Gospels were created for different window sizes of text, and the networks were analyzed for the basic features of degree and clustering coefficient distributions in order to investigate the presence of scale-free patterns of connectivity and hierarchical structures. This study also reported on the application of a hierarchical graph clustering technique to sub-networks created with different clustering-coefficient thresholds. In terms of constructing an optimal semantic network, the combination of the modularity measurement and the F measure is demonstrated to be effective in controlling the sizes of clusters.

Bibliography

Barabási, A.L., and R. Albert (1999), Emergence of Scaling in Random Networks. *Science*, 86, pp.509-512.

Dorow, B., D. Widdows, K. Ling, J. Eckmann, D. Sergi, and E. Moses (2005), Using Curvature and Markov Clustering in Graphs for Lexical Acquisition and Word Sense Discrimination. *MEANING-2005, 2nd Workshop organized by the MEANING Project*, February 3rd-4th 2005, Trento, Italy. Trento, Italy.

Jung, J., Miyake, M., Akama, H. (2006), Recurrent Markov Cluster (RMCL) Algorithm for the Refinement of the Semantic Network, In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC2006)*, pp.1428-1432.

Nestle-Aland (1987), Novum Testmentum Graece. 26th edition. Stuttgart: German Bible Society, 1987.

Miyake, M., Joyce, T (2007), Mapping out a Semantic Network of Japanese Word Associations through a Combination of Recurrent Markov Clustering and Modularity. In *Proceedings of the 3rd Language & Technology Conference (L&TC'07)*, pp.114-118.

Steyvers, M., and J. B. Tenenbaum (2005), The Large-Scale Structure of Semantic Networks: Statistical Analyses and a Model of Semantic Growth. *Cognitive Science* 29.1, pp.41-78.

Takayama, Y., R. Flournoy, S. Kaufmann, and S. Peters. Information Mapping: Concept-based Information Retrieval Based on Word Associations. http://www-csli.stanford.edu/semlab-hold/infomap/theory.html, 1998.

van Dongen, Stijn Marinus (2000), "Graph Clustering by Flow Simulation". PhD Thesis. University of Utrecht. http://igitur-archive.library.uu.nl/dissertations/1895620/inhoud.htm

Vechthomova, O., S. Robertson, and S. Jones (2003), Query Expansion With Long-Span Collocate. *Information Retrieval* 6, pp.251-273.

Watts, D., and S. Strogatz (1998), Collective Dynamics of 'Small-World' Networks. *Nature* 393, pp.440-442.

Literary Space: A New Integrated Database System for Humanities Research

Kiyoko Myojo
kiyokomyojo@gmail.com
Saitama University, Japan

Shin'ichiro Sugo
shinsugouakita@mac.com
Independent System Engineer, Japan,

We describe here a basic design of a multi-purpose database system for Humanities research that we are developing. The system named "LiterarySpace" provides a new integrated electronic workbench for carrying out various kinds of processing acts on textual study: documentation, transcription, interpretation, representation, publication and so on (Shillingsburg 2006). In other words, it has a polar range of possibilities of functions from a private tool to a publishing medium: image data of any materials such as inscribed stones, wooden tablets, papyrus rolls, paper codex etc. can be stored for private and/or for public use, and recorded text data such as the results of electronically scholarly editing and/or outcomes of hermeneutic inquiry with/without relation to the graphics, and then represent all or part of the data archives to yourself in a stand-alone computer and/or to anyone via the internet.

The essential architecture of this relational database is simple: It is constructed from three database files: file-1, graphic data file = GDF; file-2, text data file = TDF; file-3, control data file = CDF. GDF has two data fields in each of its records: the one for storing image data and the other for its unique serial numbers the system automatically issues. We named this identification number IDGraphic. A record in TDF consists of three data fields which keep three kinds of data: 1) text data, 2) its unique identification integers assigned by the system which we named IDText, 3) IDGraphic as a relational key to GDF. In terms of the relation between GDF and TDF, one distinguishing feature is that one GDF record allows an overlapped link to plural TDF records; namely the correspondence of records in GDF to ones in TDF is one-to-many and not vice versa. A CDF record contains the following two fields: 1) IDSet field in which IDGraphic or IDText is registered as a relational key, and 2)Index field which keeps numbers that have been input in order to make the sequence of the records called out by the IDSet. As for the identification of the type of IDSet, with just the two alternatives, IDGraphic or IDText, the name of the CDF serves as a mark. If the filename given by the user contains the word "GDF" or "gdf" , the file is interpreted as a set of records from GDF. Otherwise, it is being from TDF.

The powerfulness of "LiterarySpace" is caused from this simple structure which corresponds to the theoretical model we made up based on our clear understanding about space

and time in writing (Myojo & Uchiki 2006). Concerning space, written existence demands two different spaces: physical space which every written text occupies, and conceptual space which the text generates in the minds of writers and readers at the same time (Bolter 1991). On the other hand, time is time; time is never physical but ever visionary and such time in a written text can be controlled by the writer and even the reader. The three database files of our system coincide with these epistemological dimensions: GDF with the physical space, TDF with the conceptual space, and CDF with the variable time, or a little more precisely, fictional spaces in which each conducted time goes (Stanzel 1989). Within the one space of the one "LiterarySpace" there can exist only one GDF, one TDF, and plural CDF(s). The space of the single GDF is constituted of image data as non-character-coded documentation of extant materials. And the time of its space is represented by the sequential order indicated by integers assigned mechanically for each additional record. Needless to say, the order corresponds to the succession in which the records are input. The one-time-only character of the inputting acts in reality is the reason why the integer, i.e. IDGraphic is unique. So you might call the time generated by the unique numbers the "real" one. The single TDF is made up from text data as character-coded documentation of any writing, that is to say, the products of conceptual work; and the space has also the single "real" time represented in the order designated by the IDText integers, which naturally corresponds to the succession in which the data are put. The philosophical substance of CDF(s) is, however, different; its space is a meta-space which consists of data records chosen from GDF or TDF. As the space is, so to speak, fictional, it is possible to be set in imaginary time. The meta-time can be produced by the sequence of the Index-numbers issued not by machine but by the user. Because of the fictionality, one can make more than one sequence if necessary. So, unlike other types of field, the Index-field can be increased. To be sure, there also exists "real" time in accordance with the succession of the records input into the CDF and if no sequence is created, the records would be represented in the input order.

The first part of our presentation will illustrate this system design more in detail, explaining significant features including e.g. the reason of the above mentioned one-to-many correspondence of GDF to TDF in connection with the theoretical modelling (McCarty 2005).

In the second part we will demonstrate a prototype of the system which is built using the database application "FileMaker Pro" (After the specification of the software has been completed, its program will be rewritten in Common Lisp). The current main database in the prototype is Franz Kafka's database constituted of information particularly from his "Oktavhefte". One of us, Myojo is originally a Kafka researcher and the primal need of this system arose in the context of her literary research project (Myojo 2002). Myojo's study following her own methodology as a combination between "Editionswissenschaft" and "critique géntique" has always demanded a cross-referenced complicated work handling three editions simultaneously: the practical edition by Max Brod, the critical edition "Kritische Kafka-Ausgabe", and the facsimile edition "Franz Kafka-Ausgabe". So we should admit that the first attempt of our modelling was not theoretical but practical (Myojo 2004). It is, however, because our project started to grasp the actual needs of the one scholar in the Humanities that this system has grown up as a real useful entity. Using the Kafka database as an example, we would like to demonstrate not only how the system operates but also how efficiently the performance assists the investigation for work in the humanities.

The third part will be dedicated to showing the other feasibilities of this system. One of the many defining characteristics of this system is that it allows us to contain several databases at the same time. For instance in addition to Kafka's database one can also construct Shakespeare's as well as the work of some Japanese authors' etc. in the same system. Also, a more significant point is that one can input as much data as necessary without worrying about grouping or sequences. Simply put, one can make a record of Kafka's information just after inputting that of Shakespeare's. This unrestrictedness is enabled by the aforementioned epistemological data structure because the grouping and the sequence can be controlled afterwards by means of creating a meta-area, i.e. a CDF. Presently in our prototype two more databases of Japanese authors are under construction: Kenji Miyazawa (宮沢賢治)'s and Soseki Natsume (夏目漱石)'s. The strong merits of this capability will be presented as an exhibition of research results in the field of comparative literature (Myojo 2003).

One more important point we have to mention: as we noted at the beginning of this text, the system serves as a private tool and/or a publishing medium. The significance of this feature would be recognized well if you imagined the case of dealing with works protected by copyright. Of course the dual aspects of privateness and publicness correspond to the original distinguishing character of the computer itself. From not only this but the all above points of view it might be suggested that this system could become a powerful electronic substitute for a physical notebook. In the last part of the presentation we would like to discuss the multifaceted possibilities, i.e. the universality of the system.

Acknowledgments

This work was funded in part by The Japan Society for the Promotion of Science (JSPS) under Grant-in-Aid for Scientific Research (C) [18529001]. We would also like to thank Christian Wittern (Kyoto University), Tetsuya Uchiki (Saitama University) and other members of The Japanese Association for Scholarly Editing in the Digital Age (JASEDA) for their help in establishing regular and fruitful discussions.

Bibliography

Bolter, David J., *Writing Space*, Hillsdale (Lawrence Erlbaum Associates) 1991.

McCarty, Willard, *Humanities Computing*, Houndmills (Palgrave McMillan), 2005.

Myojo, Kiyoko, *Kafka Revisited*, Tokyo (Keio University Press), 2002 [in Japanese].

Myojo, Kiyoko, "Kafka und sein japanischer Zeitgenosse Kenji", *Saitama University Review* 39.2 (2003): 215-225.

Myojo, Kiyoko, "What is the Future of Computer-Assisted Scholarly Editing?" *IPSJ SIG Technical Reports* CH-62 (2004): 37-44 [in Japanese].

Myojo, Kiyoko & Tetsuya Uchiki: "A Theoretical Study on Digital-Archiving of Literary Resources", *IPSJ Symposium Series* 17 (2006): 153-160 [in Japanese].

Shillingsburg, Peter L., *From Gutenberg to Google*, Cambridge (Cambridge University Press) 2006.

Stanzel, Franz K., *Theorie des Erzählens*, Göttingen (Vandenhoec & Ruprecht), 1989 [1979].

A Collaboration System for the Philology of the Buddhist Study

Kiyonori Nagasaki
nagasaki@ypu.jp
Yamaguchi Prefectural University, Japan

In the field of the Buddhist study, especially, the philosophy of the Buddhism, there are many texts which have been passed down throughout history since around 5th century BCE. In spite of the long history, philological study started around the 18th century CE and has not yet been considered adequate. The original texts which were written in India do not remain, but many manuscripts copied by Buddhist monk scribes remain in Indic languages such as Sanskrit or Pāli or translated into Tibetan or classical Chinese. Those translated texts alone consist of huge number of pages. Some of them have been published in the scholarly editions, but many texts are not published, or are not verified as reliable texts. Moreover, sometimes old Sanskrit manuscripts may be newly discovered in Nepal, Tibet or others. Under these conditions, even if a new scholarly edition is published, it may have to be edited again with the discovery of a newly-found Indian manuscript. Therefore, in the field, a collaboration system for the sake of editing of the texts on the Internet would be effective so that reliable texts could be edited anytime, sentence-by-sentence, word-by-word, or even letter-by-letter on the basis of those witnesses. The collaboration system must be able to:

(1) represent and edit the texts and the facsimile images in a multilingual environment such as Sanskrit, Tibetan and CJK characters.

(2) store the information of the relationship between each of the content objects which are included in the each of the witnesses as text data and facsimile image.

(3) add the physical location of those objects to the information.

(4) record the name of the contributors of each piece of information.

It is not difficult to represent and edit the texts and the facsimile images in a multilingual environment by the popularization of UTF-8. However, regarding CJK characters that are either not yet in Unicode or belong to the CJK-Extension B area, the collaboration system adopts a method called "Mojiyaki" to represent the glyphs of the characters as the image files on Web browsers and a character database called "CHISE" based on the IDS (Ideographic Description Sequence). Tibetan scripts are included in Unicode and supported in Windows Vista. But the font of the scripts is not included in the older versions of Microsoft Windows. In the field, most researchers use the Tibetan texts on the computers by means of their

transliteration into ASCII characters. Thus, according to these conventions, the system must support at least ASCII characters for the Tibetan characters. Indic characters are the same as the Tibetan characters.

The content objects which are included in each of the witnesses such as the paragraphs, the sentences, the words or the letters in the text data or the fragments of the facsimile image have relationships to others in the context of the history of thought in Buddhism. The collaboration system provides a function that will describe such information about the relationships separately from the literal digital copies of the witnesses. The information at least consists of the location of the content objects in the witness such as the paragraph, the sentence or the word and the attributes of the relationship such as variants, quotations, translations, and so on. Because some relationships have a kind of a hierarchy, it must be reflected in the collaboration system. However, it is important to keep the flexibility in the methods of the description of the hierarchy because the hierarchy is often different in each tradition of the texts. One more important thing is that copyright problems might be solved by describing the information separately from the digital copies of the witnesses.

It is important to describe the physical location of the content objects by the means of the traditional methods such as page and line number so that the information of the relationship can maintain interchangeability with the traditional methods which refer to their physical location in the witness.

The collaboration system must record the name of the contributors of the information so that responsibility for the information can be shown explicitly and users can filter the unnecessary information.

The prototype of a collaboration system which implements the above functions is already completed as a Web application using the "Madhyamaka Kārikā" which is a famous philosophical Buddhist text that has been quoted or referred in other texts since about the 3rd Century. It reflects opinions of some Buddhist researchers. It is working on GNU/Linux using Apache HTTPD server and PostgreSQL and coded by PHP and AJAX so that users can do all of the works on their Web browsers. All of them consist of free software. It can be demonstrated on the Digital Humanities 2008. Moreover, At present, I am attempting to describe the relationships by use of RDF, OWL and the elements defined by the TEI. The method of the description will be also shown at the conference.

Bibliography

Caton, Paul, "Distributed Multivalent Encoding", *Digital Humanities 2007 Conference Abstracts*, pp. 33-34. (2007).

DeRose, Steven, "Overlap: A Review and a Horse", *Extreme Markup Languages 2004: Proceedings,* http://www.mulberrytech.com/Extreme/Proceedings/html/2004/DeRose01/EML2004DeRose01.html. (2004).

MORIOKA, Tomohiko, "Character processing based on character ontology", *IPSJ SIG Technical Report, 2006-CH-072*, pp. 25-32. (2006).

Nagasaki, Kiyonori, "Digital Archives of Indian Buddhist Philosophy Based on the Relationship between Content Objects", IPSJ SIG Technical Report, 2007-CH-75, pp. 31-38. (2007).

Nagasaki, Kiyonori, Makoto MINEGISHI and Izumi HOSHI, "Displaying multi-script data on the Web", *Proceedings of the Glyph and Typesetting Workshop, 21st Century COE Program "East Asian Center for Informatics in Humanities - Toward an Overall Inheritance and Development of Kanji Culture - " Kyoto University*, pp. 44-51. (2004).

Renear, Allen H., "Text Encoding", *A Companion to Digital Humanities*, Blackwell Publishing, 2004, pp. 218-239. (2004).

Steinkellner, Ernst, "Methodological Remarks On The Constituion Of Sanskrit Texts From The Buddhist Pramāṇa-Tradition", *Wiener Zeitschrift für die Kunde Südasiens*, Band XXXII, pp. 103-129. (1998).

Information Technology and the Humanities: The Experience of the Irish in Europe Project

Thomas O'Connor
thomas.oconnor@nuim.ie
National University of Ireland, Maynooth, Ireland

Mary Ann Lyons
marian.lyons@spd.dcu.ie
St. Patrick's College, Ireland

John G. Keating
john.keating@nuim.ie
National University of Ireland, Maynooth, Ireland

Irish in Europe Project

As a result of its leadership role within the field of early modern migration studies, the Irish in Europe Project, founded in NUI Maynooth in 1997, is now the custodian of a rich and constantly expanding corpus of biographical data on Irish migrants in early modern Europe. The most substantial data sets relate to Irish students in the universities of Paris and Toulouse in the seventeenth and eighteenth century, Irish patients in Parisian hospitals in the early eighteenth-century, Irish students in Leuven University; Irish students in France's leading medical faculties, and Irish regiments in eighteenth-century France and Spain.

In line with the Project's mission to make this data accessible to researchers, teachers and students and to ensure that the material is rendered responsive to the full range of users needs, the Irish in Europe Project undertook the development of an electronic research environment for the collection, management and querying of biographical data on Irish migrants in early modern Europe. This was facilitate by the award of a Government of Ireland Research Project Grant in the Humanities and Social Sciences in 2006, which accelerated progress in reaching four target objectives, namely to encode these databases in extensible Mark up language (XML), to provide internet based access to the databases, to establish a Virtual Research Environment (VRE) to facilitate research and teaching, and lastly, to manage and promote the adoption of best practice for biographical databases. In this encoded format, the data is rendered susceptible to a broad range of queries and manipulations, permitting, for instance, the automatic generation of graphics, maps, tables etc. This project will be rolled out in 2007-9. In July-November 2007 important progress was made on the project in the context of the preparing a database Kiosk for the National Library of Ireland exhibition 'Strangers to Citizens: the Irish in Europe, 1600-1800' which opened in December 2007.

Kiosk Development

The Kiosk is essentially user-friendly software providing access to research derived from four different biographical studies of migration in the Early Modern period, i.e. student migration to France (Brockliss and Ferté, 1987; Brockliss and Ferté, 2004), student migration to Louvain (Nilis, 2006), military migration to France (Ó Conaill, 2005) and military migration to Spain (Morales, 2002; Morales, 2003; Morales, 2007). Data from the associated databases were provided in Microsoft Excel format or extracted using custom developed programs, from Microsoft Word versions and Portable Document Format (PDF) versions of the research papers. XML data models, similar to those previously described by Keating et al. (2004) were used to encode the four data sets. Each dataset contained information pertinent to the profession under study, and there was some overlap, particularly related to personal data. In general, student data were associated with migration from dioceses whereas military migration data were associated with counties. Central to the software requirements of this project was the specification that common software tools should be used to visualize differing data sets whenever possible for (i) usability issues and (ii) providing comparative analyses functionality which has never been available to Early Modern migration researchers, for example, see Figure 1.

The project team designed and developed two distinct and fundamental "database inspectors" essential for data visualization: (i) an interactive vector-based "heat-map" of Ireland which provides scaled diocese or county density migration patterns for a variety of inputs selected by the user, and (ii) an interactive record inspector which provides key record information based on exact, partial or phonetic searching of individual surnames and other personal features. The database inspectors were developed using Macromedia Flex and have been tested in a variety of modern browsers. Converted datasets reside on the Irish in Europe web server (http://www.irishineurope.ie) and are accessed using optimized searching software implemented as web services. Database inspectors communicate with the web services using CGI (outward); the latter return XML which describe how the maps or record inspectors should be drawn – the drawing functions are all implemented using Macromedia Flex and Flash, as shown in Figure 2.

The Kiosk, essentially hosting a virtual exhibition, was custom developed (using Objective C) within the project using Internet browser tools provided as part of the XCode and Safari development suite. The current version provides accessibility options including user selected zoom and pan, and an audio soundtrack for each virtual exhibition page. The complete system is available for operation in Kiosk mode or can be accessed online. Overall, this project was completed in 28 person months and required a wide range of software development and information architecture skills not possessed by a single individual. We propose to present the lessons learned from the management of the project and development process, in addition to those described below.

Lessons learned from Collaboration

The collaborative exercise involved in the production of NLI/ Irish in Europe Biographical Databases Kiosk yielded three critical results:

Firstly, the exercise revealed that the software architecture of the Kiosk had to reflect both the nature and form of the historical data used and the specifications of the clients. It was the practical interaction between the data collectors/ interpreters and the information technology experts that shaped the architecture of the Kiosk. These functions can not and ought not be compartmentalized and, in project of this nature, should not be hermetically sealed off from each other.

Secondly, the solution of technical and software engineering difficulties and challenges in the project involved not only increasingly defined specifications from the data collectors and the NLI clients but also critical input from the software engineering and technical teams. For instance, important inconsistencies in the raw information came to light thanks to the interrogations of the software and technical teams, who had tested the data. This led to a feedback, which permitted the creation of a practical working relationship between the humanities team, the software and technical teams and the clients. It is highly suggestive for the development of architectures for humanities-software-client-user relations in the future.

Thirdly, the expansion in the range of functions provided by the site was driven by the dynamic interface between the software/ technical teams and the information. While it is usual for the architecture of similar sites to be dictated by purely technical criteria or costing issues, this project revealed the possibility for creative input from the technical-raw data interface. It is rare for technical teams to be afforded this level of access to the data collection and interpretation/presentation.

References

Brockliss, L.W.B. and Patrick Ferté, P. (1987) Irish clerics in France in the seventeenth and eighteenth centuries: a statistical survey. *Proceedings of the Royal Irish academy* , 87 C, pp. 527-72.

Brockliss, L.W.B. and Patrick Ferté, P. (2004). Prosopography of Irish clerics in the universities of Paris and Toulouse, 1573-1792. *Archivium Hibernicum*, lviii, pp. 7-166.

Hernan, E. G. and Morales, O. R. (2007). *Essays on the Irish military presence in early modern Spain, 1580-1818*, Eds. Madrid, Ministerio de Defensa.

Keating, J. G., Clancy, D., O'Connor, T. and Lyons, M. A. (2004) Problems with databases and the XML solution. *Archvium Hibernicum*, lviii, pp. 268-75.

Morales, O. R. (2002) *El socorro de Irlanda en 1601*, Madrid.

Morales, O. R. (2003) *España y la pérdida del Ulster,* Madrid.

Morales, O. R. (2007). *The Irish military presence in the Spanish Armies, 1580-1818*. Madrid, Ministerio de Defensa.

Nilis, J. (2006). Irish students at Leuven University, 1548-1797'. *Archivium Hibernicum,* lx, pp. 1-304.

Ó Conaill, C. (2005) "'Ruddy cheeks and strapping thighs": an analysis of the ordinary soldiers in the ranks of the Irish regiments of eighteenth-century France' in *The Irish Sword*, xxiv (2005/5), pp 411-27.

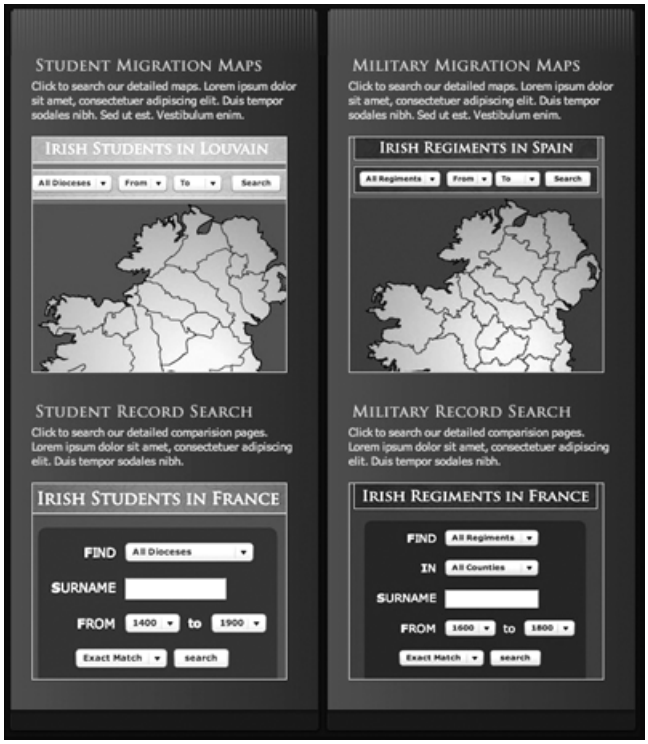


Figure 1: Extract from "Strangers to Citizens" Kiosk's Database Landing page

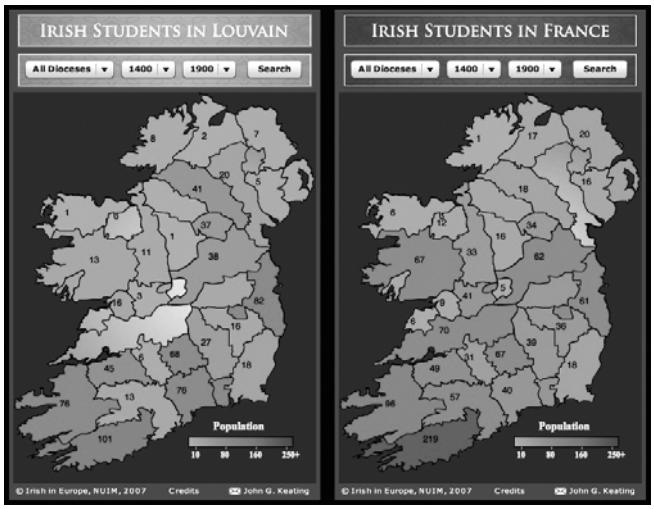


Figure 2: Extract from "Strangers to Citizens" Kiosk's Student Migration Comparison page showing two Database Inspectors (Diocese Density Maps)

Shakespeare on the tree

Giuliano Pascucci
giuliano@aconet.it
University of Rome "La Sapienza", Italy

This work illustrates some computer based tools and procedures which have been applied to the complete corpus of Shakespeare's works in order to create a phylogenetic tree[1] of his works and to discriminate between Shakespeare's and Fletcher's authorship in the writing of *All is True*, which is known to be the result of a collaboration of both authors.

The general procedure applied to this study was devised in 2001 by a group of University of Rome *La Sapienza*[2] scholars who asked themselves whether starting from a graphic character string it was possible to extract from it information such as the language in which the string itself was written, whether or not it belonged to a wider context (e.g. it is a part of a literary work), its author and its filiation from other texts.

Their method was based on the idea of linguistic entropy as it has been defined by Chaitin-Kolmogorov[3] and was applied to the translations into 55 different languages of a single text: the *Universal Declaration of Human Rights*.

Following Kolmogorov's theories and considering each of the 55 texts as a mere string of characters they decided to study the linguistic entropy of each text, that is to say the relationship between the information contained in a string of characters and the ultimate compressibility limit of the string itself. Because said limit, i.e. the complete removal of all redundancies, is also the ideal limit that a *zipping* software should reach, the authors used a compression algorithm to find out what kind of information could be extracted from a given text starting from what they could get to know about its linguistic entropy.

The algorithm they used was an appropriately modified version of LZ77[4], which was able to index the length and position of redundant strings within a text, and also, among other things, to extrapolate and collect them. Such modified version of LZ77 was called BCL1, thus named after the initials of the authors' names. This algorithm was sided by two more software programs (FITCH and CONSENSE)[5] created by Joe Felsenstein for inferences about phylogenies and made available online by the University of Washington. After processing the texts using both BCL1 and FITCH, Benedetto, Caglioti and Loreto obtained a graph in which the 55 languages were grouped in a way that matches to a surprising extent philological classifications of the same languages.

Although the authors had no specific interest in literature, at the end of their paper they expressed the hope that their method may be applied to a wider range of areas such as DNA and protein sequences, time sequences and literature.

The most relevant difference between the present case study and previous stylometric studies is that the latter have only dealt with the analysis of single words or sentences. More precisely, the studies which investigated single words especially focussed on features such as length, occurrence and frequency, whereas the works that dealt with phrase or sentence analysis especially studied features such as the avarage number of words in a sentence, the avarage length of sentences, etc.

These procedures have been followed in many different seminal studies which gave birth to modern stylometry: Ellegård's study about the *Junius Letters*[6], Mosteller's investigation of the *Federalist Papers*[7], Marriott's analysis of the *Historia Augusta*[8], last but not least Mendenhall's scrutiny of Shakespeare's works[9]. During the last decades of the last century, some new studies have been carried out using computer based tools. However, such studies do not differ from those dating back to 19th and 18th century, in that they use computers as fast helpers, instead of bringing about new hypotheses based on specific characteristics and potential of the computer. This is the case, for example, of Thomas Horton's[10] study about *function words* in Shakespeare's *All is True* and *Two Noble Kinsmen*.

On the contrary, the present study doesn't analyse function words or a particular class of words, nor does it simply deal with phrase or sentence analysis. The investigation here is based on the ratio of equal character strings shared by two or more texts. Moreover a character string cannot be identified with a single word or phrase in that it may contain diacritical signs, punctuation, blanks and even word or phrase chunks.

A few years ago Susan Hockey clearly stated that[11], if deeply investigated, a text must show on some level its author's DNA or fingerprint. It goes without saying that the greater the number of DNA sequences common to two biological entities, the greater their phenotypical resemblance is. As a consequence it is also very likely that the two entities are in someway related.

Based on this conviction, which has nowadays become self-evident both in Biology and Genetics, the present study analyses shakespearian texts as though they were mere DNA strings. Then the texts are automatically grouped into families and placed on a phylogenetic tree, so as to account both for their evolution and for their deepest similarities.

Results are surprisingly precise and the families thus created confirm the groupings which textual critics and phylologists have agreed on over the last few centuries. Other interesting similarities have also been brought to light. The algorithms, for instance, have been perfectly able to recognise and group on a single branch of the phylogenetic tree Shakespeare's so called Roman Plays, which share the same setting, some themes and a number of characters. The system also grouped together the Historical plays, works whose similarities have also been acknowledged by textual and literary criticism. Furthermore the experiment has pointed out a liguistic similarity between

the tragedy of Romeo and Juliet and some of Shakespeare's comedies. Although such similarity has never been studied in depth and certainly deserves further investigation, it would not come across as unlikely or bizarre to the shakespearian scholar.

Experiments have also been carried out to test the validity of the algorithms on shorter texts. In this second phase of the study, the complete *corpus* of Shakespeare's works was split into 16 KiB text chunks. This time the system was able to create 37 subsets of chunks, each of which coincided with a play and appropriately placed each subset on the phylogenetic tree.

In a final phase of the experiment the effectiveness of the algorithms on authorship was tested against a corpus of 1200 modern English texts by known authors with positive results and it was then applied to the play *All is True* to discriminate between Shakespere's and Fletcher's authorship vis-à-vis single parts of the play.

Whereas the results achieved during the testing phase were completely successful (100% of correct attributions) when dealing with Shakespeare's All is True they were a little less satisfactory (about 90%). Two factors may account for this: on the one hand this may be due to the fluidity of English morphology in that period; on the other hand this specific text may have suffered the intervention of other authors as a few critics have suggested during the last century.

Experiments are still being carried out to refine the procedure and make the algorithms produce better performances.

Notes

1 A phylogenetic tree is a graph used in biology and genetics to represent the profound relationship (e.g. in the DNA), between two phenotypically similar entities which belong to the same species or group.

2 D. Benedetto, E. Caglioti (department of Mathematics), V. Loreto (Department of Physics)

3 A.N. Kolmogorov, Probl. Inf. Transm. 1, 1(1965) and G.J. Chaitin, *Information Randomness and Incompleteness* (WorldScientific, Singapore, 1990), 2nd ed.

4 LZ77 is a lossless data compression algorithm published in a paper by Abraham Lempel and Jacob Ziv in 1977.

5 Both programs are based on algorithms used to build phylogenetic trees and are contained in a software package called PHYLIPS.

6 A. Ellegård, *A Statistical Method for determining Autorship: The Junius Letters 1769-1772*, Gothenburg, Gothenburg University, 1962

7 F. Mosteller, D. Wallace, *Inference and Disputed Authorship: The Federalist*, Reading (Mass.), Addison-Wesley, 1964

8 I. Marriot, "The Authorship of the Historia Augusta: Two Computer Studies", *Journal of Roman Studies,* 69, pagg. 65-77

9 T. C. Mendenhall, "A Mechanical Solution of a Literary Problem" , *The Popular Science Monthly*, 60, pagg. 97-105

10 *The Effectiveness of the Stylometry of Function Words in Discriminating Between Shakespeare and Fletcher*, Edinburgh, Department of Computer Science, 1987. This text can be found online at: http://www.shu.ac.uk/emls/iemls/shaksper/files/STYLOMET%20FLETCHER.txt

11 Hockey S., *Electronic Texts in the Humanities*, New York, Oxford University Press, 2000

XSLT (2.0) handbook for multiple hierarchies processing

Elena Pierazzo
elena.pierazzo@kcl.ac.uk
King's College London, UK

Raffaele Viglianti
raffaele.viglianti@kcl.ac.uk
King's College London, UK

In transcribing and encoding texts in XML, it is often (if not always) the case that structures and features do not nest within each other but overlap. Paraphrasing a notorious quotation from Paul Maas on the most unsolvable problem in editing, one may say that "there is no remedy for multiple hierarchies"[1]. Every year a considerable number of new papers and posters about how to deal with multiple hierarchies and overlapping structures are presented to conferences, and the new version of the TEI includes a chapter on similar matters[2]. And yet, as the liveliness of debate shows, a convenient, standard and sharable solution is still to be found.

Any kind of text potentially (and actually) contains multiple hierarchies, such as verses and syntax, or speeches and verses. Perhaps the most extreme form of this problem arises when transcribing a manuscript, assuming that the transcriber wants to describe both the content and also the physical structure and characteristics at the same time. Pages, columns and line-breaks conflict with paragraphs and other structural and non structural divisions such as quotation and reported speeches, as well as deletions, additions, and changes in scribal hand.

At present three different approaches to this problem have been proposed:

1. non-XML notation (for instance LMNL in Cowan, Tennison and Piez 2006) to be processed by specific tools which must be developed in-house;

2. pseudo-XML, such as colored XML or concurrent XML;[3]

3. full XML approaches such as milestones, stand-off markup[4], or fragmentation.

All of these approaches depend on post-processing tools. Some of these tools have been developed using scripting or other languages such as perl, Python, or Java, and others use XSLT-based approaches.

The milestone approach is the one chosen by the TEI, and, consequently, by us. Nevertheless, milestones introduce greater levels of complexity when building (X)HTML output and for this reason they might not be used in practice.

As XSLT is the most common technology used to output XML encoded texts, at CCH we experimented with different ways to deal with milestones and have developed a handful of XSLT techniques (as opposed to tools) that can be easily adapted to different circumstances.

Clashing of two hierarchies

Let us consider, for example, the TEI P5 empty element <handShift/> that delimits a change of scribal hand in a manuscript. Yet problems arise if one needs to visualize both the paragraphs and the different hands with different formatting. In case of XHTML visualization, one would almost certainly want to transform the <handShift> from an empty element to a container, but this container could well then overlap with existing block or inline elements such as <p>.

The easiest way to deal with this is to use the "disable-output-escaping" technique by outputting HTML elements as text; for instance:

```
<xsl:template match="tei:handShift">
    <xsl:text disable-output-
escaping="yes">&lt;span style="color:
red;"&gt;</xsl:text>
</xsl:template>
<xsl:template match="tei:anchor[@
type='end-handShift']">
    <xsl:text disable-output-
escaping="yes">&lt;/span&gt;</xsl:text>
</xsl:template>
```

However, this solution presents the obvious disadvantage that the output will not be well structured (X)HTML, and although browsers are often forgiving and therefore may cope with this in practice, this forgiveness cannot be relied on and so this process cannot be recommended.

A better option is to transform the area delimited by two <handShift>s in a container but fragmenting it to avoid overlapping.

One possible XSLT algorithm to expand and to fragment an empty element could be:

- Loop on the <handShift>s
- Determine the ancestor <p>

```
<xsl:variable name="cur-p"
select="generate-id(ancestor::p)"/>
```

- Determine the next <handShift>

```
<xsl:variable name="next-
hs" select="generate-
id(following::handShift)"/>
```

- Create a new non-empty element <handShift>

• Loop on all nodes after <handShift/>, up to but not including either the next <handShift/> or the end of the current <p>. This can be achieved using an XPath expression similar to the following:

```
following::*[ancestor::
p[generate-id()=$cur-p]]
    [not(preceding::handShift[generate-
id()=$next-hs])]
    |
following::text()[ancestor::
p[generate-id()=$cur-p]]
    [not(preceding::handShift[generate-
id()=$next-hs])]
```

That will return:

```
<p> … <handShift> … </handShift></p>
<p><handShift> … </handShift> … </p>
```

This resulting intermediate XML could then be used to produce XHTML that would fulfil the visualization required. However this would involve another XSL transformation and the intermediate file would not be valid against the schema, unless an ad hoc customized schema is generated for that purpose.

Nevertheless, thanks to XSLT 2 it is possible to produce a single process outputting a first transformation into a variable and then apply other templates on the variable itself using the mode attribute, thus dividing the process into steps and avoiding both external non-valid files and also modifications to the schema.

This is an example using the mode attribute.

• Declaration of variables

```
<xsl:variable name="step1">
        <xsl:call-template name="one"/>
</xsl:variable>

<xsl:variable name="step2">
        <xsl:apply-templates
select="$step1" mode="step2"/>
</xsl:variable>
```

• XML to XML transformation (Step 1)

Copying the whole XML text:

```
<xsl:template match="*" mode="step1">
     <xsl:copy>...</xsl:copy>
</xsl:template>
```

Other templates (as the ones described above) to transform <handShift/>:

```
<xsl:template match="handShift"
mode="step1">
[...]
</xsl:template>
```

Saving the elaborated file into the declared variable:

```
<xsl:template name="one" mode="step1">
        <xsl:apply-templates
select="TEI" mode="step1"/>
</xsl:template>
```

• XHTML transformation (Step 2)

```
<xsl:template match="/" mode="step2">
     <html>...</html>
</xsl:template>
```

Other templates to transform <handShift> and <p> in XHTML elements:

```
<xsl:template match="handShift"
mode="step2">
     <span class="hand">...</span>
</xsl:template>
```

• Output

```
<xsl:template match="/">
        <xsl:copy-of select="$step2"/>
</xsl:template>
```

The poster will include a comparison of the performances of the XSLT 2.0 algorithm with a sequence of XSLT 1.0 transformations.

Clashing of more than two hierarchies

It is not improbable that in complex texts such as manuscripts more than two hierarchies clash. Consequently, the difficulties of visualization in XHTML can become more complex.

During the analysis for a CCH project devoted to the digital edition of Jane Austen's manuscripts of fictional texts, the need emerged to mark up lines as block elements in order to manage them via a CSS stylesheet.

In TEI P5 lines are marked by the <lb/> empty element, and so it was necessary to transform these into containers. Therefore at least three hierarchies were clashing: <handShift/>, <lb/> and <p>.

A good way to handle the conflict could be looping on text nodes between milestones. In the following example, all the text nodes between <handShift>s are expanded into container elements and then transformed into elements carrying a class attribute. Moreover all the lines are transformed into further s using the algorithm mentioned before in order to manage them as block elements.

The following templates show a possible implementation of this method.

Step 1 XML to XML:

```
<xsl:template match="text()
[not(ancestor::teiHeader)]"
mode="step1">
<handShift>
<xsl:copy-of select="preceding::
handShift[1]/@new"/>
<xsl:value-of select="."/>
</handShift>
</xsl:template>
```

Step 2 XML to XHTML:

```
<xsl:template match="handShift">
     <span class="@new">
<xsl:apply-templates mode="step2"/>
</span>
</xsl:template>
```

Such a solution is also applicable with more than two clashing hierarchies.

Even though this approach can be applied generically, a deep analysis of the needs for representation and visualization is required in order to develop more customized features. For instance, the need to show lines as block elements has caused other hierarchical clashes that have been resolved using customized applications of the algorithms explained above. According to project requirements, in fact, if the apparently innocuous element <lb/> is used to produce non-empty elements in the output, any TEI element at a phrasal level is potentially overlapping and requires a special treatment.

The poster may be seen as providing an XSLT Cookbook for multiple hierarchies (the XSLT code will be available as just such a cookbook from the TEI wiki.) In our opinion simple recipes are better for encoding multiple hierarchies than a tool is, even a customizable one. The flexibility and the extensibility of the TEI encoding schema allows for an almost infinite combination of elements, attributes and values according to the different needs of each text. Since the first release of the TEI Guidelines, the Digital Humanities community has learnt to enjoy the flexibility of SGML/XML based text encoding, but such advantages come with a price, such as the difficulty of creating generic tools able to accommodate the specific needs of every single project.

Furthermore, even assuming that a finite combination of elements, attributes and values could be predicted at input (considerably limiting the possibilities offered by the TEI schema), the potential outputs are still infinite. This is why the most successful technology for processing text encoded in XML is either an equally flexible language – XSLT – or tools that are based on such a language but that still require a high degree of customization.

Therefore, sharing methodologies and approaches within the community, though disappointing for those looking for out-of-the-box solutions, is perhaps the most fruitful line of development in the field of multiple hierarchies.

Notes

1 "Gegen Kontamination ist kein Kraut gewachsen", in Maas 1927.

2 "Non-hierarchical structures" in Burnard and Bauman 2007 at http://www.tei-c.org/release/doc/tei-p5-doc/en/html/NH.html. Something of the kind for SGML was also in TEI P3.

3 See Sperberg McQueen 2007 for an overview.

4 Burnard and Bauman 2008, at http://www.tei-c.org/release/doc/tei-p5-doc/en/html/SA.html#SASO

References

Lou Burnard and Syd Bauman (2007) *TEI P5: Guidelines for Electronic Text Encoding and Interchange*, available at http://www.tei-c.org/release/doc/tei-p5-doc/en/html/index.html (11/12/07).

Paul Maas (1927), *Textkritik*, Leipzig.

Alexander Czmiel (2004) *XML for Overlapping Structures (XfOS) using a non XML Data Model*, available at http://www.hum.gu.se/allcach2004/AP/html/prop104.html#en (11/16/07)

John Cowan, Jeni Tennison and Wendell Piez (2006), LMNL Update. In "Extreme Markup 2006" (slides available at http://www.idealliance.org/papers/extreme/proceedings/html/2006/Cowan01/EML2006Cowan01.html or at http://lmnl.net) (11/16/07)

Patrick Durusau and Matthew Brook O'Donnell (2004) *Tabling the Overlap Discussion*, available at http://www.idealliance.org/papers/extreme/proceedings/html/2004/Durusau01/EML2004Durusau01.html (11/16/07)

Michael Sperberg McQueen (2007), Representation of Overlapping Structures. In *Proceedings of Extreme Markup* 2007 (available at http://www.idealliance.org/papers/extreme/proceedings/xslfo-pdf/2007/SperbergMcQueen01/EML2007SperbergMcQueen01.pdf) (15/03/2008)

The Russian Folk Religious Imagination

Jeanmarie Rouhier-Willoughby
j.rouhier@uky.edu
University of Kentucky, USA

Mark Lauersdorf
mrlaue2@email.uky.edu
University of Kentucky, USA

Dorothy Porter
dporter@uky.edu
University of Kentucky, USA

The "folk" and intellectual views of folk material play a profound role in Russian cultural history. The German Romantics argued that the folk conveyed the true essence of a nation's identity, a stance adopted in nineteenth century Russia. Beginning in the late 1930s, the Soviets held up the folk as heroes who conveyed the ideals of the socialist state. Both of these conceptions persist until the present day and present an attitude that differs significantly from most contemporary civil societies, where the folk are typically viewed as poor, backward people that need to be enlightened. Throughout 19th- and 20th-centuries, folk culture and "high" culture have come together in Russia in a way unknown in most European societies. Thus, it can be argued that in order to understand Russia, one must study the conception of the "folk" and their belief systems. Russian religion is no exception in this regard; to study Russian religious belief without a consideration of the folk conceptions is to overlook one of the most important sources for Russian religious ideas.

Russian Orthodoxy has been the source of a great deal of speculation about the extent of dvoeverie (dual faith). Soviet scholars argued that the Russian Orthodox religion had a strong pre-Christian base, which allowed the Soviet government to assert that religious folk tradition was actually not "religious" at all and thereby should be preserved despite the atheist policies of the state. Since the fall of the Soviet Union, scholars have undertaken the study of folk religion in earnest, but there is as yet no comprehensive study of the interrelations between various folk genres in relation to religious belief. Typically folklorists study either oral literature (e.g., legends and songs), or folk ritual and iconography. This separation of genres inhibits a full understanding of the complexity of the complete religious belief system. Our multimedia web-based critical edition, the Russian Folk Religious Imagination (RFRI), will feature an innovative cross-disciplinary approach combining the study of legends on saints and biblical figures, songs and religious rituals, and folk iconography into a single, comprehensive research project that will be published in a new digital framework designed to integrate text and multimedia into a coherent whole (http://www.rch.uky.edu/RFRI/). We are using the AXE annotation tool (created by Doug Reside at the Maryland Institute for Technology in the Humanities: http://www.mith2.umd.edu/) for encoding commentary on audio, video, images and textual materials (for an example of video annotation, see here: http://www.rch.uky.edu/RFRI/AXE-example.html). This far-reaching project will be of use to specialists in a wide range of disciplines including historians, folklorists, anthropologists, linguists, and scholars of literature and of religion, as well as to amateur historians and to the general public. Our poster presentation will showcase the achievements of the project thus far and provide demonstrations of the variety of material and techniques we are using in the development of this project.

Despite the Soviet government's tolerance of scholarly fieldwork gathering folk religious traditions, there is a paucity of in-depth research and publication in this area, due to the official policy of atheism in the Soviet Union. Folklorists collecting data on this topic often could not publish their findings, and the material has languished in archives and private collections. Even when published editions did exist, they quickly went out of print, so that specialists elsewhere were also unable to do extensive research on Russian folk religion. Our edition will provide unprecedented access for scholars and students of folklore and of Russia to materials they could not previously obtain without extensive archival work. Scholars will be able to read currently inaccessible texts, access audio and video files of song and legend texts and rituals, and consult still images of folk iconography and rituals. Users will be able to search, compare, and study in full context the entire range of text, image, audio and video materials in a way that would not be possible in a print edition. This approach not only provides a greater breadth of materials (typically only available in Russian archives) for both scholars and students, but also brings to the fore the advantages of using digital resources in the humanities. In addition, the RFRI project will serve as broad an audience as possible by providing Russian originals and English translations of both the original texts and the scholarly commentary and textual analyses. The project resulting from this expertise will significantly increase the knowledge of and scholarly interest in Russian folk belief and religion. A digital multimedia critical edition, as we have conceived it, will not only make use of the latest in digital technology, but will also feature a combination of technologies that is truly cutting-edge. Our methods and design will be a model for other scholars to follow in developing fully integrated multimedia research projects using open-source software and internationally accepted standards.

Using and Extending FRBR for the *Digital Library for the Enlightenment and the Romantic Period – The Spanish Novel (DLER-SN)*

Ana Rueda
rueda@email.uky.edu
University of Kentucky, USA

Mark Richard Lauersdorf
mrlaue2@email.uky.edu
University of Kentucky, USA

Dorothy Carr Porter
dporter@uky.edu
University of Kentucky, USA

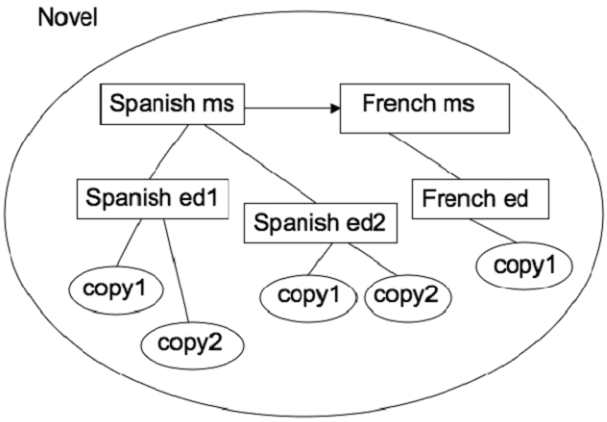
Spanish literature has been largely overlooked in the development of the canon of European fiction of the Enlightenment, and literary criticism has traditionally treated the eighteenth-century and early nineteenth-century Spanish novel with indifference, when not with hostility. While France, England, and Germany have produced print catalogs of their novelistic production for this period, Spain still lacks one. Manuals of literature constantly remind us of the void in Peninsular Spanish literature in this genre, a sentiment which has become a commonplace in literary scholarship. Montesinos, one of the better informed scholars in the field, sees nothing but sterility in the Spanish eighteenth-century novel of this "lamentable período" (*Introducción a una historia de la novela en España, en el siglo XIX*, 40), and Chandler, like many literary investigators of this period, concludes that literary efforts were reduced to servile imitations of the masters of a preceding age (*A New History of Spanish Literature,* 497). Such perceptions are changing as the Spanish Enlightenment is now becoming an important site of critical inquiry with a growing body of scholarship. Spain produced a national body of fiction that was far superior in volume to the few acclaimed fictional accounts and enjoyed great popularity among the general readership of its day.

Extra-literary factors hindered but also shaped the Spanish novel between 1700 and 1850. Given the extraordinary popularity of novels through this period over the whole of Europe and the fact that Spain closely followed French fashions after Felipe V became the first king of the Bourbon dynasty, it is not surprising that Spain translated and adapted foreign novels to its own geography, customs, and culture. The production of Spanish novels picked up considerably after 1785, propelled by the adaptation of European models published in translation. But the country's novelistic output suffered from a major set-back: religious and government censorship, which was preoccupied with the moral aspect of novel writing. Although translations of foreign novels were widely read in Spain, the erudite accused them of contributing to the corruption of the Spanish language. Further, the novel as a genre disquieted censors and moralists of the age. At best, the novel was condemned as frivolous entertainment, devoid of "useful" purpose, to the extent that Charles IV issued a decree forbidding the publication of novels in 1799. The decree, however, was not consistently enforced, but it did affect literary production. Given the dual censorship – religious and civil – existing in Spain at the time, many novels remained in manuscript form. Spain's publishing and reading practices remained largely anchored in the *ancien régime* until the 1830's, a decade that witnessed social transformations and innovations in printing technology which dramatically altered the modes of production and consumption in the literary marketplace.

The Digital Library for the Enlightenment and the Romantic Period – The Spanish Novel (DLER-SN), an international collaborative project of scholars at the University of Kentucky and the Consejo Superior de Investigaciones Científicas in Madrid, Spain, is creating an online resource center for the study of the Spanish novel of the Enlightenment that extends into the Romantic Period (1700-1850). This scholarly collection will reconstruct the canon for the eighteenth- and nineteenth-century novel in Spain and, by joining bibliographic and literary materials with an extensive critical apparatus, it will constitute a quality reference work and research tool in the field of Eighteenth- and Nineteenth-Century Studies. The overall objective is to foster new research on the significance of these Spanish novels by providing online open access to a collection of descriptive bibliographic materials, critical studies, and searchable full-text editions. The proposed poster will outline the *DLER-SN* project and showcase the first portion of it: the development and implementation of the bibliographic database that will serve as the backbone for the complete digital collection.

To organize the *DLER-SN Database*, we turn to the Functional Requirements for Bibliographic Records (FRBR), an entity-relationship model for describing the bibliographic universe, developed by the International Federation of Library Associations and Institutions (IFLA). FRBR allows us not only to describe the individual novelistic works, but also to place them within the context of the collection (i.e., within their full eighteenth and nineteenth-century literary context). In this way it is different from every other bibliographic resource of the period of which we are aware.

A single novel may exist in one manuscript, written in Spanish; a second manuscript that is a translation of the Spanish into French; two separate printed editions of the Spanish; one printed edition of the French; and several individual copies of each edition, found in different libraries in the US and Europe.



Entity: *Work*

In the FRBR model, the text of the original novel would be identified as the *work*. The FRBR guidelines are clear that the concept of *work* is in fact separate from any physical object. Although in our example the original novel is synonymous with the Spanish manuscript, it is the story/ideas/text written in the manuscript and not the manuscript itself that is the *work*.

Entity: *Expression*

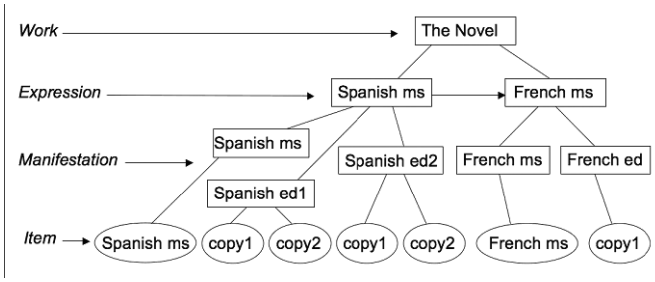
The manuscripts, both Spanish and French, are two *expressions* of the same *work*. These *expressions* have relationships among each other and with the *work*. In our example, both *expressions* represent the same *work*, and they also relate to each other: the French is a translation of the Spanish.

Entity: *Manifestation*

The editions – the actual print runs – are *manifestations* of the *expressions* on which they are based. In the case of manuscripts, the *manifestation* is the manuscript itself.

Entity: *Item*

The individual copies of the editions, as found in libraries and private collections, are *items* which exemplify the *manifestations*. Variations may occur from one *item* to another, even when the *items* exemplify the same *manifestation*, where those variations are the result of actions external to the intent of the producer of the *manifestation* (e.g., marginal notation, damage occurring after the item was produced, [re]binding performed by a library, etc.). In the case of manuscripts, the *item* is the same as the *manifestation* (i.e., the *item* is the manuscript itself).



The application of the FRBR model to the *DLER-SN Database* is fairly straightforward, however there are three issues that are specific to the Spanish novel that involve extensions to the model.

New *Work* **vs.** *Expression*

In the FRBR model there is no clear dividing line between assigning a modified text as a new *work* or as an *expression* of an existing *work*. To illustrate this problem, let's add an English version of our novel to the mix, since multiple versions of stories written in different languages are common occurrences during our period of investigation of the Spanish novel. Our hypothetical English version clearly tells the same story as both the Spanish and French *expressions*, but it is clearly not a translation of either of them. Perhaps some characters are left out and others are added, scenes are rearranged, etc. The English version could be considered a new *work* or another *expression* of the original *work* depending on the contextual relationship we wish to emphasize. In order to emphasize that there is a clear thematic relationship between the English version and the Spanish and French versions (i.e., it tells much of the same story), it could be reasonable for us to say that, in this case and for our purposes, the English version is an expression of the same *work* rather than a new *work*. However, if in assembling our FRBR-oriented database we were to choose to call the English version a new *work*, we would then, perhaps, wish to experiment with the creation of a more abstract category above *work* to illustrate that there is some relationship (in character archetypes, leitmotifs, etc.) between the different *works*.

"Masked" *Works* **and** *Expressions*

In the *DLER-SN* collection we have identified two types of "masked" *works* and *expressions*.

1. An author claims his product is a translation of another text, but it is in fact an original product (the author is claiming an *expression* of an existing *work*, but it is actually a new *work*) – this is a "masked" *work*.

2. An author claims his product is an original when it is in fact a translation of an existing text (the author is claiming a new *work*, but it is actually an *expression* of an existing *work*) – this is a "masked" *expression*.

We will need to make decisions about how to identify these products within our FRBR-based system. It may be most reasonable to identify such texts as both *work* and *expression* (there is nothing in the guidelines that disallows such multiple designations); however, if so, we will need to identify them appropriately as "masked".

Censorship & Circumvention (Its Effect on *Work*, *Expression*, *Manifestation*, and *Item*)

In some cases censorship will be an issue – i.e., changes made to texts to satisfy the requirements of the government and the church can potentially have an effect on any level of the hierarchy. If this is the case, we will need to determine a method for identifying these changes and describing them, perhaps identifying "types" of modifications:

- title changes
- prologues and prologue variations
- publication outside Spain
- personal (unofficial) reproduction and distribution

We also anticipate that there may have been illegal print runs made of uncensored texts. If so, we will want to identify these uncensored runs with a special designation in the system. An especially interesting case would be a text that appeared in both censored and uncensored printings.

The database will be an important educational tool as well because it discloses a body of fiction severely understudied (and undertaught) due to lack of documentation and inaccessibility. The *DLER-SN Database* will help scholars reassess the need for modern editions and for new studies of these forgotten texts which, nonetheless, constituted the canon of popular culture of their time. It is important to note that only a dozen of these novels enjoy modern editions. Once completed, our project will give scholars in multiple disciplines the tools necessary to begin serious work in this dynamic but underworked field.

Bibliography

Reginald F. Brown. *La novela española,* 1700-1850 (1953).

Richard E. Chandler and Kessel Schwartz, *A New History of Spanish Literature*. Revised Edition. Baton Rouge: Louisiana State University Press (1991)

José Ignacio Ferreras. *Catálogo de novelas y novelistas españoles del siglo XIX* (1979).

Functional Requirements for Bibliographic Records Final Report, IFLA Study Group on the FRBR. UBCIM Publications – New Series Vol 19. (http://www.ifla.org/VII/s13/frbr/frbr.pdf)

Dionisio Hidalgo. *Diccionario general de bibliografía española* (1867).

José Montesinos. *Introducción a una historia de la novela en España en el siglo XVIII. Seguida de un esbozo de una bibliografía española de traducciones de novelas 1800-1850* (1982).

Antonio Palau y Dulcet. *Manual del librero hispano-americano* (1923-1927).

Ángel González Palencia. *Estudio histórico sobre la censura gubernativa en España, 1800-1833* (1970, 1934)

Francisco Aguilar Piñal. *Bibliografía de autores españoles del siglo XVIII* (1981-1991).

How to convert paper archives into a digital data base? Problems and solutions in the case of the Morphology Archives of Finnish Dialects

Mari Siiroinen
mari.siiroinen@helsinki.fi
University of Helsinki, Finland

Mikko Virtanen
mikko.virtanen@helsinki.fi
University of Helsinki, Finland

Tatiana Stepanova
tatjana.stepanova@helsinki.fi
University of Helsinki, Finland

This poster presentation and computer demonstration deals with some of the problems encountered and solutions found when converting paper archives containing linguistic data into a digital database.

The Morphology Archives of Finnish Dialects

The paper archives referred to are the *Morphology Archives of Finnish Dialects,* which contain about 500,000 paper file cards classified and arranged on the basis of almost a thousand linguistically based catalogue codes. The data the file cards contain are derived from spontaneous dialectal speech by linguistically trained field-workers. The data has been analysed and encoded to determine, for example, types of inflection and word formation and their use in the sentence, sound changes in word stems, and particles and their related uses.

The data gathering was accomplished during a 30-year period (from the 1960s to the 1990s).

The paper archives are located in the *Department of Finnish Language and Literature* in the *University of Helsinki*. There are six copies of the data in the archives in six differ-ent universities and research institutions in Finland, Sweden and Norway. *The Morphology Archives of Finnish Dialects* are closely related to two other archives of Finnish dialects, namely the *Lexical Archive of Finnish Dialects (Research Institute for the Languages of Finland)* and *Syntax Archives of Finnish Dialects (University of Turku)*.

The purpose of the *Morphology Archives of Finnish Dialects* has been to facilitate research on the rich morphology of Finnish and to provide researchers with well-organised data on the dialects of different parishes. The Archives cover all the Finnish dialects quite well since it consists of 159 parish collections equally distributed among Finnish dialects. The archive collections have served as data sources for over 300 printed publications or theses.

For additional information about the Archives, see www.helsinki.fi/hum/skl/english/research/ma.htm.

The Digital Morphology Archives of Finnish Dialects

Plans to digitize the data in the Archives were first made in the 1990s. The digitization project finally got the funding in 2001. A project to create a digital database of Finnish dialects *(Digital Morphology Archives, DMA)* was then launched. The project was funded by the *Academy of Finland,* and it ended in 2005.

The digitization was not implemented simply by copying the paper archives, but the objective has been to create an independent digital archive, which also contains data not included in the paper archives, in particular to ensure sufficient regional representation.

The *Digital Morphology Archives* currently contain 138,000 clauses in context (around one million words) from 145 parish dialects of Finnish. So far a total of 497,000 morphological codes have been added to the dialectal clauses (approx. 4 codes for each clause). In the parish collections, which are coded thoroughly, each example has been assigned from 5 to 10 codes. This increase in the number of codes will improve the possibilities of using the DMA for research purposes. The *Digital Morphology Archives* are unique in that all the data is derived from spoken language.

The database was implemented using MySQL, while the search system is built on HTML. The data are stored in the *Finnish IT Centre for Science (CSC)* (http://www.csc.fi/) and has been accessible in current form via the internet to licensed users since 2005. Licences are granted to students and researchers upon request.

An internet search facility developed jointly with the *Language Bank of Finland (CSC)* allows quick and straightforward searches both from the entire material and from individual parishes or dialect areas. Searches can also be made directly from the dialect texts. These word and phrase searches can also be targeted at dialect texts without diacritic marks. Searches can also be refined by limiting them to certain linguistic categories according to a morphological classification containing 897 codes.

For additional information about the Digital Morphology Archives, see www.csc.fi/english/research/software/dma

Licence application form:

www.csc.fi/english/customers/university/useraccounts/languagebank/?searchterm=language%20bank

Problems and Solutions

One of the problems encountered during the process has been that digitizing the data manually is very slow. In fact, the data in the digital data base still only cover about 5.5% of the paper archives. Scanning of the paper file cards has been proposed as a solution. The new challenge then would be to find a powerful enough OCR program, as the paper cards have mostly been written by hand.

Another problem has been the presentation of Finno-Ugric phonetic transcription, which includes symbols and diacritics that are not part of the ISO-Latin-1 character-set. As Unicode is not yet supported by all programs, characters in the ISO-Latin-1 character set were chosen to replace some of the Finno-Ugric phonetic symbols.

The second phase of digitization was launched in September 2007, with new funding. By the end of 2009, it is estimated that at least 250,000 new dialectal clauses with linguistic coding will have been added to the digital archive.

A Bibliographic Utility for Digital Humanities Projects

James Stout
James_Stout@brown.edu
Brown University, USA

Clifford Wulfman
Clifford_Wulfman@brown.edu
Brown University, USA

Elli Mylonas
Elli_Mylonas@brown.edu
Brown University, USA

Introduction

Many, if not most, digital humanities projects have a bibliographical component. For some projects the collection, annotation and dissemination of bibliographical information on a specific topic is the primary focus. The majority, however, include bibliography as a resource. Over the years, our group has worked on several projects that have included a bibliography. In each case, we were given bibliographical information to include at the outset, which the scholar working on the project would edit and amplify over time.

In the past we wrote several one-off bibliographic management systems, each for a particular project. Each tool was tailored to and also limited by the needs of its project. Over time we ended up with several bibliographic components, each slightly different in the way it was implemented and the way it handled data formats. We decided to settle upon a general purpose tool, in order to avoid writing any further single use applications, which were difficult and time consuming to support.

We were primarily interested in a tool that could handle many forms of bibliographic entity, from conventional scholarly materials, serials and manuscripts to multimedia and websites. It had to allow scholars to interact with it easily using a web interface for single record input and editing, and also be capable of loading large numbers of prepared records at once. We were less concerned with output and text-formatting capabilities, given the ability to produce structured XML output. We expected to allow the individual projects to query the system and format the output from the bibliographic tool.

Bibliographical management systems, although tedious to implement, are a well-understood problem, and we originally hoped to use a pre-existing tool. Upon surveying the field, we realized that the available tools were primarily single-user citation managers, like EndNote, Bibtex, and Zotero. Our users were compiling and manipulating scholarly bibliographies, so we wanted something that was familiar to academics. However, we also wanted the flexibility and overall standards compatibility of the library tools.

A great deal of the difficulty of bibliographic management arises out of the complexity of bibliographic references and the problem of representing them. Many libraries use a form of MARC to identify the different pieces of information about a book. MARC records are not ideal for storing and displaying scholarly bibliography; they aren't meant to handle parts of publications, and they don't differentiate explicitly among genres of publication. Many personal bibliographical software applications store their information using the RIS format, which provides appropriate categories of information for most bibliographic genres (e.g. book, article, manuscript, etc.). Hoenicka points out the flaws in the underlying representation of RIS, but agrees that it does a good job of collecting appropriate citation information [Hoenicka2007a, Hoenicka2007b]. We wanted to be able to use the RIS categories, but not the RIS structures. We also hoped that we would be able to store our bibliographic information in XML, as we felt that it was a more versatile and appropriate format for loosely structured, repeating, ordered data.

Personal bibliography tools were not appropriate for our purposes because they were desktop applications, intended to be used by a single person, and were optimized to produce print output. At the time we were researching other tools, RefDB had a very powerful engine, but no interface to speak of. WIKINDX was not easily configurable and was not easy to embed into other projects. Both RefDB and WIKINDX have developed features over the last year, but we feel that the system we developed is different enough in ways that render it more versatile and easier to integrate into our projects. Although RefDB is the system that is closest to Biblio, it uses own SQL-based system for storing information about records, whereas we were looking for an XML based system. RefDB also allows the user to modify records in order to handle new genres of material. However, the modifications take place at the level of the user interface. As discussed below, Biblio provides full control over the way that records are being stored in the database.

Based our survey of existing tools and the needs we had identified for our projects, we decided to implement our own bibliographic utility. We made sure to have a flexible internal representation, and a usable interface for data entry and editing. We left all but the most basic display up to the project that would be using the bibliography.

Implementation

Biblio is an XForms- and XQuery- based system for managing bibliographic records using the MODS XML format. The pages in the system are served by a combination of Orbeon, an XForms implementation, and eXist, a native XML database that supports XQuery. Using a simple interface served by eXist, the user can create, edit, import, export, and delete MODS records from a central collection in the database. When the user chooses to edit a record, they are sent to an Orbeon XForms page that allows them to modify the record and save it back to the database.

By basing the backend of our system on XML, we can take advantage of its unique features. Among the most useful is the ability to easily validate records at each point changes may have occurred. Validation begins when a document is first imported or created, to ensure that no invalid document is allowed to enter the database. The document is validated against the standard MODS schema, which allows us to easily keep our system up to date with changes in the MODS format. Once a record is in the system, we must ensure that future editing maintains its integrity. Instead of simply validating upon saving a document, we use on-the-fly XForms validation to inform the user immediately of any mistakes. The notification is an unobtrusive red exclamation mark that appears next to the field containing the error.

Although validation helps prevent mistakes, it does little to protect the user from being exposed to the complexity and generality of the MODS format. Because the MODS format is designed to handle any type of bibliographic record, any single record only needs a small subset of all the available elements. Fortunately, most records can be categorized into "genres", such as "book" or "journal article". Furthermore, each of these genres will have several constant elements, such as their title.

To maximize workflow efficiency and easy of use, we have designed a highly extensible genre system that only exposes users to the options that are relevant for the genre of the record they are currently editing. The genre governs what forms the user sees on the edit page, and hence what elements they are able to insert into their MODS record. The genre definition can also set default values for elements, and can allow the user to duplicate certain elements or groups of elements (such as a <name> element holding an author). When a record is saved to the database, any unused elements are stripped away.

The genre system is also compatible with MODS records that are imported from outside the system. We simply allow the user to choose what genre best describes the imported record, which allows the system to treat the record as if it had been created by our system. The user can also select "Previously Exported" if they are importing a document that was created by our system, which will cause the appropriate genre to be automatically selected.

We have designed a simple XML format for creating genre definitions that makes it easy to add, remove, or change what genres are available for the user. The format allows the administrator to specify which elements are in a genre, what type of field should be used (including auto-complete, selection, and hidden), what the default value of a field should be, and whether the user should be allowed to duplicate a field. All of our predefined genres also use this format, so it is easy to tweak the system to fit the needs of any organization.

Finally, once users have accumulated a large set of records of various genres, we use the power of XML again to enable the user to search and sort their records on a wide and extensible

range of criteria. The administrator can easily specify new search and sort criteria for any element in the MODS schema, using the powerful XPath language.

Results

The bibliographic tool that we built provides an intelligent filter between the academic's concept of bibliography and the cataloger's concept of bibliographic record. We adopted the fundamental representation of a bibliographic entry, the MODS structure, from the library community because we want to use their tools and their modes of access. At the same time, we don't expect scholars and their students to interact directly either with MODS as an XML structure or MODS as a cataloging system, so we mediate it through a RIS-inspired set of categories that inform the user interface. The choice of an XML structure also benefits the programmer, as this makes it very easy to integrate Biblio into an existing web application or framework. The resulting bibliographic utility has an easily configurable, easy to use data entry front end, and provides a generic, standards-based, re-configurable data store for providing bibliographic information to digital projects.

Bibliography

[BibTeX] BibTeX. http://www.bibtex.org/

[Endnote] Endnote. http://www.endnote.com

[Hoenicka2007a] Marcus Hoenicka. "Deconstructing RIS (part I)" (Blog entry). http://www.mhoenicka.de/system-cgi/blog/index.php?itemid=515. Mar. 12, 2007

[Hoenicka2007b] Marcus Hoenicka. "Deconstructing RIS (part II)" (Blog entry). http://www.mhoenicka.de/system-cgi/blog/index.php?itemid=567. Apr. 23, 2007.

[MODS] MODS. http://www.loc.gov/standards/mods/

[Orbeon] Orbeon. http://www.orbeon.com/

[RefDB] RefDB. http://refdb.sourceforge.net

[RIS] RIS. http://www.refman.com/support/risformat_intro.asp

[Wikindx] WIKINDX. http://wikindx.sourceforge.net

[Zotero] Zotero http://www.zotero.org

A Digital Chronology of the Fashion, Dress and Behavior from Meiji to early Showa periods(1868-1945) in Japan

Haruko Takahashi
RXF13365@nifty.com
Osaka-Shoin Women's University, Japan

Introduction

This paper describes a digital chronology with images of fashion, dress and behavior (Hereafter called FDB) from 1868 to 1945 in Japan. This period when kimono and western style of dress contended with each other, was a very important time for Japanese clothing culture. Nevertheless there have been few timelines published of high credibility due to a lack of supporting evidence. This chronology consists of 4000 images and documents related to 8000 events of this period. It will be available on the Internet in the near future.

How this digital chronology came about

The National Museum of Ethnology Japan, the project of making a database of clothing culture in the world, which was named MCD (Minpaku Costume Database) was started in 1984. Now it is available to the public through the museum's website under the heading "CostumeDatabase" (http://www.minpaku.ac.jp).

This database consists of six sub-databases. Of these, five are reference database and one is an integrated image database of the clothes and accessories collection of The National Museum of Ethnology. The number of items is 208,000 in the database.

To make this Costume Database, many reference data have been collected and analyzed. A lot of newspaper archives from 1868 to 1945 were in this collected reference data and I thought that a digital chronology could be created by using these archives. The articles and images came from 20 different newspapers such as The Yomiuri, Asahi, Mainichi, Tokyo-nichinichi, Miyako, Kokumin, and Jiji.

The features of this digital chronology

The features of this digital chronology are summarized as following four points (fig.1).

1) The events were chosen out of newspaper archives on the basis of 140 themes. These 140 themes, detailing the acculturation of FDB for about 80 years after the Meiji

Restorationwere set up by analyzing books and papers as well as newspaper archives. They include "Society's Opinions of FDB", "Views on Health", "Views on the Body", "Dress Reform Problem and the Blending of Japanese and Western Styles", "Formal Dress", "Kimono","Makeup and Hairstyle", etc., and "Road and Lighting" (Explaining their effect on FDB). -_The book, *Acculturation of fashion culture: behavior and dress from Meiji to early Showa periods*_(Takahashi, pub. Sangensha 2005) describes the argument for selecting these 140 themes.

2) Sources are described in all respects. As a result the timeline also functions as an index tool, and most events have supporting sources attached.

3) The chronology is divided into two columns entitled "Events" and "Contemporary Conditions". These remove ambiguity in the descriptions of FDB from the general comprehensive timeline. The former, have precise dates given and therefore function in the same way as thegeneral timeline. On the other hand, "Contemporary Conditions" in the latter show clothingculture more generally and include fads which cannot be connected to exact dates.

4) Alongside the description of each year of "Contemporary Conditions", you can see images of "Scenes of Contemporary Life" "Men", "Women", "Children", and "Beautiful Women"which are typical of the time.

These images came from the illustrations of newspapers which had served the same purpose as photographs in the second half of the 19th century. These pictures were drawn to illustrate serial novels published in the papers. They portray the joy, anger, humor and pathos of the people of those days. They also represented the culture of each class, such as the life of the empoverished which photographers of those days did not capture. These illustrations were drawn to help readers better understand the text. However, if they didn't accurately portray something, then eager readers would write to the newspapers to complain. This led to a high level of accurary on the part of the artists.

The credibility of these pictures as source is discussed in the paper *Newspaper serial novel illustrations as a source of data of the Costume Image Database in the middle of the Meiji era.* The Bulletin of Japan Art Documentation Society. (Takahashi,. 2005. p.1-15).

The system of this digital chronology

Application Capabilities

This system uses Horiuchi Color's iPallet/Kumu. This application's capability is as follows.

1) It has the function which allows the user to smoothly increase the size of a detailed picture. The user can also examine fine details in the magnified picture with a drug-and-zoom function (Fig.2).

2) It supports Firefox 2 for Japanese on the Windows XP, Windows Vista, and MacOSX.

3) Users can search by title, year, and keyword.

4) The use of standard plug-ins (ex. Flash) is possible.

Display of Dates in <Contemporary Conditions>

The dates in grey on the left handside in <contemporary conditions> are approximate dates only. However, this is not the best way of representing approximate dates. Is it possible to visually express the rise and fall of fads? That's the big challenge for this system..

Conclusion

This digital chronology concretely shows the changes in FDB for about 80 years from 1868, and also faithfully details the clothing culture of those days. It would be helpful in themaking the movies and stage productions, and thus is much more than just a scientific reference text.

This research is supported by Grants-in- Aid for Scientific Research (C) from the Japan Society for the Promotion of Science (JSPS) from 2006 to 2008.

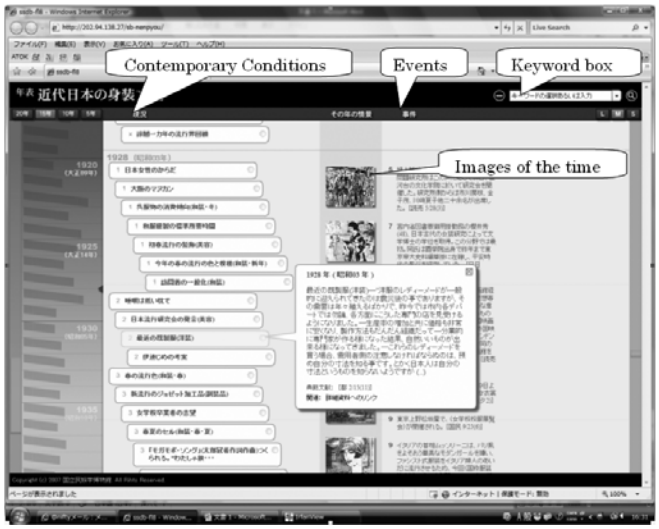


Fig.1 Example of display

Fig.2 Example of image data

Knight's Quest: A Video Game to Explore Gender and Culture in Chaucer's England

Mary L. Tripp
mtripp@mail.ucf.edu
University of Central Florida, USA

Thomas Rudy McDaniel
rudy@mail.ucf.edu
University of Central Florida, USA

Natalie Underberg
nunderbe@mail.ucf.edu
University of Central Florida, USA

Karla Kitalong
kitalong@mail.ucf.edu
University of Central Florida, USA

Steve Fiore
sfiore@ist.ucf.edu
University of Central Florida, USA

A cross-disciplinary team of researchers from the College of Arts and Humanities (English, Philosophy, and Digital Media) at the University of Central Florida are working on a virtual medieval world based on Chaucer's Canterbury Tales. The objectives are (1) that students experience a realistic virtual environment, with historical detail (buildings, music, artwork), and (2) that students view fourteenth century England from various perspectives, both as a master and the mastered, from the sacred to the profane. These objectives are accomplished through a knight's quest game based on the Wife of Bath's Tale. The Wife's tale emphasizes the battle between men's perceived authority and women's struggles for power. In her tale, a knight must go on a quest to find out what women most want. The Wife's tale leaves a narrative gap in the knight's quest, which is where our adaptation of Chaucer's work is launched.

The knight interacts with Chaucerian characters on his journey--he meets everyday people like a cook, a reeve, a miller. In order to understand the perspectives of master and mastered, sacred and profane, the knight will dress as a woman to disguise his identity; enter a cathedral to converse with a priest; disguise himself as a peasant; and, later disguise as a monk to escape danger. Other characters are scripted to react differently to the noble knight when he is in disguise. This fetishism of clothing is reminiscent of several of Shakespeare's comedies and tragedies, notably Merchant of Venice, Midsummer Night's Dream, and King Lear. There are also two narrators, the knight operating from his perspective (male and noble) and the Wife (female and a widow), interrupting and offering her opinion on the action of the quest.

The initial phase, a Game Design Document, was completed in May 2007. A complete dialogue script for the storyline is scheduled for December 2007, and initial programming and modification of the project, using the Elder Scrolls IV: Oblivion game engine to create this virtual world have also taken place through 2007. A paper prototype usability study is complete during November 2007 and a prototype visualization will be available by June 2008.

As the Canterbury Tales is traditionally taught at the high school level (Florida's curriculum includes this under 12th grade English Literature), the target audience for this prototype is 14 to 16 year old players. Eight episodes of gameplay will run an average player about two hours to complete. This format has proven suitable for classroom use, where the teacher can use the game during a computer lab session, with independent student interaction time, or as a whole-class teacher-directed activity to be completed over 3 to 4 days. The objective here is not to teach the plot of Canterbury Tales—this information is readily available from a number of different sources. Our primary focus is to immerse the user in Chaucer's historical world while educating the user on the historical facts behind the medieval era. These facts have been diluted over the years by popular fiction. An additional mission of this game is to provide the player a variety of cultural perspectives through the use of disguise and interaction with members of the Three Estates of clergy, nobility, and peasantry and the feminine estates of virgin, wife and widow. This game is not developed to "teach" students the narrative structure of the Tales. It is designed with the literary theories of New Historicism and Cultural Studies as a basis for understanding the Tales as situated in a particular history and culture.

Current research in the fields of human learning and cognitive science speak to the exceptional ability of computer games to explore identity, problem solving skills, verbal and non verbal learning, and the transfer of learned ability from one task to another (Gee, 2003; Berman & Bruckman, 2001; Cassell, 1998; Fiore, Metcalf, & McDaniel, 2007; McDaniel, 2006; Jenkins, 2006; Ryan, 2001a, 2001b; Squire, 2002). In fact, learning games in the humanities have been used for hundreds of years--Bach's Well Tempered Clavier and The Art of the Fugue are his "learning games," simple to complex musical exercises that build skill (Prensky, 2000). Virtual worlds offer a unique and interesting way in which to observe critical relationships involving race, gender, identity, community, and history. Such relationships are clearly within the territory of humanities scholarship, and we believe video games will excite and motivate students to understand and engage with these complex topics in a fashion that is intuitive and exciting for them.

In terms of humanities-specific objectives, games can provide entry points for discussions and reflections of all types of cultural and critical issues. The discussion of interactivity and the semantic quality of narrative and its application to digital media is an important aspect of computer game development. Ryan, in her article, "Beyond Myth and Metaphor—The Case of Narrative in Digital Media" concludes that computer

games offer the most suitable interface for digital narrative is compelling, especially in light of current applications in the field (2001a).

Serious games can also foster new understanding among their players. In philosophy of the mind, Gadamer describes a fusion of horizons as the way humans understand. We, as human beings, are historical and we each possess a unique horizon. Each person has his own prejudice, history, and tradition, all within the context of his language. If we approach knowledge acquisition from Gadamer's perspective, serious games are one of the best ways for students to more fully understand the humanities. Although Gadamer understands these new experiences as material interactions with the world, we propose that new understanding can take place through interactions with the virtual world as well.

Lorraine Code echoes some of the historical and cultural definitions of hermeneutical understanding that Gadamer proposes, she filters these ideas through the lens of feminism and her explanation of a mitigated definition of cultural relativism. In her essay "How to Think Globally: Stretching the Limits of Imagination," she uses Shrage's idea that relativism works mainly with 'crosscultural comparisons' to endorse her view that global understanding happens within a local context (Code, 1998). Our team proposes that serious games allow the player to construct new perceptions of global understanding from a variety of social, gendered and cultural perspectives (Berman & Bruckman, 2001; Squire, 2002). Also, the production of video games offers an important learning opportunity for students involved in the multimodal production and representation of source-specific content, narrative, and gameplay experiences.

We believe that this project, the development of a virtual medieval world of Chaucer, can challenge learners to use this path toward creating a new kind of knowledge of the humanities. This new kind of knowledge in the humanities is not the traditional memorized litany of Western icons, nor a structure of literary genres and plots, but intensive and extensive meaningful interaction with the cultural and artistic achievements of humankind. This project is really what game based learning can be at its best, meeting the needs of a changing modern paradigm of understanding and basing its development in good theoretical research.

Representative Bibliography

Berman, J. & Bruckman, A. (2001). "The Turing Game: Exploring Identity in an Online Environment." *Convergence*, 7(3), 83-102.

Cassell, Justine. (1998). "Storytelling as a Nexus of Change in the Relationship between Gender and Technology: a Feminist Approach to Software Design." *From Barbie to Mortal Kombat: Gender and Computer Games.* Cassell and Jenkins, eds. Cambridge: MIT Press.

Code, L. (1998). "How to Think Globally: Stretching the Limits of Imagination." *Hypatia*, 13(2), 73.

Fiore, S. M., Metcalf, D., & McDaniel, R. (2007). "Theoretical Foundations of Experiential Learning." In M. Silberman (Ed.), *The Experiential Learning Handbook* (pp. 33-58): John Wiley & Sons.

Gadamer H. G. (1991). *Truth and Method.* (2nd revised edition) (J. Weinsheimer & D. Marshall, Trans.). New York: Crossroad. (Original work published 1960).

Garris, R., R. Ahlers, et al. (2002). "Games, Motivation, and Learning: A Research and Practice Model." *Simulation Gaming* 33(4), 441-467.

Gee, J. P. (2003). *What Video Games Have to Teach Us About Learning and Literacy.* New York: Palgrave Macmillan.

Jenkins, H. (2006). *Convergence Culture*. New York: New York University Press.

McDaniel, R. (2006). Video Games as Text and Technology: Teaching with Multimodal Narrative. Paper presented at the 9th Annual Conference of the Association for Teachers of Technical Writing in Chicago, IL. March 22, 2006.

Prensky, M. (2001). *Digital Game-Based Learning*. New York: McGraw-Hill.

Ricci, K. S., Eduardo; Cannon-Bowers, Janis (1996). "Do Computer-Based Games Facilitate Knowledge Acquisition and Retention?" *Military Psychology* 8(4), 295-307.

Ryan, Marie-Laure. (2001a) "Beyond Myth and Metaphor—The Case of Narrative in Digital Media." *Game Studies* 1(1). 26 October 2007. <http://www.gamestudies.org/0101/ryan/>.

Ryan, Marie-Laure. (2001b) *Narrative as Virtual Reality: Immersion and Interactivity in Literature and Electronic Media.* Baltimore: Johns Hopkins University Press.

Squire, K. (2002). "Cultural Framing of Computer/Video Games." *Game Studies*, 2(1).

Tripp, M. (2007). Avatar: From Other to Self-Transcendence and Transformation. Paper presented at the Philosophy Club of Winter Park, FL, 7 August.

Underberg, N (2006). "From Cinderella to Computer Game: Traditional Narrative Meets Digital Media in the Classroom." Milwaukee, WI: American Folklore Society: 2006.

Unsworth, John. (2006). "Digital Humanities: Beyond Representation." Lecture. University of Central Florida. Orlando, FL. 13 November.

TEI by Example: Pedagogical Approaches Used in the Construction of Online Digital Humanities Tutorials

Ron Van den Branden
ron.vandenbranden@kantl.be
Centre for Scholarly Editing and Document Studies (KANTL), Belgium

Melissa Terras
m.terras@ucl.ac.uk
University College London, UK

Edward Vanhoutte
edward.vanhoutte@kantl.be
Centre for Scholarly Editing and Document Studies (KANTL), Belgium

Over the past 20 years, the TEI (Text Encoding Initiative) has developed comprehensive guidelines for scholarly text encoding (TEI, 2007). In order to expand the user base of TEI, it is important that tutorial materials are made available to scholars new to textual encoding. However, there is a paucity of stand-alone teaching materials available which support beginner's level learning of TEI. Materials which are available are not in formats which would enable tutorials to be provided in blended learning environments (Allan, 2007) such as classroom settings, for instance, as part of a University course, or allow individuals to work through graded examples in their own time: the common way of learning new computational techniques through self-directed learning.

As a result, there is an urgent need for a suite of TEI tutorials for the self directed learner. The "TEI by Example" project is currently developing a range of freely available online tutorials which will walk individuals through the different stages in marking up a document in TEI. In addition to this, the tutorials provide annotated examples of a range of texts, indicating the editorial choices that are necessary to undertake when marking up a text in TEI. Linking to real examples from projects which utilise the TEI reaffirms the advice given to learners.

The aims and focus of the project were documented in Van den Branden et al. (2007), whereas the aim of this poster is to detail the editorial, technological, and pedagogical choices the authors had to make when constructing the tutorials, to prepare stand- alone tutorials of use to the Digital Humanities audience, and beyond.

TEI by Example is effectively an implementation of problem based learning, an efficient and useful approach to teaching skills to individuals in order for them to undertake similar tasks themselves, successfully. The literature on this is wide and varied (for seminal literature regarding the effectiveness of this pedagogic approach see Norman and Schmidt (1992); Garrison (1997); and Savin-Baden and Wilkie (2006)). There has been particular consideration as to the effectiveness of example and problem based learning when learning computer programming (for example, see Mayer (1981), Mayer (1988), Kelleher and Pausch (2005)). Additionally, another wide area of academic research is how to develop online tutorial materials successfully (Stephenson, 2001; Jochems et al., 2003). Understanding the nature of online tutorials, and grappling with the pedagogical issues these technologies offer us, was a core issue when beginning to implement the TEI by Example materials.

In order to develop the TEI by Example tutorials, the team had to understand the technical possibilities and limitations afforded by the online environment, and decide how best to integrate these into the tutorial materials. By juxtaposing static (pages, articles) and dynamic (quizzes, validation) functionality, the project aims to provide a holistic learning environment for those new to the TEI. Further linking to other examples provided by the community extends the remits of the project into another, alternative viewpoint by which to start learning the TEI, aside from the TEI guidelines themselves (TEI, 2007). Additionally, the role of user testing will be explored to feature feedback and comments from the TEI user community, to aid in the development of intuitive tutorial materials.

This poster will report on progress, problems, and potential solutions in developing teaching materials for the TEI, demonstrate the tutorials developed, and highlight areas in which the TEI and Digital Humanities communities can aid in the design, and implementation, of materials for students and teachers.

References

Allan, Barbara (2007). *Blended Learning. Tools for teaching and training*. London: Facet Publishing.

Garrison, D. R. (1997). "Self Directed Learning, Towards a Comprehensive Model". *Adult Education Quarterly*, 48(1): 18-33.

Jochems, W., van Merrienboer, J. and, Koper, R. (eds.) (2003). *Integrated E-Learning: Implications for Pedagogy, Technology and Organization* (Open & Flexible Learning). London: Routledge Farmer.

Kelleher, C. and Pausch, R. (2005). "Lowering the Barriers to Programming: A Taxonomy of Programming Environments and Languages for Novice Programmers", *ACM Computing Surveys*, June 2005, 37(2): 83-137.

Mayer, R. (1981). "The Psychology of How Novices Learn Computer Programming" *ACM Computing Surveys* (CSUR), 13(1): 121 - 141.

Mayer, R. (1988). *Teaching and Learning Computer Programming: Multiple Research Perspectives*. Hillsdale, N.J: Lawrence Erlbaum Associates Inc.

Norman, G. R., and Schmidt, H. G. (1992). "The psychological basis of problem-based learning: a review of the evidence". *Acad Med,* 67(9): 557-565

Savin-Baden, M. and Wilkie, K. (2006). *Problem based learning online*. Maidenhead: Open University Press.

Stephenson, J. (ed.) (2001). *Teaching and Learning Online: Pedagogies for New Technologies*. Abingdon: Routledge.

TEI (2007). Burnard, L. and Bauman, S. (eds). "TEI P5 Guidelines for Electronic Text Encoding and Interchange" http://www.tei-c.org.uk/P5/Guidelines/index.html

Van den Branden, R., Vanhoutte, E., Terras, M. (2007). "TEI by Example". *Digital Humanities 2007, University of Illinois at Urbana-Champaign, USA, June 2007*. http://www.digitalhumanities.org/dh2007/abstracts/xhtml.xq?id=221

CLARIN: Common Language Resources and Technology Infrastructure

Martin Wynne
martin.wynne@oucs.ox.ac.uk
Oxford University

Tamás Váradi
varadi@nytud.hu
Hungarian Academy of Sciences

Peter Wittenburg
Peter.Wittenburg@mpi.nl
Max Planck Institute for Psycholinguistics

Steven Krauwer
steven.krauwer@let.uu.nl
Utrecht University

Kimmo Koskenniemi
kimmo.koskenniemi@helsinki.fi
University of Helsinki

This paper proposes the need for an infrastructure to make language resources and technology (LRT) available and readily usable to scholars of all disciplines, in particular the humanities and social sciences (HSS), and gives an overview of how the CLARIN project aims to build such an infrastructure.

Why we need a research infrastructure for language resources

Problems of standards for textual representation, interoperability of tools, and problems with licensing, access and sustainability have dogged the Humanities since the invention of the digital computer. Language resources such as text corpora exhibit a variety of forms of textual representation, metadata, annotation, and access arrangements. Tools are usually developed for ad hoc use within a particular project, or for use by one group of researchers, or for use with only one text or set of data, and are not developed sufficiently to be deployed as widely-used and sustainable services. As a result, a large amount of effort has been wasted over many years developing applications with similar functionality. Veterans of ACH and ALLC will know that the problems which are addressed by this paper are not new ones. What a persistent and sustainable infrastructure, as part of the e-Science and Cyberinfrastructure agenda, can offer is perhaps the first realistic opportunity to address these problems in a systematic, sustainable and global fashion.

The Summit on Digital Tools in the Humanities in Charlottesville, Virginia in 2006 estimated that only 6% of scholars in the Humanities go beyond general purpose information technologies (email, web browsing, word processing, spreadsheets and presentation slide software), and

suggested that revolutionary change in humanistic research is possible thanks to computational methods, but that this revolution has not yet occurred. This is an exciting time in humanities research, as the introduction of new instruments makes possible new types of research, but it is clear that new institutional structures are needed for the potential to be realised.

CLARIN is committed to boost humanities research in a multicultural and multilingual Europe, by allowing easy access and use of language resources and technology to researchers and scholars across a wide spectrum of domains in the Humanities and Social Sciences. To reach this goal, CLARIN is dedicated to establishing an active interaction with the research communities in the Humanities and Social Sciences (HSS) and to contribute to overcoming the traditional gap between the Humanities and the Language Technology communities.

The CLARIN proposal

The proposed CLARIN infrastructure is based on the belief that the days of pencil-and-paper research are numbered, even in the humanities. Computer-aided language processing is already used by a wide variety of sub-disciplines in the humanities and social sciences, addressing one or more of the multiple roles language plays, as carrier of cultural content and knowledge, instrument of communication, component of identity and object of study. Current methods and objectives in these disparate fields have a lot in common with each other. However it is evident that to reach the higher levels of analysis of texts that non-linguist scholars are typically interested in, such as their semantic and pragmatic dimensions, requires an effort of a scale that no single scholar could, or indeed, should afford.

The cost of collecting, digitising and annotating large text or speech corpora, dictionaries or language descriptions is huge in terms of time and money, and the creation of tools to manipulate these language data is very demanding in terms of skills and expertise, especially if one wants to make them accessible to professionals who are not experts in linguistics or language technology. The benefits of computer enhanced language processing become available only when a critical mass of coordinated effort is invested in building an enabling infrastructure, which can then provide services in the form of provision of all the existing tools and resources as well as training and advice across a wide span of domains. Making resources and tools easily accessible and usable is the mission of the CLARIN infrastructure initiative.

The purpose of the infrastructure is to offer persistent services that are secure and provide easy access to language processing resources. Our vision is to make available in usable formats both the resources for processing language and the data to be processed, in such a way that the tasks can be run over a distributed network from the user's desktop. The CLARIN objective is to make this vision a reality: repositories of data with standardized descriptions, language processing tools which can operate on standardized data, with a framework for th resolution of legal and access issues, and all of this available on the internet using Grid architecture.

The nature of the project is therefore primarily to turn existing, fragmented technology and resources into accessible and stable services that any user can share or customize for their own applications. This will be a new underpinning for advanced research in the humanities and social sciences - a research infrastructure.

Objectives of the current phase

CLARIN is currently in the preparatory phase, which has the aim of bringing the project to the level of legal, organisational and financial maturity required to implement the infrastructure. As the ultimate goal is the construction and operation of a shared distributed infrastructure to make language resources and technology available to the humanities and social sciences research communities at large, an approach along various dimensions is required in order to pave the way for implementation. The five main dimensions along which CLARIN will progress are the following:

- Funding and governance, bringing together the funding agencies in all participating countries and to work out a ready to sign draft agreement between them about governance, financing, construction and operation of the infrastructure.

- Technical infrastructure, defining the novel concept of a language resources and technology infrastructure, based on existing and emerging technologies (Grid, web services), to provide a detailed specification of the infrastructure, agreement on data and interoperability standards to be adopted, as well as a validated running prototype based on these specifications.

- Languages and multilinguality, populating the prototype with a selection of language resources and technologies for all participating languages, via the adaptation and integration of existing resources to the CLARIN requirements, and in a number of cases the creation of specific essential resources.

- Legal and ethical issues relating to language resources will have to be examined and thoroughly understood, and the necessary legal and administrative agreements proposed to overcome the barriers to full exploitation of the resources.

- Focus on users, the intended users being the humanities and social sciences research communities.

This final dimension is in many ways the most important, and will be explored in the most detail in this paper. In order to fully exploit the potential of what language technology has to offer, a number of actions have to be undertaken: (i) an analysis of current practice in the use of language technology in the

humanities will help to ensure that the specifications take into account the needs of the humanities, (ii) the execution of a number of exemplary humanities projects will help to validate the prototype and its specifications, (iii) the humanities and social sciences communities have to be made aware of the potential of the use of language resources and technology to enable innovation in their research, and (iv) the humanities and language technology communities have to be brought together in networks in order to ensure lasting collaborations between the communities. The objective of this cluster of activities is to ensure that the infrastructure has been demonstrated to serve the humanities and social sciences users, and that we create an informed community which is capable of exploiting and further developing the infrastructure.

Concluding remarks

CLARIN still has a long way to go, but it offers an exciting opportunity to exploit the achievements of language and speech technology over the last decades to the benefit of communities that traditionally do not maintain a close relationship with the latest technologies. In contrast to many European programmes, the main beneficiaries of this project are not expected to be the big ICT-oriented industries or the bigger language communities in Europe. CLARIN addresses the whole humanities and social sciences research community, and it very explicitly addresses all the languages of the EU and associated states, both majority and minority languages, including languages spoken and languages studied in the participating countries.

Fortune Hunting: Art, Archive, Appropriation

Lisa Young
Lisa_Young@brown.edu
Brown University, USA

James Stout
James_Stout@brown.edu
Brown University, USA

Elli Mylonas
Elli_Mylonas@brown.edu
Brown University, USA

Introduction

Although digital humanities projects have a variety of forms and emphases, a familiar type is the dataset that is prepared - encoded, structured, and classified - to allow scholars to engage with a research question. Such projects base the rational for their encodings and classification on earlier research, on disciplinary knowledge from the subject area and on encoding practice. The choices are then documented so researchers are aware of the assumptions underlying their dataset. The user interface that provides access to the data is also designed to privilege particular audiences and modes of interaction.

Most projects our group undertakes follow this model, and we are familiar with the process of elucidating the information necessary in order to implement them. Fortune Hunting, however, an art project, appropriates the forms of the archive and the methodologies of literary analysis to express a personal vision. The artist developed her representations of the fortune cookie texts without any knowledge of the work in digital humanities; she took her initial inspiration from the library catalog and simple desktop databases. As we proceeded with the web implementation of this art project, we became increasingly aware of the similarity of the artist's interactions to the encoding and classification tasks we used in other projects, and we drew on that paradigm. As she was introduced to these methodologies, classifications and analytical tools, the artist also discovered new ways of working with her texts.

Artist's Statement

The seemingly personalized fortune-cookie fortune is in fact a mass-produced item distributed to restaurants in batches. This random delivery system means that the fortune a given diner receives is mostly a function of chance. Even with this knowledge, many diners read their fortunes hoping to find an answer to a pressing question or recognize some part of themselves.

Over a period of years, I methodically saved the fortune-cookie fortunes I received while dining out. I curated my collection intuitively, keeping fortunes that presented an "accurate" description of myself ("you are contemplative and analytical by nature"), offered a sense of hope ("you will be fortunate in everything you put your hands to"), or evoked a rueful sense of misidentification ("you have great physical powers and an iron constitution").

I wanted to examine my desire to find structured meaning in what was really a haphazard or random occurrence. After some thought, I decided a digital archive that could sort the fortunes in a systematic way would be the perfect lens through which to read them. My work with the Brown University Scholarly Technology Group started with this germ of an idea. The result is the mini corpus "Fortune Hunting."

"Fortune Hunting" allows users to create an endlessly evolving series of narratives that explore both self-definition and chance, the incomplete and transitory nature of being and desired states of being, and the shifting nature of the personal pronoun "you." The anticipation, unexpected surprise (or possibly even disappointment) users experience as they construct searches continually enacts the same desiring mechanism (the need to construct meaning) that is as at the very root of the project.

As I began to work I realized that most fortunes were directed toward the diner through the use of the word "you." I sorted the fortunes into three categories: "subject" you, "object" you and "without" you. Within these categories, sorts were further specified by personal pronoun phrases. For example, all the fortunes that contain "you are" (subject you) can be seen together. Similarly, all the fortunes that contain "about you" (object you) can be seen together. A third method of searching ("without" you) captured all fortunes not containing the word "you" (example: art is the accomplice of love.)

At the same time, I became aware of the linguistic connections the search engine could manifest (for example: capturing all the fortunes that contained the word "love"). As a result, we created an individual word sort function. Words could be selected using the "exact word" search (for example: every) or a "similar word" search (which would yield every, everybody, everyone, and everything). The word sort function allowed fortunes to be sorted across multiple "you" categories.

As the database evolved, I became interested in ways that I could step outside its grammatical and linguistic parameters. At that point we developed a feature that allowed me to create "collections" of fortunes centered on my own subjective interpretations. These collections were divided into three categories: subjects (collections of fortunes on a particular topic such as travel, love or luck), portraits (clusters of fortunes that described types of individuals: writer, paranoid, neurotic overachiever), and narratives (groups of fortunes that created short stories: searching for happiness, reconciliations, revenge). Through analysis and intuition, my collection sorts created connections (both thematic and linguistic) that would not necessarily occur using either "you" or keyword searches.

Another important aspect of the database was incorporating the actual images of the fortunes. From the start, I was drawn to the ephemeral appearance of the fortunes: messages on paper scraps marked by stains and creases. With this in mind, I chose to scan the fortunes and have them appear on screen as images Instead of creating a strictly text-based interface. After users have completed their searches, they can view their results on a "picture page" which displays the images of the fortune cookie fortunes, or an "archive page" which displays the metadata attached to each fortune (allowing viewers another way to trace interconnections between fortunes). Lastly, viewers can move to a printer friendly page and print out their results, leaving each visitor with a visual record of their travels through the database.

"Fortune Hunting" is like going through a trunk in an attic, sorting through a collection of someone else's things and making your own connections and re-readings. Because all the searches are Boolean "or" searches users can cast a wide net as they comb the database. "Fortune Hunting" is constantly evolving, and each visit to the site can involve a different interpretation of the material. New fortunes and subsequent subject, portrait and narrative searches continue to be added.

Digital Humanist's Statement

Fortune Hunting is an art project; the artist was inspired by her reactions to fortune cookie texts to create a piece that would engender a similar experience in the viewer. The form she chose to exhibit her work, even before she began to work on its digital incarnation, highlighted her efforts to classify and inter-relate the fortunes. She created a large wall display, on which the fortunes were listed in columns that represented categories, with lines linking fortunes from different groups. Young's desire to explore the discovery of meaning from the randomness of the fortunes led her to quantitative tools and methods that are usually applied to carefully assembled and encoded archives. The "Fortune Hunting" database and website embody two competing modes of interaction: on the one hand, tool based interactions allow users to determine their own browsing strategy, using searches, keywords and a visualization, mimicking the artist's own experience of the materials. On the other hand, users can view the results of the artist's interpretations, a re-organization of the archive out of which she creates new meaning in the form of narratives and impressions.

Young developed tagging and classification as a way to interpret her fortune archive independently of her collaboration with a digital humanities group, but her use of these linguistic and analytical tools is essentially a misappropriation. The archive has been carefully compiled, but is based on personal evaluation, and oriented to creative ends rather than scholarship. The classifications that are applied to the texts are also subjectively

based. The formats and activities resemble conventional digital humanities research, but for a different purpose - the pleasurable discovery of meaning in the juxtaposition of essentially random artifacts. The whole analytical structure Is at the same time misused and exceedingly productive.

Like Ramsay's explorations of literary analysis [Ramsay2003], and the playful distortions of interpretation that we find in the Ivanhoe Game and Juxta [ARPProjects], this project has a liberating effect on the digital humanist. Just as Young plays with the fortune texts in order to understand how they "work," we are also playing with the subjective and exploratory application of our methods. The creative use of these tools allows us to focus on different parts of the discovery process, and we derive pleasure from the functioning of the tools themselves. In its similarity and great difference from conventional digital humanities projects, "Fortune Hunting" makes us self-conscious about our own practice.

Bibliography

[ARP] Applied Research in Patacriticism. http://www.patacriticism.org/

[ARPProjects] Applied Research in Patacriticism. Projects. http://www.patacriticism.org/projects.html

[FortuneHunting] Fortune Hunting Web Site. http://www.fortunehunting.org (site publication date: Dec.14, 2007)

[Ramsay2003] Ramsay, Stephen. "Toward an Algorithmic Criticism," Literary and Linguistic Computing 18.2 (2003)

[TAPoRTools] Tapor Tools. http://portal.tapor.ca/portal/coplets/myprojects/taporTools/

[VisualCol] Visual Collocator. http://tada.mcmaster.ca/Main/TAPoRwareVisualCollocator

Index of authors