

**The Association for Computers and the Humanities  
The Association for Literary and Linguistic Computing**

## **ACH / ALLC 2005**

**The International Conference on Humanities Computing and Digital Scholarship**

**The 17th Joint International Conference**

**University of Victoria**

**June 15 - June 18, 2005**



*Conference Abstracts (2nd Edition)*



**University of Victoria**

**Humanities Computing and Media Centre**

## **International Programme Committee Members**

- Alejandro Bia (chair) (Universidad de Alicante, Spain)
- Julia Flanders (Brown University, USA)
- Neil Fraistat (University of Maryland, USA)
- Simon Horobin (University of Glasgow, UK)
- Joseph Jones (University of British Columbia, Canada)
- Lisa Lena Opas-Hänninen (University of Oulu, Finland)
- Concha Sanz-Miguel (Universidad de Castilla La Mancha, Spain)
- Susan Schreibman (University of Maryland, USA)
- Michael Sperberg-McQueen (Association for Computing Machinery, USA)

## **Local Organizing Committee Members**

- Peter Liddell
- Scott Gerrity
- Stewart Arneil
- Martin Holmes
- Greg Newton
- Judy Nazar
- Ray Siemens
- Maire Consulting

## **Editorial Team**

- Peter Liddell
- Ray Siemens
- Alejandro Bia
- Martin Holmes
- Patricia Baer
- Greg Newton
- Stewart Arneil

2nd edition. ISBN: 1-55058-312-3

Published by

Humanities Computing and Media Centre,  
University of Victoria.

Conference logo by Richard Hunt.

Cover design by Greg Newton & Martin Holmes.

© 2005 University of Victoria and the authors.

# Introduction

To those of you reading this on arriving in Victoria for the ACH/ALLC 2005 conference — a warm westcoast welcome. While you're here, we hope you will enjoy our campus, the city of Victoria, and the enormous natural bounties of this region.

The presentations in this volume — more numerous than at previous conferences, we are told — are ample testimony to the range, depth, and frontline significance of what humanists and their colleagues are achieving with computers. The promise of how much more is to come also resonates through these pages.

Many people have contributed to the production of this book of abstracts. We would like first to thank the authors themselves. Alex Bia, as Chair of the Programme Committee, had the often unenviable task of leading the process to deliver their submissions to us for preparing this publication and its electronic 'mother'. He and his committee deserve our thanks for that, as do the authors who proof-read their abstracts and sent in corrections after the markup had been done. With no prevailing Style shaping the texts, it fell to me as chair of the Editorial Committee, with the help of Ray Siemens, Michael Best and the ever-precise Martin Holmes to try to create a more consistent 'look and feel' across the abstracts. Given the short time available, we opted for a slightly modified MLA style. What is most innovative, though, we owe to Martin, who decided to create a TEI-compliant XML data-base from which to generate this printed volume. That database, as Martin explains (over) is both the definitive text and a contribution to TEI scholarship in its own right. Martin dedicated a large portion of the months before the conference to this project, and the results speak for themselves. One afterthought: if anyone can create good tags for the humanists' beloved ambiguity and irony, we'll all be happy.

Finally I would like to express particular gratitude (I hope, on your behalf) to the local committee members who first collectively volunteered to take on the hosting when we were asked to consider it at relatively short notice. Each of them then took on specific roles, coordinated by Scott Gerrity and myself; and I'm sure they share the sense that if you never notice us during the conference, we've done a good job. The volunteers whom you will see are there to help with your technical needs and registration questions. With our best wishes for a successful conference and many fruitful insights from these pages and the presentations they represent.

Peter Liddell (chair)  
on behalf of the Local Organizing Committee



## About this abstract collection

The effort to produce this abstract book has been a true digital humanities project. All the abstracts were submitted electronically, marked up in TEI P4 XML, and stored in an XML database. All the output formats, including the PDF from which this printed volume was produced, were generated using the same XML technologies which are the subject of so many of the presentations themselves. As a primarily electronic publication, the true home of this document is its Internet location, on the conference Website (<http://web.uvic.ca/hrd/achallc2005/>). We will continue to maintain, correct and update abstracts there. This is the second edition of this book of abstracts, compiled after the end of the conference. It differs from the first, printed edition in that it includes several minor corrections; in addition, several abstracts whose authors were not actually able to attend or present their papers or posters have been removed, to reflect the actual program, and the abstracts for the two keynote presentations, by Anne Balsamo and Ian Lancashire, have been added. This electronic version of the text is the edition of record, and should be used for reference purposes.

On the Website you will also find XML, XHTML, PDF and plain text versions of all the individual abstracts, as well as the whole collection, and a basic search facility. We would like to encourage researchers to do text-analysis operations on the collection; it represents a snapshot of the state of humanities computing in 2005, and much could be learned from treating it as a textbase for analysis.

In formatting the abstracts for printing, we have done our best to follow MLA guidelines where this was practical, although there are a few cases where turning TEI XML into the exact format required by the MLA has proved beyond us, given the short time available for the production of the book. We sincerely hope the high quality of the content will mitigate the faults in its presentation.

Martin Holmes  
Humanities Computing and Media Centre



---

# Table of Contents

Introduction .....	I
<i>Peter Liddell</i>	
About this abstract collection .....	III
<i>Martin Holmes</i>	

---

## Index of Abstracts

La Docencia de la Creación Digital Emergente: La Literatura Electrónica .....	3
<i>Joan-Elies Adell</i>	
Measuring the Usefulness of Function Words for Authorship Attribution .....	4
<i>Shlomo Argamon, Shlomo Levitan</i>	
Designing Culture: A Work of the Technological Imagination .....	7
<i>Anne Balsamo</i>	
<i>HyperJournal</i> .....	8
<i>Michele Barbera, Nicolò D'Ercole</i>	
Semantic Context Visualization to Promote Vocabulary Learning .....	10
<i>Caroline Barrière, Claude St-Jacques</i>	
Playing Many Parts: Models of Collaboration in an Electronic Edition .....	13
<i>Michael Best, Jessica Slights, Peter van Hardenberg, Wendy Huot, Alan Galey</i>	
<i>Understanding Poetry Online: an Internet Application for Teaching</i> .....	17
<i>Jonathan Blake</i>	
'La Imaginación al Servicio de la Educación': un Ejemplo de <i>Work in Progress</i> .....	18
<i>Laura Borràs, Isabel Clara Moll Soldevila, Roger Canadell</i>	
Supporting Annotation as a Scholarly Tool: Experiences from the <i>Online Chopin Variorum</i> <i>Edition</i> .....	20
<i>John Bradley, Paul Vetch</i>	
Improving Access to Encoded Primary Texts .....	23
<i>Terry Butler</i>	
Learning Objects in Humanities Education .....	24
<i>Terry Butler, Catherine Caws, Norm Friesen, Scott Leslie, Griff Richards, Ray Siemens</i>	
"Temas de Literatura Universal": Usos y Aplicaciones del Hipertexto Pedagógico .....	26
<i>Roger Canadell, Laura Borràs, Isabel Clara Moll Soldevila</i>	

---

A Pilot Study for a Navajo Textbase .....	28
<i>Kip Canfield</i>	
The <i>Tibet Oral History Archive Project</i> and Digital Preservation .....	30
<i>Linda Cantara</i>	
<i>DocScapes: Visualizing Document Structures with SVG</i> .....	32
<i>Hugh Cayless</i>	
DeMeCoT, The Delftse Methode Conversation Trainer .....	35
<i>Amal Chatterjee, Piet Meijer</i>	
Reflexivity and Arts Informatics .....	36
<i>Chris Chesher</i>	
Laying that Damned Book Aside? Evaluating the Digital <i>Doctor Faustus</i> .....	38
<i>Tanya Clement</i>	
Text Modeling and Visualization with Network Graphs .....	40
<i>Aaron Coburn</i>	
In the Philosophy Room: Australian Realism and the Digital Content Object .....	42
<i>Creagh Cole, Paul Scifleet</i>	
Developing the Humanities HyperMedia Centre @ Acadia University .....	44
<i>Richard Cunningham, David Duke, John Eustace, Anna Galway, Erin Patterson</i>	
Using Ancillary Text to Index Web-Based Multimedia Objects .....	45
<i>Lyne Da Sylva, James Turner</i>	
A la Carte Schema: A Case Study Comparison of the Application of DTDs and XML Schema to the Carte Calendar Project Template .....	47
<i>Ingrid Daneker, Claire Warwick</i>	
Modelling Complex Multimedia Relationships in the Humanities Computing Context: Are Dublin Core and FRBR up to the Task? .....	50
<i>J. Stephen Downie, Allen Renear, Adam Mathes, Karen Medina, David Dubin, Jin Ha Lee</i>	
A Revolutionary Approach to Humanities Computing?: Tools Development and the <i>D2K</i> Data-Mining Framework .....	53
<i>J. Stephen Downie, John Unsworth, Bei Yu, David Tcheng, Geoffrey Rockwell, Stephen J. Ramsay</i>	
A Declarative Framework for Modeling Pronunciation and Rhyme .....	55
<i>David Dubin, David J. Birnbaum</i>	
<i>Cardplay</i> , a New Textual Instrument .....	58
<i>David Durand, Noah Wardrip-Fruin</i>	
User Generated Metadata: Creating Personalized Web Experiences .....	60
<i>Michael Fegan, Bill Hart-Davidson, Joy Palmer, Dean Rehberger</i>	
Advanced Topics in TEI .....	66
<i>Julia Flanders, Syd Bauman, Laurent Romary, David J. Birnbaum, Matthew Zimmerman</i>	
Hybrid Cyber-Librarians: The CLIR Post-Doctoral Fellowship in Scholarly Information Resources for Humanists .....	68
<i>Amanda French, John Unsworth, Susan Nutter, Sarah Michalak, Patricia Hswe, Daphnée Rentfrow</i>	



---

The Canadian Century Research Infrastructure Project and Computing in the Humanities .....	69
<i>Chad Gaffield, Marc St-Hilaire, Claude Bellavance, Gordon Darroch, Peter Baskerville</i>	
METS in Action: Standardization and Interoperability in the Digital Library .....	70
<i>Richard Gartner, Rick Beaubian, Jerome McDonough, Susan Dahl, Brian Tingle</i>	
<i>Clotel</i> : An Electronic Scholarly Edition .....	72
<i>Matthew Gibson</i>	
Words as Data? Estimating the Fiscal Conservativeness of Provincial Premiers Using the Wordscore Procedure of Content Analysis .....	73
<i>André S. Gosciniak</i>	
An Examination of the Authorship Attributions of Two Major Roman Authors .....	75
<i>Lyman W. Gurney, Penelope J. Gurney</i>	
Gottlob Frege's <i>Grundgesetze der Arithmetik</i> : Computational Linguistics Meets the Founder of Logicism .....	77
<i>Felicitas Haas, Bernhard Schröder</i>	
Delta, Delta Prime, and Modern American Poetry: Authorship Attribution Theory and Method .....	79
<i>David Hoover</i>	
The Delta Spreadsheet .....	80
<i>David Hoover</i>	
Concurrent Markup Hierarchies: a Computer Science Approach .....	82
<i>Ionut Emil Iacob, Alex Dekhtyar</i>	
<i>Edition Production Technology</i> : an Eclipse-Based Platform for Building Image-Based Electronic Editions .....	84
<i>Ionut Emil Iacob, Kevin Kiernan, Alex Dekhtyar</i>	
Animated Dynamic Highlighting .....	87
<i>Bill Janssen, Olga Gurevich, Lauri Karttunen</i>	
The ARCHway Software Infrastructure: a Demo of a Platform and Utilities for Building Applications for Electronic Editions .....	90
<i>Jerzy W. Jaromczyk, Neil Moore</i>	
In Search of Humanities Computing in Teaching, Learning and Research .....	91
<i>Martyn Jessop</i>	
Are Targeted User-Centred Interfaces the Key to Facilitating the Conversion of the Traditional Non-User to a User of Archives? .....	93
<i>Andrea Johnson</i>	
Towards an Automatic Index Generation Tool .....	95
<i>Patrick Juola</i>	
A Prototype for Authorship Attribution Software .....	97
<i>Patrick Juola</i>	

---

Social, Geographical, and Register Variation in Dutch: From Written MOGELIJK to Spoken MOK .....	100
<i>Karen Keune, Mirjam Ernestus, Roeland van Hout, Harald Baayen</i>	
The <i>Edition Production Technology (EPT)</i> and the <i>ARCHway</i> and <i>Electronic Boethius</i> Projects .....	102
<i>Kevin Kiernan, Dorothy Porter, Alex Dekhtyar, Ionut Emil Iacob, Jerzy W. Jaromczyk, Neil Moore</i>	
Historical Lexicons in Medieval and Early Modern English and French .....	109
<i>Anne Lancashire, Jennifer Roberts-Smith, Brian Merrilees</i>	
Cybertextuality and Text Analysis .....	115
<i>Ian Lancashire</i>	
Using Markup for Multivariate Analyses in the Prosopographical Study "Formation for the Public Sphere" .....	116
<i>Monica Langerth Zetterman</i>	
Markup of Educational Content .....	119
<i>Monica Langerth Zetterman</i>	
Creating an Archives Management System at the University of Maryland Libraries .....	120
<i>Jennie A. Levine, Amit Kumar, Susan Schreibman, Jennifer Evans</i>	
Story Generators: Models and Approaches for the Generation of Literary Artefacts .....	123
<i>Birte Lönneker, Jan Christoph Meister, Pablo Gervás, Federico Peinado, Michael Mateas</i>	
Mysteries in Time and Space: Historical Computing .....	131
<i>John Lutz, Patrick Dunae, John Bonnett</i>	
Reaching Out: What do Scholars Want from Electronic Resources? .....	133
<i>Shawn Martin</i>	
Human Computing: Modelling with Meaning .....	135
<i>Willard McCarty, Meurig Beynon, Steve Russ</i>	
Heraldic Applications of Computational Linguistics, Computational Geometry and Image Processing .....	142
<i>Michael McKeag</i>	
Classifying the Chimera .....	146
<i>Federico Meschini</i>	
The Computed Synoptic Table —Tele-Synopsis for Biblical Research .....	148
<i>Maki Miyake, Hiroyuki Akama, Masanori Nakagawa, Nobuyasu Makoshi</i>	
El Trabajo Final de Carrera en Filología: Perspectivas Hacia un Nuevo Horizonte .....	152
<i>Isabel Clara Moll Soldevila, Laura Borràs, Roger Canadell</i>	
An Examination of the Rhetorics of Digital Scholarship and the Emerging Digital Monograph .....	154
<i>Elli Mylonas</i>	
Testing EAD Encoding in the <i>Texas Archival Resources Online (TARO)</i> System with Textual Analysis Techniques .....	156
<i>Vidya Narayan, Patricia Galloway</i>	

---

The <i>LICHEN</i> Project: The Linguistic and Cultural Heritage Electronic Network .....	157
<i>Lisa Lena Opas-Hanninen, Jean Anderson, Ilkka Juuso, Tapio Seppänen</i>	
Multicultural Issues on the TEI's Horizon: the Case of Tibetan Texts .....	159
<i>Linda E. Patrik</i>	
Automatic Discovery of NLP Resources on the Web .....	161
<i>Viktor Pekar, Richard Evans</i>	
<i>The Robert Graves Diary (1935-39): a TEI Application using an XML Database (eXist)</i> .....	164
<i>Chris Petter, Elizabeth Grove-White, Linda Roberts, Spencer Rose, Jessica Posgate, Jillian Shoichet</i>	
An Encoding Model for Librettos: the Opera Liber DTD .....	167
<i>Elena Pierazzo</i>	
SVG Visualization of TEI Texts .....	169
<i>Wendell Piez</i>	
Regelbasierte Suche in Textdatenbanken mit Nichtstandardisierter Rechtschreibung (Rule Based Search in Text Databases with Non-Standard Orthography) .....	170
<i>Thomas Pilz, Prof. Dr. Wolfram Luther, Prof. Dr. phil. Ulrich Ammon, Prof. Dr.-Ing. Norbert Fuhr</i>	
<i>The Walt Whitman Archive: Archivist-Scholar Collaboration in Description and Representation</i> .....	174
<i>Kenneth Price, Katherine Walter, Terence Catapano, Daniel Pitti</i>	
Exhibition: A Problem for Conceptual Modeling in the Humanities .....	176
<i>Allen H. Renear, Jin Ha Lee, Yunseon Choi, Xin Xiang</i>	
L'Autoguidage: une Approche pour le Perfectionnement du Français Écrit en Milieu Minoritaire .....	180
<i>Sylvain Rheault</i>	
National Support for Humanities Computing: Different Achievements, Needs and Prospects .....	182
<i>David Robey, John Unsworth, Geoffrey Rockwell</i>	
TAPoR: Five views through a text analysis portal (COCH/COSH Allied Association Session) .....	185
<i>Geoffrey Rockwell, Stéfan Sinclair, James Chartrand</i>	
The Non-Traditional Case for the Authorship of the Twelve Disputed "Federalist" Papers: A Monument Built on Sand? .....	187
<i>Joseph Rudman</i>	
Interface Design .....	191
<i>Stan Ruecker, Stéfan Sinclair, Stephen Ramsay, Milena Radzikowska, Alan Galey</i>	
Academic Libraries and Information Communities: New Models for Supporting Digital Scholarship .....	197
<i>Christine Ruotolo</i>	
Modeling Diachrony in Dictionaries .....	198
<i>Susanne Salmon-Alt, Laurent Romary, Buchi Eva</i>	

---

Spanish Morphosyntactic Disambiguator .....	201
<i>Octavio Santana Suárez, José Rafael Pérez Aguiar, Luis Javier Losada García, Francisco Javier Carreras Riudavets</i>	
Una Herramienta de Recuperación Morfoléxica Aplicada a <i>Microsoft Word</i> .....	204
<i>Octavio Santana Suárez, Zenón Hernández Figueroa, Gustavo Rodríguez Rodríguez, Luis Losada García</i>	
A Digital Environment for Neolatin Studies .....	206
<i>Ross Scaife, Andrew Gollan, William du Cassé, Jennifer Nelson</i>	
Letters and Lacunae: Editing an Electronic Scholarly Edition of Correspondence .....	208
<i>Susan Schreibman, Gretchen Gueguen, Amit Kumar, Ann Saddlemeyer</i>	
The <i>Blackwell Companion to Digital Humanities</i> : a Roundtable Discussion .....	210
<i>Susan Schreibman, Ray Siemens, John Unsworth, Willard McCarty, Martha Nell Smith, Geoffrey Rockwell, Abby Smith, Claire Warwick, Perry Willett</i>	
Representation of Meaning: a Graphical and Interactive Approach .....	212
<i>Gary Shawver, Oliver Kennedy</i>	
Keyword Extraction in Information Retrieval .....	214
<i>Harold Short, Marilyn Deegan, Laszlo Hunyadi, Paul Baker, Dawn Archer, Tony McEney</i>	
Theory and Practise in Literary Textual Analysis Tools .....	215
<i>Ray Siemens, Geoffrey Rockwell, Susan Schreibman, Matthew Jockers</i>	
The <i>Virtual Lightbox for Museums and Archives</i> : A Distributed Solution for Structured Data Reuse Across Multiple Visual Resources .....	218
<i>Amy Smith, Brian Fuchs, Leif Isaksen</i>	
<i>Callimachus</i> : A Virtual Archivist for Electronic Markup Projects .....	220
<i>Jeff Smith, Joel Deshayes, Peter Stoicheff</i>	
Integrating a Massive Digital Video Archive into Humanities Teaching and Research .....	221
<i>Lisa Spiro, Diane Butler, Chris Pound</i>	
<i>Early Modern Literary Studies</i> : Preparing for the Long Run .....	223
<i>Matthew Steggle</i>	
Profiling Stylistic Variations in Dickens and Smollett through Correspondence Analysis of Low Frequency Words .....	224
<i>Tomoji Tabata</i>	
Disciplined: Using Curriculum Studies to Define 'Humanities Computing' .....	227
<i>Melissa Terras</i>	
Finalizing the Multiple-Text Electronic <i>King Lear</i> for Use in the Classroom .....	230
<i>Stephanie F. Thomas</i>	
The <i>MetaMap</i> , an Online Tool for Learning about Metadata .....	232
<i>James Turner</i>	
Visual Knowledge: Textual Iconography of the <i>Quixote</i> , a Hypertextual Archive .....	233
<i>Eduardo Urbina, Richard Furuta, Steven Escar Smith</i>	

---

Mining the Differences between Penninc and Vostaert .....	237
<i>Karina van Dalen-Oskam, Joris van Zundert</i>	
The e-Laborate Project and the Usability of Another Textual Paradigm .....	240
<i>Joris van Zundert, Karina van Dalen-Oskam</i>	
The <i>Abraham Lincoln Historical Digitization Project</i> , the World Wide Web, and the Public Humanities .....	243
<i>Drew VandeCreek</i>	
Databases and Prosopographies: <i>The Prosopography of Anglo-Saxon England (PASE)</i> a Case Study .....	244
<i>Hafed Walda, Alex Burghart</i>	
TM4DH (Topic Maps for Digital Humanities): Examples and an Open Source Toolkit .....	246
<i>John Walsh</i>	
Exploring the Use of Term Proximity in Collocate-Ranking for Query Expansion .....	248
<i>Ying Wang, Olga Vechtomova</i>	
User Centred Interactive Search: a Study of Humanities Researchers in a Digital Library Environment .....	250
<i>Claire Warwick, Ann Blandford, George Buchanan</i>	
Annotating Electronic Texts of Shakespeare .....	253
<i>Philip Weller</i>	
A New Methodology for Parts of the Dutch Price History, Based on Analysis of the Paalgeld Portbooks, 1771-1778 .....	254
<i>George Welling</i>	
Approaches to Searching for Language and Diversity in a 'Whitebread City' Digital Corpus: The Charlotte Conversation and Narrative Collection .....	255
<i>Stephen Westman, Boyd Davis</i>	
Action and Interaction in Music and New Media Art: Exploration of Musicians' Performative and Interactive Decisions as Evidenced by Annotated Musical Scores .....	257
<i>Megan Winget</i>	
Texttechnologie in der Universitären Lehre .....	260
<i>Andreas Witt, Dieter Metzling</i>	
From Text to Topics — Zigzagging Towards the Knowledgebase of Tang Civilization .....	262
<i>Christian Wittern</i>	
Reading Potential: The Oulipo and the Meaning of Algorithms .....	264
<i>Mark Wolff</i>	
The <i>Online Nahuatl Dictionary</i> : A Model for Interdisciplinary Multicultural Collaboration .....	266
<i>Stephanie Wood, Judith Musick, William Henderson</i>	
English Usage Comparison between Native and non-Native English Speakers in Academic Writing .....	268
<i>Bei Yu, Qiaozhu Mei, Chengxiang Zhai</i>	

---

Play and Code in Humanist Research ..... 271  
*Vika Zafrin*

---

Index of Presenters ..... 274

Index of Topic Keywords ..... 277

# *Abstracts*





# La Docencia de la Creación Digital Emergente: La Literatura Electrónica

Joan-Elies Adell ([jadellp@uoc.edu](mailto:jadellp@uoc.edu))

Universitat Oberta de Catalunya

En estos últimos años parece que la investigación en literatura haya empezado a darse cuenta de la necesidad de estudiar los resultados de la unión entre tecnologías digitales y la creación literaria. Son muchos los autores que se han referido con anterioridad a los nuevos escenarios que, con la llegada de Internet y las tecnologías informáticas, se han abierto para lo literario, tanto desde el punto de vista de la difusión de la literatura en soporte tradicional, que tiene como formato básico de difusión el libro impreso en papel, como para las nuevas obras creadas exclusivamente desde un medio digital y que deben ser leídas obligatoriamente desde un soporte informático.

No obstante, también es bien conocido el rechazo ancestral que lo nuevo produce a muchos de los estudiosos de la literatura. Parece como si únicamente el profesor de literatura pudiese ostentar con orgullo una especie de 'desvalorización del presente', transformando un síntoma psicológico de miedo a la contemporaneidad en un aparente valor positivo de autoridad y saber. No me imagino a los profesores y críticos de otros ámbitos de la ciencia jactándose de desconocer todo lo que está ocurriendo a su alrededor.

¿Cómo no podía la literatura participar de la ampliación de nuestros sentidos para dinamitar sus límites tradicionales? El impacto de las tecnologías digitales, de la informática, ha convertido también el campo de la literatura en un espacio de tanteo constante donde las posibilidades de creación de nuevos lenguajes y nuevas concepciones de lo poético ofrece un territorio fértil y expectante, para su estudio. Como es bien sabido, la tecnología digital ofrece oportunidades insospechadas para la producción, el almacenaje, la difusión y la docencia de textos literarios escritos inicialmente para ser difundidos en un soporte impreso, pero si queremos ser serios y rigurosos con nuestro presente, también deberemos enfrentarnos a todo el potencial que estos textos electrónicos nos ofrecen, analizar el cambio que supone para la literatura la aparición del espacio electrónico que, gracias a la informática, internet y el ciberespacio, se ha convertido en un lugar de 'poesis', como lo llama Loss Pequeño Glazier en su libro *Digital Poetics: The*

*Making of E-Poetries*, esto es, un no-lugar con unas condiciones específicas de textualidad.

Esta nueva condición específica de la textualidad electrónica hace que nos preguntemos si son necesarias nuevas herramientas teóricas, otros conceptos que nos sean útiles para leer críticamente este tipo de literatura, para saberla explorar desde el punto de vista del análisis, para tratar de averiguar y conocer en qué medida estas nuevas prácticas literarias nos formulan preguntas antes impensables sobre nuestra propia tradición y sobre nuestra conceptualización de lo literario. Si tenemos en cuenta las propiedades específicas de la escritura electrónica incluyen el hecho de que el texto es no es únicamente físico, tocable, sino que es mostrado, expuesto, exhibido, nos daremos cuenta que introduce elementos de reflexión que hace que la práctica literaria digital se acerque en algún punto a algunas de las características propias del discurso cinematográfico: donde el movimiento, el sonido y el montaje son determinantes. Con todo, hay quien piensa, por ejemplo Espen Aarseth que no es nada obvio que conceptos tales como 'literatura informática' o 'textualidad electrónica' merezcan ser defendidos teóricamente, que haya que concederles un estatuto axiomático, de la misma manera que piensa que la idea que el ordenador "sea capaz de producir un cambio histórico y social por sí mismo es una idea errónea que resulta extrañamente ahistórica y antropomórfica." (Aarseth 135)

Nuestra opinión, como docentes de una universidad virtual, es del todo diferente. Se nos hace urgente y necesaria la investigación sobre estas prácticas literarias emergentes, más allá de su interés por estudiar las posibilidades de la tecnología como herramienta para la enseñanza de la literatura, ya que nos sitúan ante un primer reto que es de tener que ofrecer un primer diagnóstico, necesariamente apresurado y con poca perspectiva histórica, sobre muchas de las nuevas creaciones que se vienen produciendo en el entorno electrónico.

Desde la Universidad Oberta de Catalunya, por ello, hemos asumido el reto no sólo de reflexionar histórica y teóricamente sobre estas nuevas maneras de ser de la literatura, sino también plantearnos las dificultades que supone su análisis y su enseñanza. Con esta comunicación, pues, querría explicar el proceso y diseño de una asignatura, abierta y dinámica, 'Estudios literarios y tecnologías digitales', pensada para enriquecer la oferta de literatura más estrictamente contemporánea, una especie de 'trabajo de campo' sobre el estado actual de la literatura digital. La dificultad, como hemos explicado, radica en que el objeto de análisis del curso es tan cercano y en parte aún in progress, que exige una dinámica de funcionamiento con los alumnos que debe tener un aire general de reader, de investigación colectiva en marcha, abierta y dialógica. Los objetivos de este curso son los siguientes:

1. Vencer la reticencia habitual de los estudios de literatura a los fenómenos más estrictamente contemporáneos, tratando

de que los alumnos se aproximen sin prejuicios al estudio de la literatura digital.

2. Conseguir que el estudiante sistematice una serie de conocimientos teóricos y un utillaje crítico suficiente para poder empezar a hablar significativamente sobre los textos estudiados, al tiempo que puedan adquirir un conocimiento básico sobre estas nuevas formas de ser de la literatura, y los cambios epistemológicos que ello comporta.
3. Aprender a aprovechar los recursos de la red para explorar y adentrarse en el estudio y la lectura de este tipo de obras.

El objetivo final de esta experiencia, pues, podría resumirse en un intento de creación colectiva de un aparato crítico capaz de orientarnos en la lectura de las obras literarias digitales emergentes, de intentar llegar a un consenso a través de la discusión constante y productiva entre profesores y alumnos, sobre cuales deben ser los aspectos más remarcables que han de ser valorados y subrayados en la creación digital. Se trata, en definitiva, de poder debatir y reflexionar sobre una literatura que aún no ha sido jerarquizada y que su valoración crítica todavía no ha cristalizado, que aún se encuentra en un estado virgen de atribución canónica. Como profesores de teoría de la literatura hemos intentado asumir, con el diseño y la puesta en marcha de esta asignatura, el riesgo de reflexionar junto con nuestros alumnos sobre las 'condiciones de legibilidad' de la literatura digital, de la literatura del presente.

## Bibliography

Aarseth, Espen. *Literatura y cibercultura*. Madrid: Arco Libros, 2004.

Glazier, Loss Pequeño. *Digital Poetics: The Making of E-Poetries*. Tuscaloosa: University of Alabama Press, 2002.

---

# Measuring the Usefulness of Function Words for Authorship Attribution

---

*Shlomo Argamon* (*argamon@iit.edu*)

*Illinois Institute of Technology*

*Shlomo Levitan* (*levishl@iit.edu*)

*Illinois Institute of Technology*

---

## Introduction

Some forty years ago, Mosteller and Wallace suggested in their influential work on the Federalist Papers that a small number of the most frequent words in a language ('function words') could usefully serve as indicators of authorial style. The decades since have seen this work taken up in many ways including both the use of new analysis techniques (discriminant analysis, PCA, neural networks, and more), as well as the search for more sophisticated features by which to capture stylistic properties of texts. Interestingly, while use of more sophisticated models and algorithms has often led to more reliable and generally applicable results, it has proven quite difficult to improve on the general usefulness of function words for stylistic attribution. Indeed, John F. Burrows, in his seminal work on Jane Austen, has demonstrated that function words can be quite effectively used for attributing text passages to different authors, novels, or individual characters.

The intuition behind the utility of function words for stylistic attribution is as follows. Due to their high frequency in the language and highly grammaticalized roles, function words are very unlikely to be subject to conscious control by the author. At the same time, the frequencies of different function words vary greatly across different authors and genres of text - hence the expectation that modeling the interdependence of different function word frequencies with style will result in effective attribution. However, the highly reductionistic nature of such features seems unsatisfying, as they rarely give good insight into underlying stylistic issues, hence the various efforts at developing more complex textual features while respecting constraints on computational feasibility.

One especially promising line of work in this regard has been the examination of frequent word sequences and collocations for stylistic attribution, particularly Hoover's recent (2004) systematic work on clustering analysis of several text collections

using frequent word collocations. A "word collocation" is defined as a certain pair of words occurring within a given threshold distance of each other (such as "is" and "certain" appearing within 5 words of each other in this sentence). Given such a threshold, the most frequent such collocations are determined over the entire corpus, and their frequencies in each text constitute its features for analysis. Hoover's analyses show the superiority, for his data set, of using frequent word collocations (for certain window sizes) over using frequent words or pairs of adjacent words.

We contend, however, that by using such a small data set (twenty samples of 10,000 words each, in one case), the discriminating power of a model based on function words will be much reduced, and so the comparison may not be fair. As has been shown for other computational linguistic tasks (see, e.g., Banko & Brill), even simple language modeling techniques can greatly improve in effectiveness when larger quantities of data are applied. We have therefore explored the relative effectiveness of frequent words compared to frequent pairs and collocations, for attribution of both author identity and national origin, increasing the number of text passages considered over earlier work.

We performed classification experiments on the twenty novels considered by Hoover, treating each separate chapter of each book as a separate text (rather than using just the first 10,000 words of each novel as a single text). Table 1 gives the full list with numbers of chapters and average number of words per chapter. We used a standard state-of-the-art machine learning technique to derive linear discrimination models between pairs of authors. This procedure gave results that clearly show a superiority of function words over collocations as stylistic features. Qualitatively similar results were obtained for the two-class problem of attributing the national origin (American or British) of a text's author. We conclude from this that larger and more detailed studies need to be done to effectively validate the use of a given feature type for authorship attribution.

Author	Book	# Chapters	Avg. Words
Cather	<i>My Antonia</i>	45	1826
	<i>Song of the Lark</i>	60	2581
	<i>The Professor's House</i>	28	2172
Conrad	<i>Lord Jim</i>	45	2913
	<i>The Nigger of the Narcissus</i>	5	10592
Hardy	<i>Jude the Obscure</i>	53	2765
	<i>The Mayor of Casterbridge</i>	45	2615
	<i>Tess of the d'Urbervilles</i>	58	2605
James	<i>The Europeans</i>	12	5003

	<i>The Ambassadors</i>	36	4584
Kipling	<i>The Jungle Book</i>	13	3980
	<i>Kim</i>	15	7167
Lewis	<i>Babbitt</i>	34	3693
	<i>Main Street</i>	34	4994
	<i>Our Mr. Wrenn</i>	19	4126
London	<i>The Call of The Wild</i>	7	4589
	<i>The Sea Wolf</i>	39	2739
	<i>White Fang</i>	25	2917
Wells	<i>The Invisible Man</i>	28	1756
	<i>The War Of The Worlds</i>	27	2241

Table 1. Corpus composition.

## Methodology

Given each particular feature set (frequent words, pairs, or collocations), the method was to represent each document as a numerical vector, each of whose elements is the frequency of a particular feature of the text. We then applied the SMO learning algorithm (Platt) with default parameters, which gives a model linearly weighting the various text features. SMO is a support vector machine (SVM) algorithm; SVMs have been applied successfully to a wide variety of text categorization problems (Joachims).

Generalization accuracy was measured using 20-fold cross-validation, in which the 633 chapters were divided into 20 subsets of nearly equal size (3 or 4 texts per subset). Training was performed 20 times, each time leaving out one of the subsets, and then using the omitted subset for testing. The overall classification error rate was estimated as the average error rate over all 20 runs. This method gives a reasonable estimate of the expected error rate of the learning method for each given feature set and target task (Goutte).

## Results

Results of measuring generalization accuracy for different feature sets are summarized in Tables 2 and 3, which clearly show that using the most frequent words in the corpus as features for stylistic text classification gives the highest overall discrimination for both author and nationality attribution tasks.

Feature Set	Author	Nationality
Freq. Words	<b>99.00%</b>	<b>93.50%</b>
Freq. Pairs	91.60%	91.30%
Freq. Coll. (k=5)	88.94%	90.20%
Freq. Coll. (k=10)	84.00%	87.20%

Table 2. 20-fold cross-validation results for 200 most frequent words, pairs, and collocations (window size  $k = 5$  or 10).

Feature Set	Author	Nationality
Freq. Words	93.20%	<b>93.50%</b>
Freq. Pairs	90.00%	88.60%
Freq. Coll. (k=5)	91.50%	92.10%
Freq. Coll. (k=10)	94.00%	92.10%

Table 3. 20-fold cross-validation results for 500 most frequent words, pairs, and collocations (window size 5 or 10).

## Discussion

Our study here reinforces many others over the years in showing the surprising resilience of frequently-occurring words as indicators of the stylistic character of a text. Our results show frequent words enabling more accurate text attribution than features such as word pairs or collocations, surprisingly contradicting recent results as well as the intuition that pairs or collocations should be more informative. The success of this study at showing the power of frequent words we mainly attribute to the use of more data, in the form of entire novels, broken down by chapters. The more fine-grained breakdown of text samples for each author enables more accurate determination of a good decision surface for the problem, thus better utilizing the power of all features in the feature set. Furthermore, using more training texts than features seriously reduces the likelihood of overfitting the model to the training data, improving the reliability of results.

It is indeed possible that collocations may be better than function words for different stylistic classification tasks; however such a claim remains to be proven. A more general interpretation of our results is that since a set of frequent collocations of a given size will contain fewer different words than a set of frequent words of the same size, it may possess less discriminatory power. At the same time, though, such a feature set will be less subject to overfitting, and so may appear better when very small sets of texts are studied (as in previous studies). Our results thus lead us to believe that most of the discriminating power of collocations is due to the frequent words they contain (and not the collocations themselves), thus

frequent words outperformed collocations, given sufficient data.

## Conclusions

Function words still prove surprisingly useful as features for stylistic text attribution, even after many decades of research on features and algorithms for stylometric analysis. We believe that significant progress is likely to come from fundamental advances in computational linguistics which allow automated extraction of more linguistically motivated features, such as recent work on extracting rhetorical relations in a text (Marcu).

More generally, our results argue for the importance of using larger data sets for evaluating the relative utility of different attribution feature sets or techniques. As in our case of comparing frequent words with frequent collocations, changing the scale of the data set may affect the relative power of different techniques, thus leading to different conclusions. We suggest that the authorship attribution community should now work towards developing a large suite of corpora and testbed tasks, to allow more rigorous and standardized comparisons of alternative approaches.

## Bibliography

- Baayen, H., H. van Halteren, and F. Tweedie. "Outside the cave of shadows: Using syntactic annotation to enhance authorship attribution." *Literary and Linguistic Computing* 11.3 (1996): 121-132.
- Banko, M., and E. Brill. "Scaling to very very large corpora for natural language disambiguation." *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*. 2001. 26-33.
- Biber, D., S. Conrad, and R. Reppen. *Corpus Linguistics: Investigating Language Structure and Use*. Cambridge: Cambridge University Press, 1998.
- Burrows, J. *Computation into Criticism: A Study of Jane Austen's Novels and an Experiment in Method*. Oxford: Clarendon Press, 1987.
- Goutte, C. "Note on free lunches and cross-validation." *Neural Computation* 9.6 (1997): 1246-9.
- Hoover, D.L. "Frequent collocations and authorial style." *Literary and Linguistic Computing* 18.3 (2004): 261-28.
- Joachims, T. "Text categorization with Support Vector Machines: Learning with many relevant features." *Machine*

*Learning: ECML-98, Tenth European Conference on Machine Learning*. 1998. 137-142.

Marcu, D. "The Rhetorical Parsing of Unrestricted Texts: A Surface-Based Approach." *Comp. Ling.* 26.3 (2000): 395-448.

Matthews, R., and T. Merriam. "Neural computation in stylometry: An application to the works of Shakespeare and Fletcher." *Literary and Linguistic Computing* 8.4 (1993): 203-209.

Mosteller, F., and D.L. Wallace. *Inference and Disputed Authorship: The Federalist*. Reading, Mass: Addison Wesley, 1964.

Platt, J. *Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector Machines*. Microsoft Research Technical Report MSR-TR-98-14, 1998. Accessed 2005-03-08. <ftp://ftp.research.microsoft.com/pub/tr/tr-98-14.pdf>

Stamatatos, E., N. Fakotakis, and G. Kokkinakis. "Computer-based authorship attribution without lexical measures." *Computers and the Humanities* 35 (2001): 193-214.

---

## Designing Culture: A Work of the Technological Imagination

---

*Anne Balsamo* ([abalsamo@annenbergs.edu](mailto:abalsamo@annenbergs.edu))

*USC; Keynote Speaker, sponsored by Canadian Institute for Advanced Research*

---

To stimulate a discussion about the many ways in which culture influences the practices of technology design, I present examples of technologies and digital applications whose designs were explicitly informed by cultural theory. In 1999, a research group at Xerox PARC built an interactive museum exhibit called "XFR: Experiments in the Future of Reading." The resulting exhibit explored different facets of the nature of reading in a digital culture. In describing those moments when cultural theory, values, and conventions become an explicit part of the design process, I reflect on the 'technological imagination' at work, and how the exercise of this imagination, in turn, results in the development of new literacies, modes of expression, as well as devices and digital artifacts.

## HyperJournal

---

**Michele Barbera** (*barbera@netseven.it*)

University of Bologna

**Nicolò D'Ercole** (*nicolo.dercole@virgilio.it*)

University of Pisa

---

**D**uring the last decades a market failure known as serial price crisis has led to a market configuration where the major editorial groups retain oligopolistic control over the editorial market. This market situation is reflected in the research community which largely depends on public founding.

The access to research output in form of articles has become difficult because articles are often not easy to find and (even worst) subscription fees are often unaffordable for a large number of libraries.

This situation has potentially devastating consequences over the publicly founded research field and threatens the freedom to share scientific knowledge (Guédon). During the last 5 years the Open Access movement have proposed some solutions to the problem (Suber). Some of the Open Access initiatives gained wide acceptance in natural sciences (Public Library of Science; Los Alamos ArXiv). Unfortunately within the Humanities the situation is worst than in the natural sciences, as the Open Access movement received only partial attention and little have been done for the cause (Di Donato). A point of peculiar interest is the small number of humanities-related open access journals if compared with its natural science counterpart. The reasons for the little adoption of open access journals could be identified in both the less founding the Humanities receives and the traditional resistance for the use of computers and the Internet as research tools.

*HyperJournal* (Hyperjournal) is a web application that facilitates the administration of academic journals on the Web. Conceived for researchers in the Humanities and designed according to an easy-to-use and elegant layout, it permits the installation, personalization, and administration of a dedicated Web site at extremely low cost and without the need for special IT-competence. *HyperJournal* can be used not only to establish an online version of an existing paper periodical, but also to create an entirely new, solely electronic journal. In comparison with existing software applications, *HyperJournal* introduces four major innovations:

1. *Dynamic contextualization* automatically transforms cross-references contained in journal articles into

hypertextual, bidirectional links. When the reader views an article published in *HyperJournal*, a contextualization bar provides immediate access to a) all the articles the author has cited, and b) all the articles that cite the article currently being viewed.

2. *The HyperJournal Network*. Dynamic contextualization is not limited to one journal only: it connects all the journals that use the *HyperJournal* software in a distributed, semantically structured and scaleable peer-to-peer network (Hyperjournal). Additionally, Compatibility with the *Protocol for Metadata Harvesting of the Open Archives Initiative* ensures maximal interoperability between the *HyperJournal Network* and other electronic publications. The *HyperJournal Network* thereby creates a space in which knowledge is freely shared and readily accessible. Rather than using mere keyword searching or importing artificial conceptual tables to organize this space, *HyperJournal* transposes the time-honored system of scholarly citation into an electronic environment.
3. *HyperJournals versus core journals*. By clicking on an author's name, the *HyperJournal* system automatically searches the entire *HyperJournal* network and produces a citation list that includes all the articles written by the author, all the articles the author has cited, and all the articles that cite the author. Comprehensive bibliometric lists can thereby be composed without the need to rely on the manual consultation of a small set of core journals, often exclusively in English. In this system, by contrast, it will be the actual give-and-take of academic discourse, registered automatically on the network through citations, which will signal the prestige of a journal (even of small niche journals written in so-called minor languages) and establish the reputation of scholars. In addition, through the use of (*Semantic Web*) RDF describers, bibliometric lists can be constructed that distinguish, for example, between positive and negative citations (Barbera and Di Donato).
4. *Structured vs. Opaque Formats*. Although *HyperJournal* let the editorial board choose which document formats are acceptable for submission, *HyperJournal* offers to the authors all the tools they need to use structured formats for writing their articles. The adoption of structured formats such as XML has enormous advantages over unstructured or opaque ones (such as *MS Word* or PDF) (Hockey). One of the major advantages is that structured formats are machine understandable thus perfectly suited to be used in conjunction with *Semantic Web* technologies. The most widely adopted structured format is undoubtedly *LATEX* which is wide spread within the scientific community. Unfortunately its usage within the Humanities is very limited. On one hand this is a disadvantage, on the other hand it leaves space for the diffusion of XML (who has even nicer computability properties than *LATEX*) as the format of choice. Initiatives such as TEI has already gained wide

acceptance among Humanities Scholars. TEI and other XML dialects such as DOCBOOK have the potential to be used directly to author articles, not only to encode existing texts (Piez). For this reason the *HyperJournal* developer's community is customizing and adapting some XML editors to facilitate the authors in their work. If the adoption of XML as a format for writing articles will be successful we can expect searches to be easier and much more powerful than today's heuristic search techniques and even to greatly reduce the cost of paper publication, as transforming XML to other formats suited for paper printing is a trivial task.

## Free access and respect for copyright: legal framework

*HyperJournal* aims at contributing to academic research on Internet publishing and encourages the birth of scholarly communities on the Internet. In order to achieve this goal, *HyperJournal* not only delivers IT solutions, but also tries to offer models for the independent organization and governance of scholarly communities, to develop systems for Internet peer-review, and to establish a legal framework for the free diffusion of knowledge on the Web that respects the principles of copyright. The documentation accompanying the software describes and comments on several models for the statutes of scholarly communities (the presence or absence of an Editorial Board; the constitution of the scholarly community by election or by other means; peer review and anonymity policies; criteria for publication; etc.), the administration of which is supported by the software. In addition to the licences provided by the Creative Commons initiative( <http://creativecommons.org/> ) *HyperJournal* contains three models of copyleft legal licenses (FreeKnowledge, OpenKnowledge, LimitedKnowledge <http://www.hypernetzsche.org/licenses/en/index.html> ) designed to reconcile the goal of open access to scholarly articles with the need to protect against plagiarism and to respect the moral right of the author.

## Founding and Distribution

The *HyperJournal Software* has been initially founded by the Groupement de Recherche Européen (GDREplus) *Hyper-Learning. Modèles ouverts de recherche et d'enseignement sur internet* which is a multidisciplinary research infrastructure promoted by the Centre National de la Recherche Scientifique (CNRS) regrouping 29 partners of 9 countries (universities and research centers, a large corporation (IBM), and three small enterprises). The software is currently being developed by both project members and volunteers and it is supported by *Dipartimento di scienze della politica*,

University of Pisa. *HyperJournal* is scaleable modular software distributed freely with an Open Source license. For these legal and technical reasons it is free to use and easy to modify and so can be adapted to the exigencies of a large number of scholarly communities. A prototype of *HyperJournal* has been released in February 2005.

## Bibliography

- arxiv.org*. Accessed 2005-03-21. <http://www.arxiv.org>
- Barbera, Michele, and Francesca Di Donato. "Open Access and Semantic Web. Software applications for Open Publishing." *Proceedings Semantic Web Applications and Perspectives (SWAP). 1st Italian Semantic Web Workshop*. Ed. Giovanni Tummarello and Christian Morbidoni. Ancona, Italy, 2004. 126-128. Accessed 2004. <http://eprints.rclis.org/archive/00002858/>
- creativecommons.org*. Accessed 2005-03-21. <http://creativecommons.org/>
- Di Donato, Francesca. "Verso uno "European Citation Index for the Humanities" Che cosa possono fare i ricercatori per la comunicazione scientifica." *Bollettino Telematico di Filosofia Politica* (2004). Accessed 2005-03-21. <http://bfp.sp.unipi.it/rete/ecih.html>
- Guédon, J.C. "In Oldenburg's Long Shadow: Librarians, Research Scientists, Publishers, and the Control of Scientific Publishing, Association of Research Libraries." *Proceedings of the 138th Annual Meeting*. 2001. Accessed 2005-03-21. <http://www.arl.org/arl/proceedings/138/guedon.html>
- hjournal.org*. Accessed 2005-03-21. <http://www.hjournal.org>
- Hockey, Susan. "The Robert Busa Award Lecture 2004." Paper delivered at the ALLC/ACH Conference 2004, Göteborg. 2004. video available at <http://www.hum.gu.se/allcach2004/>
- Hyper-Learning. Modèles ouverts de recherche et d'enseignement sur internet*. Groupement de Recherche Européen(GDREplus). Accessed 2004--03-21. <http://www.hyperl.org/>
- hypernetzsche.org*. Accessed 2004-12-20. <http://www.hypernetzsche.org/licenses/en/index.html>
- Piez, Wendell. "Authoring Scholarly Articles: TEI or Not TEI?" Paper delivered at the ALLC/ACH 2004 Conference, Göteborg. 2004. Accessed 2004. <http://www.hum.gu.se/allcach2004/AP/html/prop124.html>

*plos.org*. Accessed 2005-03-18. <<http://www.plos.org/>>

Suber, Peter. *Open Access Overview*. Accessed 2005-03-20. <<http://www.earlham.edu/~peters/fos/overview.htm>>

Tummarello, Giovanni, et al. "RDFGrowth, a P2P annotation exchange algorithm for scalable Semantic Web applications." *Proceedings of P2PKM 2004*. Boston, 2004. Accessed 2005-03-21. <[http://www.p2pkm.org/2004/Camera\\_Ready/1568938872.pdf](http://www.p2pkm.org/2004/Camera_Ready/1568938872.pdf)>

---

## Semantic Context Visualization to Promote Vocabulary Learning

---

**Caroline Barrière**

([Caroline.Barriere@nrc-cnrc.gc.ca](mailto:Caroline.Barriere@nrc-cnrc.gc.ca))

Conseil National de Recherche du Canada

**Claude St-Jacques**

([Claude.St-Jacques@nrc-cnrc.gc.ca](mailto:Claude.St-Jacques@nrc-cnrc.gc.ca))

Conseil National de Recherche du Canada

---

Traditional access through the alphabetically organized macrostructure (words defined) of the dictionary was convenient in a printed form. Online versions of dictionaries lead us to rethink our access approach and to exploit, as suggested by Humblé, the increased value of the dictionary in computer assisted language instruction. In a self-learning environment, a situation favored by today's wide-spread access to computers, an online dictionary is a valuable resource for reading comprehension. However, learners can quickly become discouraged if the information about a word searched for is buried among too much other information, e.g. the many definitions listed for highly polysemous words. Our current research suggests a method for providing specific guidance to a user to ease his access to information and promote vocabulary learning during his dictionary searches.

We present a tool, *REFLEX*, built on a mathematical model of a fuzzy logic search engine. We suggest that the microstructure (the content of the entries) of a dictionary be considered as a corpus. From this corpus, using fuzzy operators, we can calculate a similarity matrix, called a fuzzy pseudo-thesaurus (Miyamoto), expressing the degree of association between each pair of lemma (base forms of words) found in the corpus. This similarity calculation is based on the tendency of two words to co-occur within sentences. The fuzzy pseudo-thesaurus is pre-calculated and used at the search (query) time (Klir and Yuan).

Presently, the tool is for English learners of French and uses a learner's dictionary called *DAFLES* (*Dictionnaire de l'Apprenant du Français Langue Seconde et Étrangère*) (Verlinde et al.) merging all its defining sentences to build a corpus. Portability to languages other than French would require work at the pre-processing stage, for example to tokenize the sentences (split them into word units) and lemmatize the words (e.g. lemma should be found for diverse forms of nouns and verbs). Portability to other types of dictionaries would require



an understanding of the different types of information (definitions, examples, notes) contained before merging them into a corpus. Each type of information might be of different value to a learner, and then could be given more or less weight in our model.

*REFLEX* provides a graphical visualization of a word's semantic context. Figure 1 shows the portion of the pseudo-thesaurus relating to the word "scientifique" (scientific) as automatically built from *DAFLES*. The surrounding words (*chercheur* – researcher, *étudier* – study, *observation* – observation, *théorique* – theoretical, *rechercher* – research, etc.), express words related to *scientifique*, their graphical distance being proportional to their calculated distance.

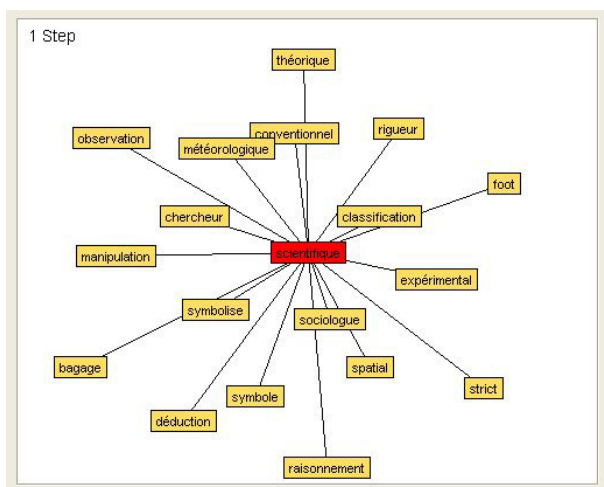


Figure 1: Semantic context for scientifique

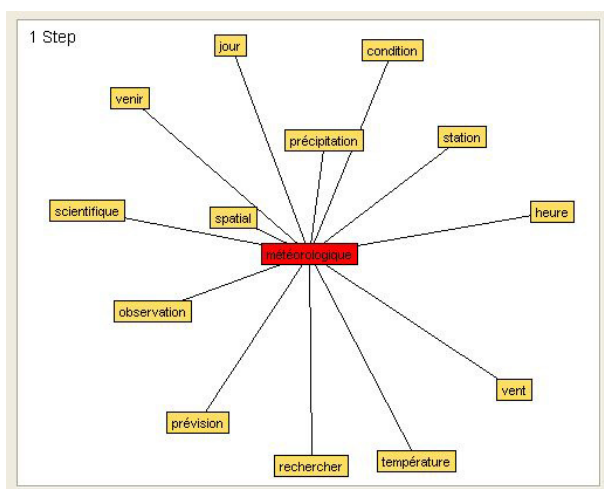


Figure 2: Semantic context for météorologique (meteorological)

Semantic contexts give a graphical representation similar to semantic maps which are widely used in language teaching to favor vocabulary acquisition (Brown and Perry; Carrell; Crow and Quigley). Semantic maps are usually provided in classroom settings and are manually constructed by instructors

interactively with the students. Here semantic contexts are produced in real-time, and *REFLEX* provides navigation capabilities for the user to easily move to other portions of the pseudo-thesaurus. For example, a click of a mouse on the word *météorologique* within the semantic context of Figure 1 brings the learner to the semantic context shown in Figure 2.

The implicit relations shown through the arcs linking the concepts are not limited to paradigmatic relations as found in *WordNet* (Fellbaum) or *MindNet* (Richardson et al.)<sup>1</sup> For example, the association between *observation* and *researcher* in Figure 1 could not be labeled with a paradigmatic relation. The complex relation, *typical activity*, is closer to a lexical function as defined by Mel'cuk.

*REFLEX* also serves as a guide for dictionary searches by making use of information found in the context of occurrence of the unknown word. It leads the learner not necessarily to the entry corresponding to that unknown word, but to any information from dictionary entries likely to help the understanding of the word within that particular context. This provides a clear filter for polysemous words and helps the learner identify among all the possible definitions and examples, the ones most relevant to their current reading situation.

Let us take for example the French polysemous word *culture*, which could relate to *cultivate* (flowers or vegetables) or *culture* (social knowledge). The following sentence is taken from a text "L'école dans les deux langues" shown in a reading comprehension software (*DidaLect*, Duquette et al.): "En étudiant la langue et la culture de l'autre, Anglo-Saxons et Latinos apprennent en même temps à se connaître et à se respecter."

Assuming the word *culture* is unknown to a learner, she launches a search. The context window of occurrence of that word is analyzed to obtain a query vector made up of the searched word and its neighbours (*culture*, *étudier*, *langue*, *apprendre*, *connaître*, *respecter*)<sup>2</sup>. This query vector is expanded on via the fuzzy pseudo-thesaurus to include related words, each with a weight corresponding to their similarity to the original query vector (*culture* 1.0, *étudier* 1.0, *langue* 1.0, *apprendre* 1.0, *connaître* 1.0, *respecter* 1.0, *didactique* 0.1, *pédagogique* 0.09, *éducatif* 0.09, *créatif* 0.08, *race* 0.08, *impoli* 0.06, *curiosité* 0.06, *brassage* 0.06, *fermier* 0.06, *formateur* 0.06)<sup>3</sup>. The enlarged vector then represents an extended context of a word, and is used in an information retrieval task to extract interesting definitions and examples pertaining to that context from the corpus of sentences. These sentences are presented to the user in decreasing order of relevance, as shown in Table 1 (relevance is given in first column), and come from various polysemous entries.

Rel.	Sentence	Entry	Polysemy of that entry
0.6	<i>Lorsqu'une personne étrangère s'assimile (à une société, une culture), elle s'intègre dans cette société, dans cette culture, elle adopte...</i>	<i>assimiler</i>	3 other meanings (confound, understand, group)
0.4	<i>Le brassage d'idées, de plusieurs choses, personnes, cultures ou races est le mélange ou la combinaison d'idées, de plusieurs choses, personnes...</i>	<i>brassage</i>	1 other meaning (beer brewing)
0.4	<i>Lorsqu'un élève, un étudiant révise un cours, il étudie, il parcourt à nouveau un cours qu'il a déjà appris.</i>	<i>réviser</i>	2 other meanings (reviewing a text inspect a vehicle)
0.4	<i>Lorsque quelque chose appartient à un endroit, à une période, à une culture, à quelque chose, cette chose est caractéristique de cet endroit, ...</i>	<i>appartenir</i>	5 other meanings (belong to a group, own something, have to do something, have responsibility of something, part of a machine)

Table 1 – Relevant sentences for the understanding of the word culture in context

Using the extended context of a word, combined with knowledge of word association as pre-calculated in the pseudo-thesaurus, *REFLEX* gathers the distributed information from the dictionary to help the understanding of a word (often polysemous) as expressed in a specific reading context.

Yet, free browsing is not necessarily a bad thing for learners. In fact, researchers in second language learning (Aston) debate on the usefulness of corpora in which learners could be left to wander among sentences by themselves. On one hand, free browsing mode encourages autonomous discovery of different or new senses of words, but on the other hand, too much discovery within a reading comprehension task might take the learner too far from the original goal. Furthermore, both activities (free or guided exploration) are more or less appropriate depending on the learner's profile (efficient or less efficient learner).

Although intended to guide, *REFLEX* is easily adaptable toward a discovery mode by simply removing the contextual cues from the query vector. The flexibility of *REFLEX* can also be seen in the presentation of the semantic maps. For efficient learners, larger maps can be shown, such as presented in Figure 1, but for less efficient learners, maps can be restricted to only contain the few closest words.

In conclusion, *REFLEX* shows much potential for online dictionary exploration. It is constructed on sound mathematical principles and has potential for adaptability to different learning purposes and types of learners. As our next step, we will integrate *REFLEX* as a module within *DidaLect*, a reading comprehension software, with focus on adaptability parameters. Investigation into automatic labeling of associations is an ambitious longer term research goal. Finally, we hope for future explorations with other dictionaries and other languages.

1. Paradigmatic relations, such as hyperonymy, meronymy, synonymy, are of great value in language learning and do provide an explicit way for learners to organize the new vocabulary in their own knowledge stores. The purpose of semantic contexts is complementary to this organization by providing a semantic field around a word.
2. Lemmatized forms of content common nouns are taken. We do not use proper nouns (they are not in the dictionary) or function words.
3. We only show values larger than 0.06, but this threshold could be adjusted.

## Bibliography

- Aston, G. *Learning with Corpora*. Houston: Athelstan, 2003.
- Brown, T.S., and F.L. Perry Jr. "A comparison of three learning strategies for ESL vocabulary acquisition." *TESOL Quarterly* 25 (1991): 655-670.
- Carrell, P.L. "Content and Formal schemata in ESL reading." *TESOL Quarterly* 21 (1987): 461-481.
- Crow, J.T., and J.R. Quigley. "A semantic field approach to passive vocabulary acquisition for reading comprehension." *TESOL Quarterly* 19 (1985): 497-513.
- Duquette, L., A. Desrochers, and S. Szpakowicz. "Adaptive Courseware for Reading Comprehension in French as a Second Language : The Challenges of Multidisciplinarity in CALL." *Proceedings of the eleventh International CALL conference, University of Antwerp*. Antwerp, 5-7 September 2004. 85-92.
- Fellbaum, C. *WordNet: An Electronic Lexical Database*. Cambridge, Mass: MIT Press, 1998.
- Humblé, P. *Dictionaries and Language Learners*. Frankfurt am Main, Germany: Haag + Herchen Verlag GmbH., 2001.
- Klir, G.J., and B. Yuan. *Fuzzy Sets and Fuzzy Logic*. Upper Saddle River, NJ: Prentice Hall, 1990.
- Mel'cuk, I. "Semantic description of lexical units in an explanatory combinatorial dictionary: Basic principles and

heuristic criteria." *International Journal of Lexicography* 1.3 (1988): 165-188.

Miyamoto, S. *Fuzzy Sets in Information Retrieval and Cluster Analysis*. Dordrecht, Netherlands: Kluwer Academic Publishers, 1990.

Richardson, S.D., W.B. Dolan, and L. Vanderwende. "MindNet: Acquiring and Structuring Semantic Information from Text." *Proceedings of the ACL'98*. Montreal, 1998. 1098-1102.

St-Jacques, C., and C. Barrière. "L'inférence dictionnaire : de la créativité poétique à celle du raisonnement flou." *Cahiers de lexicologie* 85 (2004): 129-155.

Verlinde, S., , and GRELEP (Groupe de Recherche en Lexicographie Pédagogique). *Dafles (Dictionnaire d'apprentissage du français langue étrangère ou seconde)*. . Accessed 2005-04-11. <<http://www.kuleuven.ac.be/dafles/acces.php?id=>>

---

## Playing Many Parts: Models of Collaboration in an Electronic Edition

---

**Michael Best** ([mbest1@uvic.ca](mailto:mbest1@uvic.ca))

University of Victoria

**Jessica Slights** ([jessica.slights@acadiiau.ca](mailto:jessica.slights@acadiiau.ca))

Acadia University

**Peter van Hardenberg** ([pvh@uvic.ca](mailto:pvh@uvic.ca))

University of Victoria

**Wendy Huot** ([wendy.huot@gmail.com](mailto:wendy.huot@gmail.com))

University of British Columbia

**Alan Gale** ([agaley@uwo.ca](mailto:agaley@uwo.ca))

University of Western Ontario

---

### Panel Abstract

**E**ditorial theory has in recent years been much concerned with re-evaluating the assumptions that lie behind the practice of editing, in light of the arguments promulgated by McGann and others that the text we edit is of necessity constructed through social interaction. A parallel argument makes the claim that the electronic medium is uniquely situated to put into practice the claims of the theorists (exemplary articles are included in the collection edited by Landow). While creative collaboration is not a norm in the humanities, it is very much at the centre of disciplines in the performing arts: drama, in particular, involves the interaction of many kinds of creative endeavour before a play is staged. The *Internet Shakespeare Editions* are uniquely in a position to explore varieties of collaboration that borrow from other disciplines: the open source movement in software development, and the well-developed traditions of collective creativity in the performance arts, since the concept of 'text' in a play by Shakespeare can - and arguably should - be extended to include the history of his plays in performance. This panel will provide a forum to discuss the electronic edition as an exemplar of the creation of a social/performance text as the various contributors to an edition interact and learn from each other.

Many of the collaborative activities in the preparation of an edition for a print press are so familiar as to be transparent: peer review, copy editing, data entry, printing, binding,

distributing and so on. The print interface is also very stable, with minimal experimentation with the appearance of the print on the page (but see Jowett, Grusin and Bolter). Those working in the electronic medium, however, are still experimenting with ways of displaying the text and its associated navigational structures. The result is that the lonely textual scholar will ideally become involved in the process of creating much more than simply a word-processor file to send off to the publisher. The team of technical designers and programmers are unlikely to understand what the editor sees as important or what a typical user of the edition will be seeking, and the textual scholar is unlikely to know what opportunities the full capabilities of the medium offer: thus the most effective edition will be the result of a deep and interdisciplinary collaboration between the various creators.

The *Internet Shakespeare Editions* is a collaboration in many ways. An academic Editorial Board oversees general standards; plays are edited by individual editors, or (in several cases) further collaboration between scholars; the General Textual Editor works with the editors to ensure quality and consistency; and technical experts work on separate but interrelated tasks - the development of the XML structure that is the basis for textual materials on the site, the creation of databases to display images of the texts and of performance materials, the graphic design of the site, and the stylesheets that determine the overall look of both static and dynamic pages.

The analogy with the performing arts is persuasive, since they typically involve a similar interaction between specialists (set design, costume design, choreography, lighting), creative artists (actors, the author perhaps) and administrative/creative personnel (director, producer). Publishers and actors are similarly concerned to reach and engage their audiences. Where the analogy fails, however, is in the contrast between the maturity of the theatre community and the inexperience of those of us working to create electronic texts. Much in the way that the print medium has evolved transparent signposts for navigation within the book, theatre companies have evolved generally consistent and effective structures to manage the collaborative endeavour of producing a play. The challenge that lies ahead for organizations like the *Internet Shakespeare Editions* is to develop similarly coherent and effective management of collaboration. How can we best ensure that scholars are not intimidated by technical demands, and that the programmers are aware of the full potential of the materials they are responsible for displaying?

As well as providing a potential model for collaboration in the creation of online texts, the analogy with stage performance also provides a warning. Performance (unless on film) is evanescent, leaving traces only in reviews and theatre archives. Methods for preserving print are highly developed, and there is every reason to expect that a book published today will

survive for several hundred years if it is judged worth the storage space. But techniques for archiving and for ensuring the permanence of electronic data are also evolving as file formats and the medium itself evolve.

In a recent essay, W.B. Worthen asks

If we take the printform of a work to be like a performance, materializing a historically contingent, socially inscribed instance of the work . . . we may be able to seize a more dynamic sense of the changing interplay between these two enduring, and volatile, modes of production.

This panel will involve a representative group of those working on different activities within the structure of the *Internet Shakespeare Editions* as they explore the interplay between multiple modes of production.

Panel members will give short presentations (maximum ten minutes), in order to leave sufficient time for questions both within the panel and from the floor.

**Michael Best**, Coordinating Editor of the *Internet Shakespeare Editions*, will chair the panel, and begin by providing an overview of the academic and administrative structure of the *ISE*. He will also discuss the audience for which the editions are designed, and the potential of content management software and other software packages that facilitate collaboration.

**Jessica Slights**, one of two editors collaborating on *Othello*, will discuss the challenges facing scholars trained in literary studies as they work with the additional demands made by the electronic text, both in the sheer quantity of material the electronic space makes available, and in dealing with the demands of preparing a text in e-format.

**Peter van Hardenberg** will discuss the need for the XML structures chosen for the texts to reflect adequately the needs of computing humanists.

**Wendy Huot** is responsible for designing the databases for images and performance materials that will be integral to the editions. Her presentation will explore strategies for developing, in consultation, an overall data model that can organize and interrelate a wide variety of binary and textual objects.

**Alan Galey**, an editor with extensive experience in programming, will discuss current thinking on archiving and long-term preservation of electronic artifacts.

### **A Kind of Yeasty Collection: Organizing Collaboration in the Internet Shakespeare Editions**

**Michael Best**

In 1996, when I first began work on the *Internet Shakespeare Editions*, I chose to describe my role as that of 'Coordinating

Editor' rather than the more usual 'General Editor'. Coordination and collaboration are necessarily at the centre of the development of a major scholarly site that involves both academic and technical organization. Academically, there is an Editorial Board to oversee general standards, a General Textual Editor to crack the whip, and an extensive team of editors, since each play is being edited by one or more scholars. On the technical side the demands are no less complex. Gone are the days that a scholar can, as a sideline, whip up some reasonably effective HTML code and publish it; Web users now expect attractive, professionally designed pages, intuitive navigation, and full searching capabilities. In addition, sites are increasingly being generated from centralized relational databases requiring sophisticated programming skills. All this costs money, and it is still true, in Canada at any rate, that granting agencies tend to be frugal in apportioning funds for what seem more like computer science than Humanities activities. One solution is to apply for funding for student assistants — coop positions, MA or PhD fellowships and so on — since these can be more readily justified as academic expenses than the fees of professional programmers.

The challenge thus becomes the management of two teams, largely separate in their activities, but requiring collaboration before the editors' texts can be effectively displayed by the structures created by the technical team. In his paper, Peter van Hardenberg will discuss the problems that he faces in creating an effective XML schema for the complex documents the editors produce as they transcribe, modernize, collate, and annotate texts that often have multiple origins; Wendy Huot will discuss the process by which she is designing a database that will respond flexibly to a wide range of textual and multimedia artifacts created by the performance of Shakespeare. On the academic side, Jessica Slights will discuss the learning curve that a textual scholar faces when preparing materials for electronic publication, and Alan Galey will bring the perspective of that rare bird, a textual scholar who is also familiar with programming, to the discussion of the long-term viability of the texts we are publishing.

The technical team also requires a high level of internal collaboration. Academic editors are used to working on their own, but the various activities of the programmers — designing templates for a consistent look, developing XML, and designing the database — are deeply interconnected. There are also important external consultants on graphic and interface design. To facilitate collaboration, we are using standard processes — email lists, a *Yahoo* discussion group, and conference calls — but we are also looking at possibilities for software solutions: a content management system that would permit flexible access to the site, at the same time as making networks of communication available to those working on parallel projects.

## **Made Tame and Most Familiar: Adapting to the Medium of the E-text**

**Jessica Slights**

This paper will address the topic of editorial collaboration from the perspective of a literary scholar trained in Shakespeare studies and now facing the multiple challenges of helping to prepare an electronic edition of *Othello* for a modern readership. The paper will argue, following Jeffrey Masten, that collaboration was a prevalent model of textual production during Shakespeare's lifetime, and that it is therefore a particularly appropriate model for us to adopt in the 21st century as we move to bring his plays to new reading audiences. I will offer my own experiences in collaborative editing as a test case for this claim since the *Othello* edition involves not simply the traditional challenges of editorial partnership, but also a commitment to new technologies that require a degree of teamwork with which most scholars in the humanities are probably unfamiliar.

## **Hierarchies and Treespaces: A Proposed Extension of XML**

**Peter van Hardenberg**

Those in the Humanities Computing community will be very much aware of the limitations of XML as a markup language for working with literary texts, especially in its awkwardness in dealing with documents that require some method of encoding overlapping hierarchies. As is the case in many other projects, texts prepared for the *Internet Shakespeare Editions* must at the very least record hierarchies that represent the separate conceptual and physical divisions of the text. A third independent hierarchy, tentatively labelled 'annotational' would also simplify much of the difficulty in describing the complexities of textual variants and other scholarly apparatus. This paper will propose a solution to this problem through a simple extension to the standard format of XML documents.

The necessary flexibility is accomplished by relaxing XML's requirement that all element tags must form a single tree, to the requirement that nesting must only be preserved within a given *treospace*. A document may have multiple treespaces, each with its own DTD or schema. Immediate advantages include elegant decomposition of complex DTDs into modular component DTDs, and the abolishment of stopgap tricks like span tags. This change has many ramifications and consequences to explore, including validation techniques, combining documents, extensions to the *DOM* (Document Object Model), and backwards compatibility with vanilla XML formats. Application domains are not limited to the text encoding community and could potentially be realised in many fields including bioinformatics, word processing file formats,

and any other field where tagging applies to a stream of character data.

## Communicating with the Ivory Tower: Modeling Humanities Multimedia Data

Wendy Huot

Creating a data model for Humanities multimedia and data requires scholarly understanding of the content itself and a technical understanding of database design. This can lead to collaboration between humanities scholars and database technicians; experts that may be largely ignorant of the other's realm.

The major communication challenge in such a collaboration is defining the functional requirements of the data model. Misunderstandings abound due to a lack of shared terminology, the scholar's unfamiliarity with the needs and capabilities of the technology, and the technician's ignorance of the significant characteristics of the data to be modeled. Special cases and extremes within the data provide for a difficult interdependency: the technician may need to be warned of problematic special cases by the scholar, but only the technician may understand what kinds of special cases are problematic.

The *Internet Shakespeare Edition's* development of a data model (and resulting database) for text and multimedia performance materials inspired some strategies for managing collaboration. These strategies included describing data content with samples in hand, early development of speculative designs, and conducting technical discussion using media — such as instant messaging — that record a text log of the conversation for later review.

## Collaborating with the Future: Shakespeare and Preservation

Alan Galey

"Thy easy numbers flow," writes John Milton in an early poetic commentary on the preservation and transmission of Shakespeare's works: "each heart / Hath from the leaves of thy unvalued book, / Those Delphic lines with deep impression took" (10-2). Unusual in its praise for print as a preservation format, Milton's poem prefaces the first reprinting of Shakespeare's first chief textual archive, the collection of the plays in folio (1623, rpt. in 1632, 1663-4, and 1685). Since the advent of electronic Shakespearean editing and text analysis, Milton's words have acquired unintended resonance. As encoded alphanumeric data, Shakespeare's works easily flow into new digital forms and objects of analysis. (Indeed, Shakespearean compatibility with hypermedia is almost a truism now.) But Milton's poem also serves to remind projects like the *ISE* of two vital points: that a distinctly Shakespearean subculture of

textual archiving predates us by centuries; and that even in 1632 this subculture had articulated the importance of both collaboration and remediation.

This brief paper will attempt to bridge the distance between digital preservation and software longevity practices, on the one hand, and Shakespearean editing and textual studies, on the other. Both traditions of thought bear upon the *ISE's* electronic transcriptions of plays and poems from the 1623 Folio and the early quartos. In developing encoding strategies for these complex, historically loaded texts, the *ISE* must weigh present software needs against future interoperability, and *TEI* compliance against editorial responsibility. As Milton understood, Shakespeare has no perfect archive for "transcendental data" (as Alan Liu terms it); his works persist only through renewal and collaboration with generations - and encoding formats - yet unknown.

## Bibliography

- Grusin, Richard, and J. David Bolter. *Remediation: Understanding New Media*. Cambridge, MA: MIT Press, 1999.
- Jowett, John. "Addressing Adaptation: Measure for Measure and Sir Thomas More." *Textual Performances*. Ed. Lucas Erne and Margaret Jane Kidnie. Cambridge: Cambridge University Press, 2004. 63-76.
- Landow, George P., ed. *Hyper/Text/Theory*. Baltimore, MD: Johns Hopkins University Press, 1994.
- Liu, Alan. "Transcendental Data: Toward a Cultural History and Aesthetics of the New Encoded Discourse." *Critical Enquiry* 31 (2004): 49-81.
- McGann, Jerome J. *A Critique of Modern Textual Criticism*. Chicago: University of Chicago Press, 1983.
- McGann, Jerome J. *The Textual Condition*. Princeton, NJ: Princeton University Press, 1991.
- McGann, Jerome J. *The Rationale of Hypertext*. Institute for Advanced Technology, University of Virginia, 6 May 1995. Accessed 2005-03-23. <<http://www.iath.virginia.edu/public/jjm2f/rationale.html>>
- McGann, Jerome J. *Textonics: Literary and Cultural Studies in a Quantum World*. National Humanities Center, Revised: October 1995. Accessed 2005-03-23. <<http://www.nhc.rtp.nc.us:8080/newsrel2002/mcgannwebcast.htm>>
- Milton, John. "On Shakespeare. 1630." *The Complete Poems*. Ed. John Leonard. London: Penguin, 1998. 19.

Worthen, W.B. "The Imprint of Performance." *Theorizing Practice: Redefining Theatre History*. Ed. W.B. Worthen and Peter Holland. Basingstoke and New York: Palgrave Macmillan, 2003. 213-234.

## ***Understanding Poetry Online: an Internet Application for Teaching***

---

***Jonathan Blake*** ([jonathan.blake@vanderbilt.edu](mailto:jonathan.blake@vanderbilt.edu))  
*Vanderbilt University*

---

**T**his poster session offers participants a thorough overview of the genesis, process, development, and utilization of *Understanding Poetry Online*, an application for teaching formal and analytical approaches to verse. Developed at Vanderbilt University, this application offers instructors and students new and different ways of approaching fundamental issues in college-level instruction in poetry including the ability to create, share, and deliver, and take lessons online. Participants will learn how this application originated, how it has been supported and developed, and how it continues to be used, assessed, and refined. In this way, the session will offer both an introduction to a useful new technology, and a discussion of how such technologies can be developed.

The teaching of poetry to undergraduates has two components: training students to recognize basic poetic forms, techniques, meters, and rhymes, and helping students to develop and defend thematic observations about the poetry. The challenge to teachers at all levels is to integrate these two aspects of teaching, so that students can use their acquired knowledge of scansion, rhyme, imagery, and so forth to deepen and enrich their understanding of the themes and contexts of the poetry. We have designed the *Understanding Poetry Online* application to help train students in both aspects of poetic analysis and especially to integrate those aspects.

The *Understanding Poetry Online* project has elucidated students' learning processes and progress by providing a rich, interactive multimedia platform for the acquisition and application of both the technical and analytical reading of poetry at the university level. Our *Sonnet Scansion and Analysis* program is a web-based interface for designing and implementing curricula that facilitate assessment of students' learning of critical poetry reading skills in the following ways:

- providing an online, on-demand environment for reviewing and acquiring the skills of metrical scansion, rhyme scheme analysis, and analytical reasoning;
- making the processes of reading and analysis visible with on-screen, color-coded, computer-assisted poetic scansion tools;

- providing a facility for user-stored and instructor/user-retrievable in-depth analyses of poetic forms and content;
- creating links from the online environment to the classroom and student study groups with printable versions of all online activities and work;
- encouraging students to integrate their online work with essays they are writing for classes.

The application continues to evolve, and has grown from a basic web-page into a fully-automated instructional technology.

---

## 'La Imaginación al Servicio de la Educación': un Ejemplo de *Work in Progress*

---

*Laura Borràs (lborras@uoc.edu)*

*UOC/Hermeneia*

*Isabel Clara Moll Soldevila (imoll@uoc.edu)*

*Universitat Oberta de Catalunya*

*Roger Canadell (rcanadell@uoc.edu)*

*Universitat Oberta de Catalunya*

---

## La imaginación al servicio de la educación

**E**ste conocido lema del Rector de la Universitat Oberta de Catalunya, que le llevó a impulsar una aventura universitaria que este año cumple 10 años de existencia, ha sido una y otra vez codiciado, explotado y reinventado por cuantos trabajamos en esta universidad. Mediante este artículo nos gustaría compartir el proceso de elaboración de la asignatura *Estudios Literarios y Tecnologías Digitales* que ha de impartirse por vez primera el próximo curso 2005-2006 y sobre la que hemos estado trabajando desde el año 2002. Se trata, por tanto, de mostrar una parte del proceso de creación de una asignatura optativa que queríamos que permitiera a nuestros estudiantes —que presuponemos mínimamente alfabetizados digitalmente teniendo en cuenta que estudian en una universidad enteramente virtual— ejercer de críticos y poner en práctica las herramientas teórico-críticas que les han sido transmitidas en otras asignaturas de sus estudios. Nuestro propósito inicial, pues, consistía en tratar de forzar al estudiante a ubicarse en la nueva dimensión creativa que nos ofrecen las tecnologías digitales y penetrar, a fondo y mediante un completo ejercicio crítico que fuera más allá de la habitual crítica formalista y descriptiva de estas nuevas obras, en las dificultades y particularidades de la literatura electrónica.

Estos objetivos iniciales suponían también una serie de estreñimientos. Debía ser una asignatura eminentemente práctica: había que leer literatura digital. Pero en este caso teníamos otro problema previo considerable en la medida que las obras de creación auténticamente digital existentes en la red (la mayoría en inglés o francés) desafían la competencia lingüística de la mayoría de nuestros estudiantes. Por ello hemos



diseñado una obra de literatura digital con distintas posibilidades de navegación contribuyendo, de este modo, a la salud creativa de la literatura digital en lengua catalana y española (existen dos versiones lingüísticas del cibertexto) a la par que fomentando una consideración temática alrededor de algunos de los problemas que están concentrando la atención de la crítica en los últimos tiempos.

El *Diario de una ausencia* es un proyecto surgido de los resultados de un congreso internacional organizado por el grupo de investigación *HERMENEIA*, que estudia las confluencias entre los estudios literarios y las tecnologías digitales, celebrado en Barcelona en abril de 2004, uno de cuyos resultados es este producto literario digital concebido para un uso didáctico y en red (otro de los resultados que ya puede consultarse en red es la edición digital de un monográfico de la prestigiosa revista internacional *Dichtung Digital*: <http://www.brown.edu/Research/dichtung-digital/english.htm>).

Nuestro *Diario de una ausencia* se presenta como un producto eminentemente textual, ubicado en una dimensión visual y musical determinante, que el lector puede recorrer como si de un peregrinaje se tratara. Hemos optado por un uso de la tecnología con una finalidad estética, narrativa, semiótica y hermenéutica: no se hallará aquí una eclosión de los medios más avanzados como escaparate tecnológico sin otra finalidad que la del mero lucimiento de recursos, sino una convergencia de medios informáticos al servicio de un artefacto estético digital. Su naturaleza fundamentalmente hipertextual evoca el aspecto más secreto e íntimo del hipertexto —un dato que todavía cobra más relevancia teniendo en cuenta el componente de confesión, de autoconfesión que lleva implícito un diario privado. Sin embargo, no cabe llevarse a engaño: el *Diario* es algo más que un hipertexto, cuando menos no es un hipertexto 'puro'. Lo cierto es que se combinan las posibilidades creativas sinérgicas de lenguajes artísticos como la imagen o la música. Se trata de un producto múltiple y complejo por cuanto el texto se ve apostillado por los comentarios de un lector 'modelo' —en este caso otro profesor de la UOC— que lo atraviesan y lo penetran a la búsqueda de significado. La mirada crítica de un ojo externo, que se sumerge en esta crónica del dolor y de la ausencia, se concreta en un primer ejercicio interpretativo que emplaza a sus lectores a pronunciarse, al tiempo que ofrece posibles claves de lectura sobre las que discutir.

## Un espacio híbrido y fecundo

**Y** todo ello sucede en la inasible virtualidad de una pantalla. En este entorno creativo, culminada ya la fase inicial -necesaria e inevitable- de descubrimiento de un nuevo medio; superado pues un tiempo en el que aprender a habitar en una nueva dimensión textual, poco los creadores han ido

desarrollando las potencialidades que son propias del medio. Así, la conjunción de códigos informáticos y de lenguaje, la colaboración entre escritores y programadores ha dado a luz una retórica, una gramática y una sintaxis específicas del entorno digital que, de algún modo —como el *Diario* también pretende ejemplificar—, ha transformado también lo que representa el acto de lectura y la consideración de lo literario.

## Niveles de navegación: a la conquista de un territorio íntimo

**L**a disposición espacial de los textos digitales despierta un conjunto de metáforas que hacen posible la lectura bajo un nuevo prisma, el de la navegación. En este sentido, el lector que accede al *Diario* percibe con nitidez que la casa es un territorio por explorar donde las palabras actúan como reclamo a la vez que funcionan como capciosas contraseñas que nos permiten acceder a un texto que se oculta, que se esconde porque secreto es su contenido y como un secreto le es confiado al lector. La idea central de esta propuesta metodológica es la de concebir la lectura como una *queste*, como una búsqueda. Así, del mismo modo que en los romances medievales la búsqueda es la organizadora del relato y la *queste* es, por lo tanto, el esquema narrativo a seguir, podemos llegar a establecer una identidad absoluta entre búsqueda y organización de los contenidos en el espacio digital. El *Diario de una ausencia* presenta cinco modos de acercamiento posible al texto y, en consecuencia, cinco posibilidades distintas de lectura.

1. En primer lugar está la inmersión explorativa en la casa-escenario que nos ofrece, por estancias, palabras que actúan de contraseña y puerta de acceso a los distintos días del texto. Es ésta una vía donde la curiosidad y el azar de las elecciones nos ubican ante palabras que ejercen de reclamo para permitir que aflore el texto que esconden y al que se refieren, al que pertenecen, en definitiva, porque de él han salido.
2. En segundo lugar, está la posibilidad de realizar una lectura numérica, buscando en el 'mapa del tesoro' los números que permiten una lectura cronológica de las páginas del diario.
3. Teniendo en cuenta de que cada página del diario se corresponde con un día del mes, también es posible efectuar una tercera lectura a partir de los textos que ya han sido descubiertos/revelados en el orden 'lineal' del calendario, eso sí, recordando que hasta que no haya sido desvelado el contenido de una palabra no será visible la correspondencia entre la palabra seleccionada y el día del mes al que hace referencia.
4. Otra vía de lectura posible es la que permite consultar el mapa o plano de la casa y sus inmediateces y ver qué

objeto conlleva una vía de acceso al texto y dónde está ubicado.

5. Por último, también el mapa ubica las distintas músicas que conforman esta dolorosa melodía en el espacio y uno puede dejarse llevar por lo que la música le sugiere para, a partir de la música evocada ir al texto correspondiente. Palabras, números, objetos, músicas y azar son los médiums de este particular viaje a las tinieblas del alma.

## Una nueva experiencia de lectura

**E**n esta tesitura, queremos que surja la pregunta ¿dónde está la lectura? O mejor ¿dónde está el texto que ha de leerse si éste es invisible y no aparece más que a partir de una interacción —mínima, pero interacción al fin y al cabo— como es una elección, un clic, el establecimiento de un itinerario? ¿Cómo se lee una obra como esta, sin texto —a primera vista—, sin un orden aparente, sin páginas que pasar, sin ver en qué punto del relato estamos, sin saber cuánto nos falta para llegar al final, si es que lo hay? ¿Será la lectura algo susceptible de ser compartido? ¿Habrán leído todo el mundo el mismo *Diario* o será la lectura de cada cual forzosamente distinta? ¿Incluso uno mismo puede leerlo de distintas maneras en distintas ocasiones? ¿Hasta qué punto divergirá la lectura lineal, la lectura más 'literaria' y menos 'lúdico-explorativa' del *Diario de una ausencia*, la lectura que también es posible realizar en papel de las 'lecturas' fragmentadas y aleatorias de páginas sueltas de un diario? ¿Representará la elección de la música, del ambiente paisajístico en el que se hallan las palabras, los objetos y los textos, la lectura digital, en definitiva, un añadido, una experiencia estética diferente a la lectura analógica y convencional del *Diario* como un diario? ¿Qué modo de lectura emerge en un contexto tan cambiante, con tantos sentidos implicados, con tantas posibilidades de elección para el lector? Éstas y otras muchas preguntas sólo podrán ser contestadas por los estudiantes después de enfrentarse con el *Diario*. Son posibles puntos de reflexión para un diálogo abierto respecto de una materia que se halla en proceso de construcción.

---

## Supporting Annotation as a Scholarly Tool: Experiences from the *Online Chopin Variorum Edition*

---

*John Bradley* ([john.bradley@kcl.ac.uk](mailto:john.bradley@kcl.ac.uk))

*King's College London*

*Paul Vetch* ([paul.vetch@kcl.ac.uk](mailto:paul.vetch@kcl.ac.uk))

*King's College London*

---

**M**embers of the Centre for Computing in the Humanities, King's College London, in partnership with researchers at Royal Holloway College, have been investigating the use of some text/image comparison enhancement technologies for the creation of an *Online Chopin Variorum Edition (OCVE)*. The project's primary research goal is to explore how one might provide improved support for the comparative scholarly analysis of (in this case musical) source materials (manuscripts, first impressions of first editions, and later impressions which often contain variants), using as a basis the music of the famous 19th century composer Frédéric Chopin.

*OCVE* provides access to images of the music directly, rather than to symbolic representations of that music. In this light, it investigated three kinds of image manipulation:

1. superimposition: *OCVE* looked at the laying of images from two defined filiation chains over top of each other in order to reveal variants,
2. juxtaposition: *OCVE* explored the provision of tools to facilitate the comparison of variants on a bar-by-bar basis, and
3. combination/interpolation: *OCVE* considered the utility of allowing users to create purposeful collations assembled from the variants.

Figures 1 and 2 illustrate superimposition and juxtaposition in the context of the *OCVE* project.

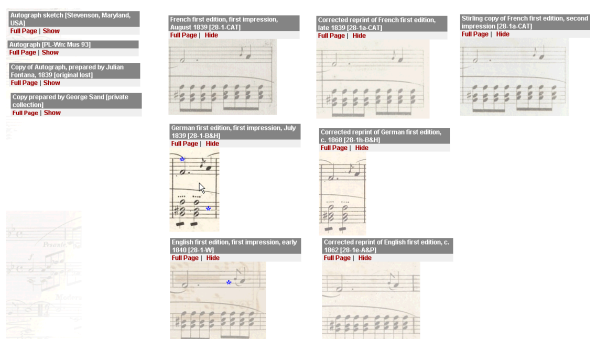


Figure 1: OCVE Juxtaposition view

### Overlay of E<sup>1</sup> on c.1868 reprint at 50% opacity (both scores visible)

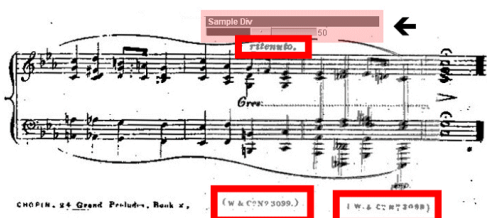


Figure 2: OCVE Superimposition

At first glance perhaps *OCVE* looks like another digital archive project, but a part of the goal of the project was to probe the edges of established digital archive practice to see how that model might be extended and made more useful to scholarship.

Digital library and archive research has been underway in information and computing science for more than ten years now, and started full of promise about the expected benefits to scholars. After ten years, however, it would seem that the humanities should be reaping these expected benefits that were set out with such enthusiasm 10 years ago. Indeed, of course there *are* clear tangible benefits from using digital archives and similar resources. Note, for example, a recent report in the *AHDS Newsletter* about the launch of the Early English Books Online (EEBO) resource as a cheaply available resource for colleges and universities in the UK. There a reader in History of Early Modern Ideas at Royal Holloway (University of London) is reported as saying

Once you get a taste of what research can be like with EEBO you want more. It transforms how you work. I can work at 2am. I can scribble on my printouts. I'm not restricted by library opening times. I've cut my transport costs and time. It's simply more efficient.

(Leon)

Clearly, the potential of simply making rare material more or less instantly accessible on demand is a clear benefit, and one

that one hopes will be available to more and more researchers over time.

At the same time, digital libraries in the 1990s seemed to promise more than improved access alone, and in these other areas they have proven to be somewhat of a disappointment. See the recent analysis of the impact of digital libraries on humanities research from the perspective of several librarians who did some research on their impact:

While digital resources are becoming more visible in the humanities, use of these resources by scholars remains limited. Humanists have come to rely on computers and electronic communication for some of their daily work, but the use of digital information resources has yet to become routine. Digitization projects are bringing texts, data sources, sound, and images to the scholar's desktop; however, the functions on which research in the humanities depend are neither well understood nor well supported by librarians.

(Brockman et al.)

The Brockmann et al. report goes on to examine some aspects of humanities research in general, and proposes some common elements that appear in the work of many scholars, and that perhaps should influence future technical developments in digital libraries. One of the common elements is 'annotation'. Humanists write in their books, they scribble notes on photocopies; they print out material from online sources and write on it as well. The report claims that the process of writing these annotations, and the recording and organising of notes that arise from this work supports the scholarly research process for many researchers. It is interesting to note that the EEBO user mentions this ("scribble[ing] on my printouts") explicitly as one of the benefits of using EEBO as well.

There has been work in computing science developing models of the role of annotation to support reading and research. Much of this has evidently been linked to the development of tablet computers where it is possible for the user to write on a digital copy in the same way that they might want to write on a printed one. Catherine Marshall's article "Towards an Ecology of Hypertext Annotation" reports on a study of annotations in textbooks, and begins the process of developing some models of annotations, based upon the how the annotator intends to use them in the future. In Marshall et al. there is a report on how annotations on tablet computers supported the research aims of a reading group. Bradley, in Bradley 2004, applied and extended some of the models outlined by Marshall in her papers to the task of humanities textual scholarship. This paper will extend some of the issues presented in part there.

There is also some critical literature on the role of annotation-like note taking and organisation, and an analysis of the impact of computing support for these tasks in the social sciences. There is some discussion of this literature, and its possible relevance to humanities text-based scholarship in

Bradley 2003, and this article draws attention to the potential of providing computing support for such materials that goes beyond merely supporting the *representation* (in something like TEI) of notes and annotations to also developing tools to support the *process* of building and organising the annotations, and then using them to support study of the materials they are linked to.

As a response to this work, we created a formal model for annotations, a prototype annotation tool and an annotation presentation environment in *OCVE*. In this paper we shall discuss how we arrived at the model we had for *OCVE*, where it supported (and where it failed to support) the act of analysis, and we will provide some thoughts on how these tools might be improved to better support creation and analysis. In addition, *OCVE* annotations contained in their model some sense of annotations for public and for private use. The paper will also discuss some of the implications we noted in using annotations in a public vs. a private manner, and how public annotations relate to recent developments in public collaboration software such as wikis.

Marshall, Catherine C., G. Golovchinsky, M. Price, and Bill Schilit. "Introducing a digital library reading appliance into a reading group." *Proceedings DL* (1999): 77-84.

*Online Chopin Variorum Edition Pilot Project*. Accessed 2005-03-03. <<http://www.kcl.ac.uk/humanities/cch/ocve/final/content/index.html>>

## Bibliography

Bradley, John. "Finding a Middle Ground between 'Determinism' and 'Aesthetic Indeterminacy': a Model for Text Analysis Tools." *Literary and Linguistic Computing* 18.2 (2003): 185-207.

Bradley, John. "Highlighting the Past: Annotation of historical texts to support Humanities Scholarship." Presentation given at the University of Kentucky and MITH, University of Maryland. April 2004. Accessed 2005-04-06. <<http://www.kcl.ac.uk/humanities/cch/jdb/papers/ken-tucky.pdf>>

Brockman, William S., Laura Neumann, Carole L. Palmer, and Tonyia J. Tidline. *Scholarly Work in the Humanities and the Evolving Information Environment, a report from the Council on Library and Information Resources*. Washington: Council on Library and Information Resources, December 2001. Accessed 2005-03-21. <<http://www.clir.org/pubs/reports/pub104/pub104.pdf>>

Leon, Pat. "Early English Books Online: The Holy Grail of online resources?" *Arts and Humanities Data Service Newsletter* (Autumn/Winter 2004). Accessed 2005-03-21. <[http://www.jisc.ac.uk/index.cfm?name=news\\_eebo](http://www.jisc.ac.uk/index.cfm?name=news_eebo)>

Marshall, Catherine C. "Towards an Ecology of Hypertext Annotation." *Hypertext* 98 (1998): 40-49.

---

## Improving Access to Encoded Primary Texts

---

*Terry Butler (Terry.Butler@UAlberta.ca)*  
*University of Alberta*

---

### The Access Problem

An impressive amount of our literary heritage has now been put into digital editions. Much of it is encoded in XML, often using recognized standards for encoding such as the TEI. One of the primary scholarly goals behind this activity has been to increase access to the texts - by publishing them on-line, and by making the text amenable to searching. The XML tagging provides further added value for searching and display. Metadata, where it exists at all, is mostly at the collection level, or provides only a broad guide to the contents of a specific work.

Between high-level metadata access, and a direct search on the word forms of the text, there is little help for the reader. Due to the immense labour involved in creating detailed subject indexing, very few scholarly electronic texts have indexes or finding aids which would draw the reader to specific sections of the work.

To address this deficiency, a first trial has been made at automatic indexing of a substantial non-fiction work. The notebooks of Samuel Taylor Coleridge are a rich treasury of the thought of one of the 19th century's most important intellectuals. Comprised of over 6,500 individual entries (in scope ranging from a single phrase to complete essays), they are a valuable record of his thought and the active intellectual currents of the time. We have captured the text of the notebooks in electronic form, encoded with TEI. As a first step to building a coherent subject index to this material, we have generated a mapping between this material and a contemporary subject index (Roget's first edition of his celebrated *Thesaurus*).

Our strategy has been to construct connections between the conceptual categories in the *Thesaurus* and Coleridge's individual notes, based upon a weighted measure of similarity between the words of the note and the terms and sub-terms in the *Thesaurus*. Common words are weighted lightly; rarely used words heavily. Using this measure, we can connect each note to one or more thesaurus entries, which then makes the note accessible to searching through the thesaural categories. Implementing these connections through topic map technology,

we have a stand-off tagging structure that relates these two encoded works but still leaves both of them unchanged, available to be delivered and shared with colleagues.

This presentation will describe the process by which we create an appropriate mapping between Coleridge's text and Roget's hierarchy, demonstrate the environment for creating and managing the stand-off tagging, and describe the utility of the resulting product.

The resulting edifice illustrates three important advantages for access to scholarly text: the index connects and relates sections of the text to larger, consistent conceptual categories; it provides access for searching that is complementary to the texts' own idiosyncratic terminology; it uses stand-off tagging to provide access without direct intervention in the electronic source text. This indexing structure, of value to researchers in its own right, is also the scaffolding upon which we will construct our subject index of the *Notebooks*, using modern terminology and accessible conceptual categories.

### Background to the Project

The notebooks of Samuel Taylor Coleridge are a valuable and almost unknown resource. Much of Coleridge's work as poet, philosopher, scientist, linguist, and theologian was published only partially and fitfully in his time; the notebooks contain some of his most innovative and interesting work. They have been published in print in five large double volumes (text and notes) by Princeton University Press, with indexes to selected titles, names, and places; but there is no subject index. The intention for the series was to publish a thematic index to the whole, as volume 6. However, we argued (to the Canadian Social Sciences and Humanities Research Council, who are funding this work) that at the present time an electronic index to the work would be of much greater utility to scholars who wish to know how Coleridge's thought emerged and developed over the 40 years which these notebooks cover.

The overall goals for the project include:

- creating an accurate electronic text of the entire notebook corpus;
- creating an index and thesaurus for the notebooks which will be a start to a synthetic index to Coleridge's thought;
- providing a web-based search and discovery system which will meet the needs of scholars, making his thought on a vast variety of topics more accessible.

## Bibliography

Coleridge, S.T. Ed. Kathryn Coburn. *The Collected Works of Samuel Taylor Coleridge*. Bollingen Series 75. Princeton: Princeton University Press, 1969.

Hüllen, W. *A history of Roget's thesaurus: origins, development, and design*. Oxford: Oxford University Press, 2004.

Pepper, S. *The TAO of Topic Maps*. 2001. Accessed 2005-03-15. <<http://www.ontopia.net/topicmaps/materials/tao.html>>

Sebastiani, F. "Machine learning in automated text categorization." *ACM Computing Surveys* 34.1 (2002): 1-47.

Thompson, H.S., and D. McKelvie. *Hyperlink semantics for standoff markup of read-only documents*. Language Technology Group, HCRC, University of Edinburgh, 1997. Accessed 2005-03-15. <<http://www.ltg.ed.ac.uk/~ht/sgm1eu97.html>>

---

## Learning Objects in Humanities Education

---

**Terry Butler** ([Terry.Butler@UAlberta.ca](mailto:Terry.Butler@UAlberta.ca))  
University of Alberta

**Catherine Caws** ([ccaws@uvic.ca](mailto:ccaws@uvic.ca))  
University of Victoria

**Norm Friesen** ([cmns007@sfu.ca](mailto:cmns007@sfu.ca))  
Simon Fraser University

**Scott Leslie** ([leslies@island.net](mailto:leslies@island.net))  
BCcampus

**Griff Richards** ([griff@sfu.ca](mailto:griff@sfu.ca))  
Simon Fraser University

**Ray Siemens** ([siemens@uvic.ca](mailto:siemens@uvic.ca))  
University of Victoria

---

**I**n *Understanding Media*, Marshal McLuhan suggests that the content of any new medium, at least initially, is provided by the medium that it is in the process of supplanting. For example, the content of early writing, as in Homer's *Odyssey*, is the spoken word, and the content of early cinema was theatre or vaudeville. Developments in Web technology and the use of this technology in education also seem to follow this pattern. Exclusive concern with document appearance and presentation — characteristics inherited from the print world — have gradually given way on the Web to dynamic and multimedia formats, and to distributed organizational mechanisms. Similarly, in distance education and educational technology, the Web initially took as its content the lectures, overheads, discussions and other aspects of the traditional classroom. Many of these aspects — down to the closed classroom door, the grade book and the classroom whiteboard — have been faithfully transferred onto the Web via password-protected course management systems like *WebCT* and *Blackboard*. However, attempts to replicate the face-to-face classroom seem to be giving way to distributed systems of 'Learning Objects' that exploit the intrinsically decentralized and decomposable nature of Web-based content, and that lend themselves to both 'blended' and 'distance' learning approaches.

'Learning Objects' is a term used to describe resources that can be used, shared and reused across a wide variety of educational contexts. Such resources can include images, video, Flash animations, text and HTML documents, as well as more

complex aggregations of this content. These resources can take the form of cultural content (e.g. articles, broadcast clips, or Websites) that has been adapted for use in educational contexts. The use of the term 'object' is an intentional reference to object-oriented programming and design, which has made use of modularity, hierarchical content structures and standardized interfaces to promote the use and reuse of programming resources in software development. With Learning Objects, it is hoped that some of these advantages can accrue also to resources used in education. Like content developed through object-oriented design, these Learning Objects will hopefully benefit from the congruence of their nature with the fundamental characteristics of the Web: its distributed nature, the modular, or decomposable nature of its content, and its use of agreed upon or de facto technical standards for file formats, descriptive information, and connectivity protocols (e.g. XML, Dublin Core, http).

A wide variety of projects in which Learning Objects play a central role have been underway both in Canada and abroad. These include the recently completed, pan-Canadian *eduSource* project, which has produced infrastructure and support mechanisms that are being further utilized in the *Lionshare*, *Apollo* and the *Eisenhower National Clearinghouse* projects. Many of these large-scale projects collect resources across the humanities, social and natural sciences. Other 'repository' or Learning Object collection projects, such as *FLORE*, *MusicGrid* and *Internet Shakespeare Editions* focus specifically on language learning, and humanities subjects.

The proposed panel will critically assess the opportunities and challenges presented by a variety of LO initiatives that are provincial, national and international in scope, and that reflect the trends described above. Each panel member and project represented brings a different emphasis, representative of a different cross-section of users, relating and contributing in different ways to humanities education. Each member also brings a different background from humanities education, and the extensive experience in the use of digital content in this area. Discussion among the panelists and with the audience will focus on the issues, advantages and challenges of this approach in humanities education:

**Terry Butler** is the Director of Research Computing in the Faculty of Arts at the University of Alberta, and recently completed a stint as Interim Director of Academic Technologies for Learning. Terry works closely with faculty both in the humanities and other discipline areas in the integration of computer technology in research and teaching. As lead on the *Technology Edge* project, he has researched and developed multimedia content to improve the information technology skills of liberal arts students.

**Catherine Caws** is an assistant professor in both the Department of French and the Department of Curriculum and

Instruction at the University of Victoria. Dr. Caws conducts research in collaborative learning in higher education, computer-assisted language learning, and computer networking, and she plays a leadership role in the *FLORE* repository of French Language Learning Objects.

**Norm Friesen** is currently Director of the *CanCore Metadata Initiative*, and principal investigator in the SSHRC-sponsored *learningspaces.org* project. He is also a visiting Scholar at the School of Communications at Simon Fraser University, and a member of the Canadian delegation for the ISO sub-committee on "Information Technology for Learning, Education and Training."

**Scott Leslie** is the Manager of the BCcampus Learning Resources Centre, a multi-disciplinary 'open content' repository. In addition, he researches course management systems, repository and eportfolio software as part of the Western Cooperative on Educational Telecommunications' *EduTools.info* team.

**Griff Richards** concerns himself with the convivial use of technology to promote the creation, management and transfer of human knowledge. Griff has a Ph.D. in Educational Technology from Concordia, and has been active in the research, development and implementation of computers in education and training for 25 years. His most recent work has been in the context of the Mellon Foundation *Lionshare Peer to Peer Learning Object Repository* project ( <http://lionshare.its.psu.edu/main/> ).

**Ray Siemens** is Canada Research Chair in Humanities Computing and Associate Professor of English at the University of Victoria. Director of the Digital Humanities / Humanities Computing Summer Institute, founder of Malaspina U-C's Centre for Digital Humanities Innovation, and founding editor of the electronic scholarly journal *Early Modern Literary Studies*, he is also author of many articles focusing on areas where literary studies and computational methods intersect.

## "Temas de Literatura Universal": Usos y Aplicaciones del Hipertexto Pedagógico

---

**Roger Canadell** (*rcanadell@uoc.edu*)

*Universitat Oberta de Catalunya*

**Laura Borràs** (*lborras@uoc.edu*)

*UOC/Hermeneia*

**Isabel Clara Moll Soldevila** (*imoll@uoc.edu*)

*Universitat Oberta de Catalunya*

---

La introducción de las tecnologías digitales en los procesos de aprendizaje ha significado la creación de nuevos espacios educativos basados en la no-presencialidad y en la asincronía. Pensamos que la tecnología ofrece mecanismos que permiten crear marcos de aprendizaje nuevos que deben llenarse de contenido y de humanidad. En este sentido, en los estudios de Filología de la Universitat Oberta de Catalunya trabajamos para que el e-learning sea, también, una manifestación del e-living. Ante las voces que defienden que el entorno digital no tiene ninguna especificidad didáctica o pedagógica queremos mostrar como aquello que es realmente revolucionario -más allá de si añadimos, o no, una e al concepto aprendizaje- son los nuevos mecanismos docentes.

El European Plan of Action 'eLearning 2001' define el e-learning como el uso de tecnologías y de Internet para mejorar la calidad docente, y en nuestro caso creemos que esta mejora viene condicionada por la humanización del acompañamiento docente. El uso de sistemas complejos en un campus virtual permite y estimula el intercambio y la colaboración, los cuales conllevan unos resultados más que satisfactorios en las asignaturas que ofrecemos en la UOC.

La Universitat Oberta de Catalunya nació en el 1995 como la primera universidad absolutamente virtual del mundo, y por ello era necesario redefinir conceptos como docencia, aprendizaje o estudio, ya que al crear un entorno en el cual las aulas son sustituidas por la virtualidad hay que tener en cuenta los cimientos en que se ha sustentado siempre el aprendizaje y la docencia, y sumarle las distintas posibilidades que la nueva era nos ofrece.

Si tomamos como ejemplo la asignatura 'Temas de Literatura Universal' se puede ver perfectamente de qué manera se ha organizado la materia. Partiendo de una serie de textos

relacionados entre sí por el hecho de compartir algunos de los tópicos de la literatura universal, se ha elaborado un material hipertextual que combina la lectura lineal con la lectura secuencial o fragmentaria propia del hipertexto, así como el vídeo o los recursos sonoros. Lo que ofrecemos son lecturas de algunos textos literarios cruciales a través de una doble ruta: la estrictamente literaria, la pictórica o la cinematográfica, de tal forma que durante el proceso se analiza la obra teniendo en cuenta otros textos precedentes que la han influenciado. Así, los itinerarios de lectura atraviesan períodos y tradiciones culturales distantes en el tiempo y en el espacio mediante conexiones intertextuales que nos guían por el corpus hipertextual. Los estudiantes, después de familiarizarse con los textos del material docente, deben seleccionar un tema y construir su propio corpus hipertextual con la ayuda constante del profesor con el que están siempre en contacto.

La apuesta de nuestro trabajo diario como profesores virtuales se basa en la utilización de materiales didácticos electrónicos, recursos on-line, bibliotecas digitales, websites de referencia, exposiciones virtuales, etc. y un taller -muy bien valorado por los estudiantes- que les permite comparar sus ejercicios con los de otros compañeros y beneficiarse de sus correcciones. Además de utilizar estos materiales, resulta del todo necesario considerar que el acto de enseñanza on-line que utiliza estos recursos debe transformar su discurso y utilizar nuevas técnicas comunicativas. El hipertexto muestra una nueva forma de textualidad basada en la capacidad de penetración de un texto marcado con enlaces que le abren ventanas a nuevos horizontes interpretativos y a nuevos sentidos. Con el hipertexto se desvanece toda pretensión de control y la seducción es la única motivación que se busca durante la navegación, en la cual resulta imprescindible el acompañamiento a lo largo del proceso de aprendizaje mediante la pizarra virtual, la incitación al debate, el forum virtual, la corrección personalizada de ejercicios, la respuesta rápida a las dudas, etc. En definitiva, el proceso de enseñanza-aprendizaje concebido de esta manera acaba convirtiéndose una tarea holística que es beneficiosa para los estudiantes, ya que tienen que leer, comparar, escuchar a sus compañeros y al profesor, participar organizar sus ideas de forma lógica y coherente. Al fin y al cabo, el estudiante organiza y construye su propio proceso de aprendizaje de manera subjetiva y utilizando sus propias capacidades e iniciativas, de tal manera que la docencia se transforma completamente respecto a la universidad presencial y eje de todo el proceso docente se centra en el estudiante más que en el profesor.

La puesta en funcionamiento de la asignatura Temas de literatura universal se produjo el año 2001, justo en el momento en que intentamos ver si los estudiantes de la UOC eran capaces de enfrentarse de una manera radical al nuevo entorno desde su mismo lenguaje: el hipertexto. Éste era el reto, pero al mismo tiempo también había que ver si nosotros, como profesores, estábamos preparados para desarrollar nuestro rol a partir de



la especificidad de transmisión de contenidos que nos ofrecía el entorno digital. No se trataba sólo de tener webs u otros espacios virtuales que complementarían nuestra tarea docente dentro del aula, ni de elaborar materiales en papel, es decir, manuales sobre una determinada materia y preparar actividades para los estudiantes a distancia; hacía falta ver si podíamos sustituirnos o, cuando menos, trasladarnos a la deslocalización opaca de un hipertexto. Comprobar si sabíamos des-organizar nuestro discurso académico tan aparentemente organizado. Como es lógico, usamos esta peculiar forma textual que el hipertexto impone a las obras que leemos e interpretamos, una forma que puede juzgarse como fragmentaria. Fragmentaria porque practicar el arte del screen writing implica necesariamente una determinada disposición espacial de los contenidos que no es discursivamente lineal (en el sentido que el discurso se atomiza y se fragmenta en ítems: las lexias, que forman parte de bloques textuales enlazados pero contiguos). Y también fragmentaria desde el punto de vista de la accesibilidad a los textos, ya que sólo habían sido seleccionados fragmentos textuales y no obras enteras. Una fragmentariedad, pues, que es la consecuencia de una determinada estrategia didáctica, pero que también refleja cambios en la mentalidad y en la percepción de la realidad que afectan a la lectura y a la escritura, y en general a la producción cultural.

Finalmente surgieron tres propuestas diferentes de uso hipertextual que fueron desarrolladas con el objetivo de fomentar las perspectivas críticas que reconstruyen la historia de la literatura según unos itinerarios y cánones densamente personalizados. Y para hacerlo había que establecer una fuerte alianza entre las dos metodologías usadas para tal finalidad: el principio de intertextualidad (en la elaboración de los materiales) y la interacción dialógica (en los debates del foro).

Esta asignatura se propone invertir la relación convencional entre el lector/crítico y el texto literario, transformando al primero en sujeto de una exploración que libremente transcurre de una obra hacia la otra, deshilvanando un hilo temático que él mismo establece. El formato hipertextual de los materiales responde, pues, en este tipo de lectura discontinúa que pasa de una obra a la otra siguiendo una temática preestablecida, en lugar de agotar cada texto en su integridad. De este modo, los tres bloques en los que se dividen los materiales representan tres antologías personales de fragmentos de obras (literarias o no) seleccionados por su conexión, especialmente representativo del tema que cada autor ha escogido, y comentados a partir de esta conexión. La característica esencial de estos materiales es la de presentar textos y comentarios (escogidos a partir de temas muy generales en uno formato hipertextual, o sea según recorridos de lecturas que no son secuenciales (un texto tras otro), sino intertextuales (un texto dentro de otro). Los que hemos trabajado en su diseño hemos pensado en las nuevas formas de lectura que Internet y la informática han creado, por las cuales los textos ya no están encerrados en sí mismos, y se

abren a un infinito juego de reescrituras e interpretaciones. Al mismo tiempo el sujeto que lee (o que estudia) es mucho más libre de seguir su propio itinerario, pasando de un texto a otro según las asociaciones intertextuales que su experiencia y su intuición le sugieren. En estos coexisten tres ejemplos de itinerario intertextual (relativamente) arbitrario y personal en la medida que, como ya hemos avanzado, el objetivo de la asignatura es que cada uno de los estudiantes haga lo mismo al final del semestre (construyendo un corpus textual que constituirá el trabajo final del curso).

Cabe destacar que en los tres cursos que se ha impartido esta asignatura, el aumento de estudiantes ha sido exponencial, y según el análisis de las estadísticas se puede decir que ninguna otra asignatura de literatura de la UOC ha tenido un éxito parecido. Los estudiantes manifiestan leer, estudiar y aprender mucho, a pesar de su sorpresa y desorientación inicial (cuando ven que el material de la asignatura tiene un formato hipertextual). Esta confusión inicial siempre acaba transformándose en un reconocimiento de las especificidades positivas del medio y la constatación de los beneficios del abordaje hipertextual llega cuando muchos de los estudiantes presentan su trabajo final utilizando los mismos recursos multimedia que se utilizaron en la creación de los materiales.

# A Pilot Study for a Navajo Textbase

*Kip Canfield* ([canfield@umbc.edu](mailto:canfield@umbc.edu))  
 University of Maryland

## Introduction

There are a large number of collected and written texts in various Athabascan languages that form a substantial literature that could be used for both scholarship and education. This is especially true of the Navajo language for which there are a large number of written texts, many that are public domain or out of copyright protection. This paper describes and evaluates a project to acquire these texts in electronic format, in the standard orthography, and develop a dictionary lookup tool for use with these texts. Collected texts can take many forms and use many different orthographies. For this pilot study, the Navajo texts are typewriter written with a non-standard orthography. The Navajo language has a polysynthetic structure that poses special problems for dictionary lookup.

## Methods

The following steps were used for this project and are detailed below:

1. Scanner acquisition of images of the original texts.
2. Optical character recognition of the text images and post-edit.
3. XML encoding of the texts using the Text Encoding Initiative.
4. Use of an XSLT stylesheet for web display of the texts.
5. Development of an automated look-up tool for the lexicon.

Texts collected in 1929 from the book *Navaho Texts* by Sapir and Hoijer (1942) are used for this pilot project. Figure 1 shows a page fragment from this work that uses a non-standard orthography. Figure 3 shows that same fragment with the standard orthography after acquisition.

### III. PERSONAL NARRATIVES

#### 29. The Story of a Navaho Woman Captured by the Utes

ʋaʎkʎidq̄ʋ, ʎdaʎoʎʋo'dq̄ʋ, naʋniʎka'dgo, nɔ'daʋe ʎi-  
 kiʎiʋ iʎʋ biʎha'ʎq̄e'ʋ. ʎe'ʋʎko ʎaʋ siʎiʎ. ʎiʎiʎʋq̄-  
 yidaʋni'ʎce'dgo ya'ndi'kaʎ. ʎi ʎeʎiʎe' ʎiʎd'ʎ. ʎaʋ ʎda-  
 ʋaʎʋabgo ʎaʋʎe'ya ʎiʎʋ yʎiʎiʎdaʋe'ni'ʎ. ʋe'do' ʋaʎiʋ ʎda-  
 yi'eʎe dɔ'ba'hda ʋadayi'ʎa'. ʋe'do' daʋi'dq̄ʋ. ʋaʎiʎq̄-  
 daʋi'dq̄ʋgo, ʎiʎʋ yʎiʎiʎdaʎdaʋaʎiʎ. ʎiʎiʎʋ yʎiʎi'hq̄ ʋatah  
 nikiʋʎiʎka'd. ʎiʎiʎkiʎda'.

Figure 1. A page image fragment from Navaho Texts

After the image of a page has been scanned, it must be recognized using optical character recognition (OCR). For this project, the open-source *Gamera* system was used which is written in the Python language. *Gamera* allows arbitrary characters to be trained using an implementation of the k-nearest neighbor algorithm whose weights are optimized using a genetic algorithm. Figure 2 shows the Gamera interface that allows iterative classification of characters from actual text images and supports training until the error rate is acceptable.

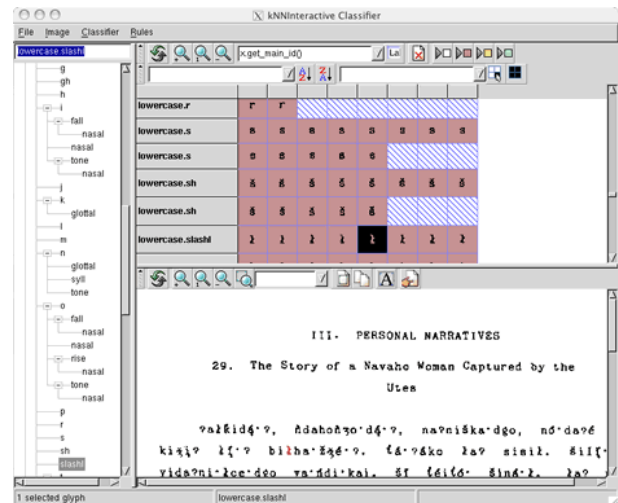


Figure 2. A Gamera Screen shot

The output configuration for a particular project requires Python programming to define the character mappings. The output is a mapping from the recognized characters of the text image to text in the Times New Roman Navajo font. The Text Encoding Initiative (*TEI*) is used for encoding that output. A sample of *Navaho Texts* that has been encoded using *TEI* is shown in Figure 3. It is transformed to HTML using the XSLT stylesheet that is available at the *TEI* website, augmented with a CSS file that includes the Navajo font.

The final step in the workflow is to develop and use a lookup tool for the lexicon that allows a user to click on a word and see the correct dictionary page or easily navigate to it. A major work for the Navajo lexicon is the *Analytical Lexicon* by Young

and Morgan (1992). There is also a project to put the *Analytical Lexicon (AL)* on-line that is partially completed and available at <http://www.speech.cs.cmu.edu/egads/navajo/>. The dictionary lookup tool developed here tries to map a verb stem parsed from a word to a page (URL) in this on-line AL. The problem is that a morpheme (the stem) must be extracted from the complex morphology of the verb for lookup which is typically a difficult task for users.

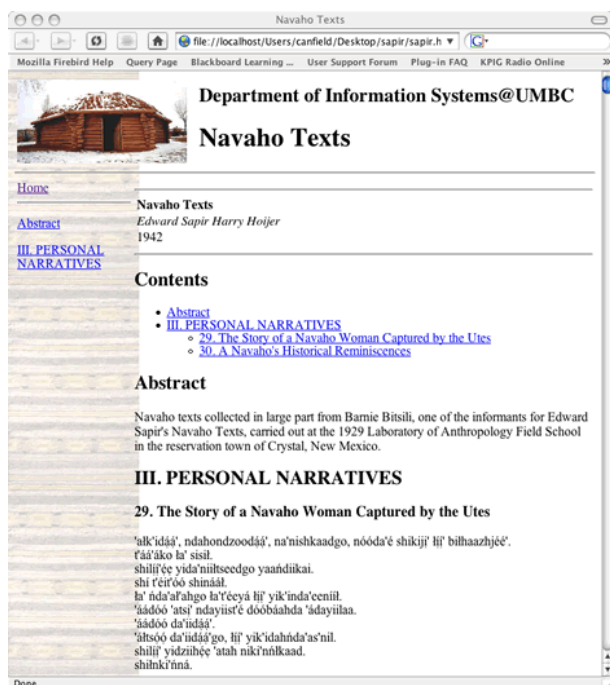


Figure 3. The TEI Encoding

A very simple example of a morphological parser for dictionary lookup is presented here. Pseudo-code for the algorithm is shown below. The algorithm is implemented in the Perl programming language.

1. Get the word
2. Look in a list for direct word lookup (no parsing)
3. If found, display the lexical entry
4. Otherwise:
  - (a) Parse the word (assume it is a verb)
  - (b) Match the longest common substring to a list of all stem shapes
  - (c) Score each match
  - (d) Rank the matches by score
  - (e) Link each stem match to the URL for the corresponding root in the AL

For step 4a, each substring of the verb is compared to a list of all stem shapes. A simple score is attached to each match in step 4c where:

$$\text{score} = (\text{index position of the substring}) * (\text{length of the substring})$$

This privileges matches that are towards the end of the word and longer substring matches. The ranked matches are displayed with the recommended one being the one with the highest score. Example output from the batch version of the Perl program that shows a correct parse is shown below.

- Word=na'nishkaadgo :
- nish - (12) - <http://www.speech.cs.cmu.edu/egads2/navajo/entry?nish>
- kaad - (28) - <http://www.speech.cs.cmu.edu/egads2/navajo/entry?kaad>
- na' - (0) - <http://www.speech.cs.cmu.edu/egads2/navajo/entry?na%27>
- ni - (6) - <http://www.speech.cs.cmu.edu/egads2/navajo/entry?ni>
- The recommended stem is kaad

The highest scored match (28) is the correctly recommended stem. Note that the URL to the on-line AL contains still another encoding for Navajo characters (a custom Latin1 mapping that is also URL encoded) and the Perl program must also translate between the standard orthography and this custom mapping.

## Results

The scanning procedure is very simple and does not require specialized equipment. The process of scanning does not require any special linguistic expertise and can be carried out as a batch job that produces the image files. The OCR training and classification process using the *Gamera* system is fairly straightforward and with the output programming pieces pre-done for a project, it can be performed by domain experts. The author found that using a training level with an approximately 15% error rate, he could do all acquisition steps of the workflow in under 20 minutes for a physical page and batch pre-scanning would have significantly reduced this time. The TEI encoding is simple and ensures that the textbase will be archival.

A formal evaluation of the dictionary lookup tool was performed. A sample of the first 300 words of the text shown in Figure 1 was selected and the Perl program parser was run against this sample in batch mode. This resulted in a list of 300 outputs such as that above. The author then checked each of the 300 parses for accuracy. Three categories were used to evaluate this output: correct, incorrect, and non-verb. The

non-verb category was used for adverbials, nouns, etc. that do not transparently map to verb stems. The result of this evaluation was that the parser was 92% accurate for Navajo verbs with a breakdown of: correct=124, incorrect=10, and non-verb=166. 45% of the sample is non-verb. The remaining problem is how to generally discriminate between verbs and non-verbs for lookup.

## Conclusions

**T**he workflow introduced here appears to be useful for domain experts that are trying to create on-line textbases for Athabascan literature. Ultimately these textbases could be used in the schools for language and cultural studies since they are easily implemented as web pages. The dictionary lookup tool goes at least some distance in solving the long-standing problem of helping users to navigate the complex Navajo lexicon. The link to the on-line *AL* is a simple example of cross-project interaction in the computational humanities. An updated programmatic interface to the *AL* would be a significant one.

## Bibliography

Sapir, Edward, and Harry Hoijer, eds. *Navaho Texts*. Iowa City: Linguistic Society of America, 1942.

Young, Robert, William Morgan Sr., and Sally Midgett. *Analytical Lexicon of Navajo*. Albuquerque: University of New Mexico Press, 1992.

---

## The Tibet Oral History Archive Project and Digital Preservation

---

*Linda Cantara* ([linda.cantara@case.edu](mailto:linda.cantara@case.edu))  
Case Western Reserve University

---

**T**he *Tibet Oral History Archive Project*<sup>1</sup> (*TOHAP*) is part of the research and education program of the Center for Research on Tibet in the Department of Anthropology at Case Western Reserve University.<sup>2</sup> The Center was created in 1987 by Melvyn Goldstein, John Reynold Harkness Professor of Anthropology, and Cynthia Beall, Sarah Idell Pyle Professor of Anthropology, to generate and disseminate new knowledge about Tibetan culture, society, and history, and was the academic pioneer in opening Tibet to in-depth anthropological and historical research. The *TOHAP* builds on a series of fieldwork-based studies that have examined the adaptation of Tibetans to high altitude, and the changes that have occurred since Tibet's incorporation into the People's Republic of China in 1951.

The *Tibet Oral History Archive* includes three primary collections:

- *The Common Folk Oral History Collection*: nearly 2,000 hours of interviews with hundreds of ordinary rural and urban Tibetans about their life experiences. Since the number of individuals in Tibet who were adults in 1959 -- the end of the traditional era -- is rapidly dwindling, there is particular urgency to document the voices of ordinary Tibetans in order understand the diversity of life as it was lived in Tibet as well as the way the salient historical events played out among the different strata of society.
- *The Political History Collection*: approximately 400 hours of historical interviews with former Tibetan government officials who played important roles in modern Tibetan history, including His Holiness the Dalai Lama. These interviews cover the traditional period before Tibet was incorporated into the People's Republic of China (1913-1951) and the subsequent period up to the end of the Cultural Revolution in 1976.
- *The Drepung Monastery Collection*: approximately 350 hours of interviews with about one hundred monks who were members of Drepung Monastery, Tibet's largest monastery, at the end of the traditional era. These interviews are unique in that they provide the only in-depth window into large-scale monasticism in traditional Tibetan society.

Conducted primarily in the Tibetan language, the interviews were taped on audio cassettes which have subsequently been digitized in three formats: archival WAVE files, medium format QuickTime files, and compressed delivery MP3 (MPEG) files. The interviews have been transcribed and translated into English and were initially saved as Microsoft Word documents. Professor Goldstein, Editor of the Archive, has partnered with Kelvin Smith Library to prepare the audio files and transcripts for online dissemination and long-term preservation. For online dissemination via the World Wide Web, we are converting the Word documents to plain text and encoding them in XML using the Text Encoding Initiative (TEI) Document Type Definition (DTD) for Transcriptions of Speech.<sup>3</sup> To facilitate understanding, the Archive will also include a glossary of terms, encoded in XML using the TEI-DTD for Printed Dictionaries.<sup>4</sup> A programmer has been hired to create a Web-based tool for creating the glossary and an application for automatically encoding extended pointer notation to link terms in the transcripts to their definitions in the glossary. Work is also underway to design an end user interface which will include browse and search functions. In the meantime, we are temporarily transforming the XML files to XHTML and using the *Greenstone Digital Library Software* to facilitate local access.<sup>5</sup>

A larger concern, however, is how to ensure long-term preservation of and access to the Archive. In 1996, the Commission on Preservation and Access (CPA) and Research Library Group (RLG) Task Force on Archiving of Digital Information published a seminal report on the long-term preservation of digital resources.<sup>6</sup> Since then, virtually every significant publication about digital preservation has indicated that primary responsibility for initiation and management of the metadata necessary to ensure long-term access to digital resources begins with the creator of the resource. Traditionally, it has been the role of librarians and archivists to ensure long-term viability of and access to cultural heritage materials, but this is not within the realm of expertise of the majority of scholars in the humanities and social sciences. Thus, if the creators of digital resources are responsible for initiating lifecycle documentation of the descriptive, administrative, and structural metadata necessary to migrate, emulate, or otherwise translate existing resources to future hardware and software configurations -- a task foreign to most discipline-based scholars -- close collaboration with information technology professionals early in a project is imperative.

Protocols and standards for digital preservation are now under vigorous development, yet there are still many unknowns. For the short-term, multiple copies of the audio and XML files will be maintained in multiple locations at Case Western Reserve University, both at the *Center for Research on Tibet* as well as in *Digital Case*, Kelvin Smith Library's *Fedora* repository.<sup>7</sup> For the long-term, the Asian Division of the Library of Congress

has expressed interest in hosting the completed Archive. To prepare the *Tibet Oral History Archive* for deposit with the Library of Congress, we are creating a Submission Information Package (SIP) in compliance with the *Reference Model for an Open Archival Information System (OAIS)*,<sup>8</sup> using the Metadata Encoding and Transmission Standard (*METS*), a metadata standard for encoding descriptive, administrative, and structural metadata regarding objects within a digital library.<sup>9</sup> This paper will present a prototype for scholar-librarian collaboration in the digital preservation of multimedia resources, including a discussion of the practical aspects of constructing a *METS* document for the *Tibet Oral History Archive*, with particular attention to the multiple metadata standards that must be bundled with the digital files to create a robust Submission Information Package.

1. This project is sponsored by the Henry Luce Foundation with additional support from the National Endowment for the Humanities (grant no. RZ-20585-00) and the National Geographic Society.
2. The Center for Research on Tibet Web Site is <http://www.case.edu/affil/tibet/index.htm> .
3. Chapter 11 of the *TEI Guidelines* (P4); see <http://www.tei-c.org/P4X/TS.html> .
4. Chapter 12 of the *TEI Guidelines* (P4); see <http://www.tei-c.org/P4X/DI.html> .
5. *Greenstone* is open source software for building and distributing digital library collections, produced by the *New Zealand Digital Library Project* at the University of Waikato, and developed and distributed in cooperation with UNESCO and the *Human Info NGO*. See <http://www.greenstone.org> .
6. Commission on Preservation and Access (CPA) and Research Library Group (RLG). *Preserving Digital Information: Report of the Task Force on Archiving of Digital Information*. May 1996. Online at <http://www.rlg.org/legacy/ftp/pub/archtf/final-report.pdf> .
7. *Fedora™ Flexible and Extensible Digital Object Repository Architecture* -- is an open source digital repository management system, developed by Cornell University and the University of Virginia, available at <http://www.fedora.info> .
8. A SIP is "an information package that is delivered by the producer [of a digital object] to the OAIS for use in the construction of one or more AIPs [Archival Information Packages]." See "OAIS Terms". *Digital Preservation Management: Implementing Short-term Strategies for Long-term Problems*. Cornell University Library, 2003. Online at <http://www.library.cornell.edu/iris/dpworkshop/working/terminology/oais.html> . See also, *Consultative Committee for Space Data Systems (CCSDS). Reference Model for an Open Archival Information System (OAIS)*. CCSDS 650.0-B-1. ISO 14721:2003. January 2002. Online at <http://ssdoo.gsfc.nasa.gov/nost/wwwclassic/documents/pdf/CCSDS-650.0-B-1.pdf> .

9. *METS* is maintained in the Network Development and MARC Standards Office of the Library of Congress, and is being developed as an initiative of the *Digital Library Federation*. See <http://www.loc.gov/standards/mets> .

## Bibliography

- The Center for Research on Tibet's Web Site*. Accessed 2005-03-29. <http://www.case.edu/affil/tibet/index.htm>
- "Chapter 11: Transcriptions of Speech." *TEI Guidelines (P4)*. Text-Encoding Initiative. Accessed 2005-03-29. <http://www.tei-c.org/P4X/TS.html>
- "Chapter 12: Print Dictionaries." *TEI Guidelines (P4)*. Text-Encoding Initiative. Accessed 2005-03-29. <http://www.tei-c.org/P4X/DI.html>
- Fedora*. Cornell University and the University of Virginia. Accessed 2005-03-29. <http://www.fedora.info>
- Greenstone*. University of Waikato. Accessed 2004-07-16. <http://www.greenstone.org>
- METS*. Digital Library Federation. Accessed 2005-01-25. <http://www.loc.gov/standards/mets>
- "OAIS Terms." *Digital Preservation Management: Implementing Short-term Strategies for Long-term Problems*. Cornell University Library. Accessed 2005-03-29. <http://www.library.cornell.edu/iris/dpworkshop/working/terminology/oais.html>
- Reference Model for an Open Archival Information System (OAIS)*. CCSDS Secretariat. Accessed 2002-01. <http://ssdoo.gsfc.nasa.gov/nost/wwwclassic/documents/pdf/CCSDS-650.0-B-1.pdf>
- Waters, Donald, and John Garrett. *Preserving Digital Information: Report of the Task Force on Archiving of Digital Information*. Accessed 2005-03-29. <http://www.rlg.org/legacy/ftp/pub/archtf/final-report.pdf>

---

## *DocScapes: Visualizing Document Structures with SVG*

---

*Hugh Cayless* ([hcayless@lulu.com](mailto:hcayless@lulu.com))

*Lulu* ( <http://lulu.com> )

---

The task of searching for and browsing documents online can be a frustrating one. Documents in search results are typically treated as atomic units rather than structured collections of information. This paper proposes some ideas for enhancing search and browsing by producing graphical 'document-scapes' that summarize document characteristics and provide links into the content of documents. The advantage of this type of summary is that it can compensate for some of the visual cues (available when browsing bookshelves) that are lost in the digital environment. It is possible to visually summarize document size, structure, density, and the presence of metadata in such a way that users will be able to tell, at a glance, the difference between (for example) an interview and a monograph, or a play and a catalog. The work in this paper focuses on a particular vocabulary of document markup, TEI, and a particular collection, *Documenting the American South* at the University of North Carolina at Chapel Hill ( <http://docsouth.unc.edu> ).

A great deal of work has been done on the visualization of collections and search results (see <http://www.cs.umd.edu/hcil/research/visualization.shtml> for a summary of online material). There is, however, a remarkable paucity of scholarship focusing on the visualization of documents themselves. No doubt this has to do with the difficulties of dealing with heterogeneous collections. Comparing the varying structures of text, XML, and PDF documents, for example, might not be an especially useful exercise. The technique discussed in this paper can easily be applied to relatively homogeneous collections of XML documents, however, and could in theory be generalized to other document types.

The techniques used in this project are relatively simple. Essentially, what is involved is the transformation of XML from one vocabulary to another; in this case TEI to SVG. Scalable Vector Graphics is an XML application that allows for the representation of vector graphics in an XML format. This means that the structure of a document in, for example, TEI, can be turned into an image via the same processes used to display the document in HTML or to convert it to PDF for printing. Since other document formats can be parsed to generate SAX (Simple API for XML) events, they too could

be fed into an XML processing pipeline and turned into *DocScape* images.

There are a number of variables which may be used to distinguish documents marked up in TEI without recourse to semantic distinctions like subject vocabularies. Since TEI documents are subdivided by division (<div>, <divN>, <front>, <back>, etc.), each document has its own internal structure. Different types of document may have very different internal structures. For example, a dictionary will consist of a set of entries (<entry> tags) inside its divisions while a monograph will contain chapters, sections, and paragraphs (<p>). The relative size and structure of nested divisions can be represented graphically in a fairly compact space. Differing types of content, on the other hand, can be represented using color.

TEI documents also differ in size (obviously) and this can be an important metric. Size can be represented visually in a number of ways. *DocSouth's* collection varies widely in terms of absolute size, from short pamphlets to large books and government documents (up to 800 pages in length). The representation of relative size must therefore be considered quite carefully. The first iteration of *DocScapes* did this using border thickness. A pixel was added to the border width for each 100 pages. This sort of scale does not help in handling the important distinction between the moderately sized (10-50 page) document, and the very short (1-2 pages), a distinction which encompasses important differences of genre. The next generation of *DocScapes* will use more complex SVG capabilities, such as drop shadows to indicate relative size.

Another important metric is the relative size and complexity of the TEI Header metadata. *DocSouth*, whose documents are largely derived from catalogued library holdings, has very detailed and thorough header information. By contrast, a TEI document that was 'born digital' might have fairly minimal metadata. A visual distinction of different levels of metadata density will be useful for collection managers and searchers alike.

A *DocScape* image is composed of the elements outlined above: the document itself, any header metadata and structural container elements (e.g. <div>s in TEI, <section>s in *DocBook*, etc). The four TEI Header sections are represented by blocks of color at the top of the image. The nested divisions are visualized as nested blocks, moving first left-to-right then top-to-bottom, and so on. The nested blocks start from different ends of the light/dark scale, so top-level containers are light green, then their children are dark green, etc. In addition, the image attempts to quantify the number of paragraphs per page or section using color saturation. The relative size of the document is indicated by the border thickness of the entire image (see figures 1 and 2).

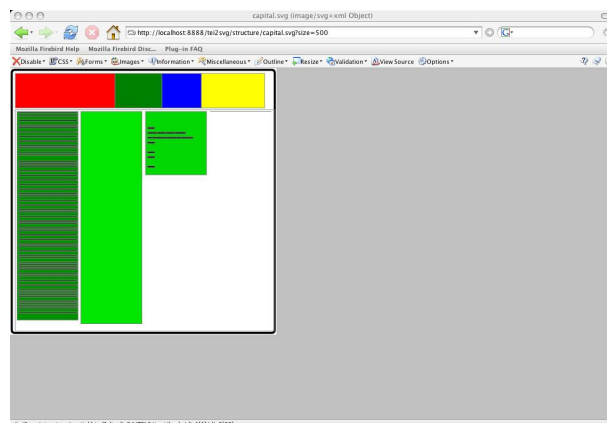


Figure 1

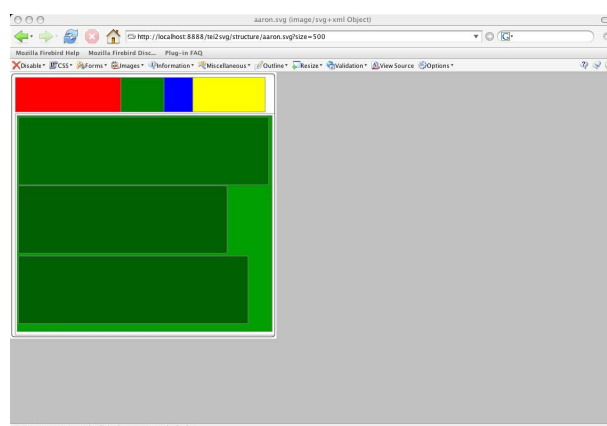


Figure 2

Figure 1 provides a nice example of a document with a very heterogeneous internal structure. The first section is a catalog, with many nested TEI <div>s, while the following divisions are more narrative in nature. Figure 2, on the other hand, represents an interview. The more densely packed paragraph structure in this document is represented by the lighter shade of green in the nested sections.

In addition to these basic elements, it is possible to use the capabilities of SVG to group many documents on a single page and dynamically zoom into the ones that are of interest. The document sections may also be linked to the documents themselves, so that it is possible to drill into the texts from their visual representations. Finally, it is possible to layer other information, such as the occurrence of search terms onto the documents. Figure 3 is an example of a *DocScape* with personal names, locations, and dates plotted on the image surface. My paper will outline the techniques and principles involved in developing *DocScape* visualizations and will discuss ways in which they may be used in digital libraries as a means to browse textual content.

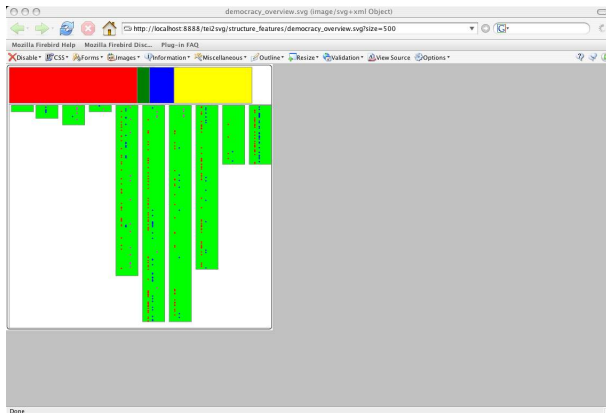


Figure 3

[106.ibm.com/developerworks/web/library/x-svgint/](http://106.ibm.com/developerworks/web/library/x-svgint/)

Visualization. Human-Computer Interaction Lab / University of Maryland. Accessed 2005-03-15. <<http://www.cs.umd.edu/hcil/research/visualization.shtml>>

## Bibliography

Börner, K. "Extracting and Visualizing Semantic Structures in Retrieval Results for Browsing." *Proceedings of the fifth ACM Conference on Digital Libraries*. 2000. 234-235.

Campeseto, O. *Fundamentals of SVG Programming: Concepts to Source Code*. Hingham, MA: Charles River Media, Inc., 2003.

Clark, James. *Transformations (XSLT), Version 1.0 (W3C Recommendation)*. W3C, 1999. Accessed 2005-03-15. <<http://www.w3.org/TR/1999/REC-xslt-19991116>>

Clark, James, and Steve DeRose. *XML Path Language (XPath), Version 1.0 (W3C Recommendation)*. W3C, 1999. Accessed 2005-03-15. <<http://www.w3.org/TR/1999/REC-xpath-19991116>>

Clark, James, Jun Fujisawa, and Dean Jackson. *Scalable Vector Graphics (SVG) 1.1 Specification (W3C Recommendation)*. W3C, 2003. Accessed 2005-03-15. <<http://www.w3.org/TR/2003/REC-SVG11-20030114>>

*Documenting the American South*. University of North Carolina at Chapel Hill. Accessed 2005-03-15. <<http://docsouth.unc.edu>>

Hornbæk, K., and Erik Frøkjær. "Reading Patterns and Usability in Visualizations of Electronic Documents." *ACM Transactions on Computer-Human Interaction (TOCHI)* 10.2 (2003): 119-149.

Sperberg-McQueen, C.M., and L. Burnard, eds. *TEI P4: Guidelines for Electronic Text Encoding and Interchange*. Text Encoding Initiative Consortium, 2002.

Venn, B. *Add Interactivity to Your SVG*. IBM developerWorks, 11 December 2003. Accessed 2005-03-15. <<http://www->



## DeMeCoT, The Delftse Methode Conversation Trainer

*Amal Chatterjee (a.k. chatterjee@tbm.tudelft.nl)*

*Delft University of Technology, The Netherlands*

*Piet Meijer (p.j.meijer@tbm.tudelft.nl)*

*Delft University of Technology, The Netherlands*

The two factors that played the most important role in the development of the DeMeCoT bot were feedback and the acquisition of speaking and conversation skills. Feedback, crucial in developing learners' hypotheses about correctness, is central in recent approaches to language teaching/learning like the natural (Krashen & Terrell; Yalden), which recommend 'natural' feedback (i.e. feedback in a communicative context). The theory is, the more frequent the feedback, the better. The introduction of computers in language learning has significantly extended the possibility of feedback for learners, and the possibility of providing it on an individual level (e.g., Van Der Linden).

The second element, the acquisition of speaking and conversational skills, is important because conversation training supports word retention and the development of grammatical correctness and, as a result, general language proficiency. It typically takes the form of classroom activities where students either interact with other learners while being occasionally observed and/or corrected by the instructor, or interact directly with the instructor playing an important role. However, these interactions can never fully meet the need for truly extensive practice due to time constraints. Again, computers, and bots in particular (e.g., the Dave ESL bot<sup>1</sup>), have shown promise as a potential solution.

Preliminary trials and assessments of a range of bots (Jabberwacky<sup>2</sup>, ALICE<sup>3</sup> and its various spin-offs, Pandorabots<sup>4</sup>) demonstrated significant limitations as (successful) language training bots, viz:

- i. interaction with them is prone to breakdown;
- ii. they lack the knowledge necessary for topic-based conversations;
- iii. they are insufficiently capable of eliciting specific language use;
- iv. grading is irregular (if present at all);
- v. changes in register are often non-standard or inconsistent.

Further consideration revealed, however, that several of the limitations, e.g., their defined (and definable) vocabularies and range of responses, were harnessable in the creation a conversation partner capable of encouraging extensive practice on selected subjects while providing feedback on the particular linguistic issues under consideration (ie being learned).

The resulting DeMeCoT Chatbot is a conversation trainer, a spin-off of ALICE (Artificial Linguistic Internet Computer Entity of the Artificial Intelligence Foundation) using AIML (Artificial Intelligence Markup Language) for use as a partner in an interactive, repeatable and graded conversation/dialogue that can be conducted at a speed appropriate to, and/or comfortable for, the individual learner. Complementary to the Delftse Methode Dutch course (co-authored by Piet Meijer), it provides students the opportunity to practice dialogue with continuous accurate responses and feedback and supplements the classroom speaking activities.

Being a teaching implementation, the DeMeCoT is deliberately and consciously limited in its function: its purpose is to enable students to practice linguistic structures and forms in a simulated, repeatable and limited (in terms of vocabulary, structure and content) conversation. Not a competitor for the Loebner Prize<sup>5</sup> or a restricted Turing test<sup>6</sup>, its users are instructed to limit themselves to the conversation topic in order to improving their accuracy in using the defined range of vocabulary and structures. As a result of these built-in and, in its case, desirable limitations, the DeMeCoT is also unusual in the deliberate "shallowness" of the personalities being built, the aim is to enable students to encounter different personalities in different situations.

The paper will describe the reasons for choosing a bot (repetition and variation in conversation, restrainable interactivity, logging, retrainability, variety and variability in design, platform independence), and how the limitations of bot technology have proved to be the strengths of the DeMeCoT. It will also elaborate on other factors that have affected the design and presentation, including the intended limitations, the need to define and regulate error tolerance and the type and quantity of feedback. The bot will be demonstrated (in the original Dutch and in English) and its future described, including the integration of a larger corpus of knowledge (the whole Delftse Methode course), strategies for structuring and grading learning using menus, setting up a system of learner 'promotion', provision of more detailed linguistic and grammatical feedback, linking directly to the study materials, the development of a 'non-specialist-user-friendly' instructor interface and research into the effectiveness of the bot and the method using control groups.

1. Dave ESL Bot: <<http://www.alicebot.org/dave.html>> (November 2004)
2. Jabberwacky: <<http://www.jabberwacky.com>> (November 2004)
3. ALICE: <<http://www.alicebot.org>> (November 2004)
4. Pandorabots: <<http://www.pandorabots.com>> (November 2004)
5. The Loebner Prize: <<http://www.loebner.net>> (November 2004)
6. The Turing Testpage: <<http://cogsci.ucsd.edu/~asaygin/tt/ttest.html>> (November 2004)

## Bibliography

- Krashen, S.D., and Tracy D. Terrell. *The natural approach: Language acquisition in the classroom*. Hemel Hempstead: Prentice Hall, 1983.
- Van Der Linden, E. "Does feedback enhance computer-assisted language learning?" *Computers and Education* 21 (1993): 61-65.
- Yalden, Janice. *The communicative syllabus: Evolution, design and implementation*. Englewood Cliffs, N.J.: Prentice-Hall International, 1987.

---

## Reflexivity and Arts Informatics

---

*Chris Chesher* ([chris.chesher@arts.usyd.edu.au](mailto:chris.chesher@arts.usyd.edu.au))  
*University of Sydney*

---

**A**rts Informatics is a recently introduced cross-disciplinary program at the University of Sydney. Students in this undergraduate degree take a major in the Faculty of Arts, and a major in Information Systems in the Department of Information Science in the Faculty of Science. This paper reports on the pedagogical and theoretical questions we are facing in building this program. As an academic unit operating as the intermediary between two faculties, we are the articulation point between two different academic worlds. Our approach is to read the Humanities as (cultural) technologies, and to unpack the humanity (and social construction) of information systems.

The Humanities has always been technological, even if that hasn't always been acknowledged. Arts Informatics draws heavily from the parts of the humanities tradition that do address these questions, from Plato's famous critique of writing, to Derrida's deconstruction and Marshall McLuhan's Gutenberg galaxy. The recent literature in new media studies directly addresses the implications of computers for knowledge and cultural practices (Manovich; Barret and Redmond; Everett and Caldwell; Wardrupp-Fruin and Montford).

In the other direction, social studies of technology offer the Arts Informatics program resources to analyse the ways that information technologies are embedded in wider cultural and societal processes. Actor network theory's adaptation of post-structural semiotics to technological change is particularly useful in offering a symmetrical approach to the human and non-human components in sociotechnical assemblages.

One of our central themes is the question of reflexivity. Anywhere computer technologies participate in traditional humanities practices (research, interpretation, textual production, communication, teaching), necessarily means qualitative transformations in that practice. The day-to-day materiality of work itself changes. Knowledge is acquired, handled, produced and communicated in different ways. As Michael Heim's analysis of word processing (1987, 1993) shows, computers are an additional component in daily practices, not simply a potentially intelligent opponent. While some of the claims behind the avante-garde experiments of 1990s hypertext theorists (Landow) seem somewhat overblown, the growing role of the Internet in everyday teaching, publishing and research over the past decade is incontrovertible.

Of course this recent experience of technocultural change is not exclusive to Humanities scholars and students. Our students have recent experience with the web, electronic mail, multiplayer computer games, DVD, SMS, digital television, and new cinematic paradigms. Their familiarity with such developments equips them to begin to understand the interweaving of technological and cultural transformations.

Students are not as well equipped to deal with the cultural differences between computer science and the humanities. The Humanities critiques of science and technology (Heidegger; Virilio; Coyne) are difficult to reconcile with scientific conceptions of humanities practices (Holtzman). Each of these areas places quite different, and often directly conflicting discourses, techniques and systems of value. It is important to acknowledge and investigate these differences. Even within the Humanities, there are very contrasting models for integrating new media technologies into teaching, theory and research.

Even outside these conflicts, teaching in this area seems to demand constant revision and updating. The only thing that changes more quickly than new media technologies themselves are the concepts used to describe them. Terminology seems to go in and out of fashion more quickly than new standards for data storage. Terms such as virtual reality, multimedia, hypertext, telepresence and artificial intelligence have controversial histories. The conflicts surrounding these terms have served to establish a vocabulary for discussing some of the key cultural changes associated with technological change.

The challenge is to remain open to interdisciplinary and transdisciplinary paradigms, while offering students a strong enough grounding in traditional disciplines to have some historical and epistemological orientation.

## Bibliography

- Barrett, Edward, and Marie Redmond. *Contextual media*. Cambridge, Mass: MIT Press, 1997.
- Coyne, Richard. *Technoromanticism. Digital narrative, holism, and the romance of the real*. Cambridge, Mass: MIT Press, 1999.
- Everett, Anna, and John T.Caldwell. *New media: theories and practices of digitextuality*. New York and London: Routledge, 2003.
- Heidegger, Martin. *The question concerning technology, and other essays*. New York: Garland Pub, 1977.
- Heim, Michael. *Electric language: a philosophical study of word processing*. New Haven and London: Yale University Press, 1987.
- Heim, Michael. *The metaphysics of virtual reality*. New York and Oxford: Oxford University Press, 1999.
- Holtzman, Steven R. *Digital mantras. The languages of abstract and virtual worlds*. Cambridge, Mass. & London: MIT Press, 1994.
- Landow, George P. *Hypertext. the convergence of contemporary critical theory and technology*. Baltimore & London: The John Hopkins University Press, 1992.
- Lister, Martin, Jon Dovey, Seth Giddings, Iain Grant, and Kieran Kelly. *New media: a critical introduction*. London: Routledge, 2003.
- Manovich, Lev. *The language of New Media*. Cambridge, Mass.: MIT Press, 2001.
- McLuhan, Marshall. *The Gutenberg galaxy*. Toronto Buffalo London: University of Toronto Press, 1962.
- Poster, Mark. *The second media age*. Cambridge: Polity Press, 1995.
- Virilio, Paul. *Lost dimension*. New York: Semiotext(e) (Autonomea), 1991.
- Wardrip-Fruin, Noah, and Nick Montfort, eds. *The New Media Reader*. Boston: The MIT Press, 2003.

## Laying that Damned Book Aside? Evaluating the Digital *Doctor Faustus*

---

**Tanya Clement** ([tclement@umd.edu](mailto:tclement@umd.edu))

Maryland Institute for Technology in the  
Humanities

---

### Good Angel

O Faustus, lay that damned book aside,  
And gaze not on it, lest it tempt thy soul,  
And heap Gods heavy wrath upon thy head,  
Read, read the scriptures, that is blasphemy.

### Evil Angel

Go forward, Faustus, in that famous art,  
Wherein all nature's treasury is contained:  
Be thou on earth as Jove is in the sky,  
Lord and commander of these elements. Exeunt.

(Act 1, Scene 1, Lines 69-76. The Perseus Project, *Tragedie of  
Doctor Faustus (B text)* (ed. Hilary Binda))

In her introduction to *Electronic Text: Investigations in  
Method and Theory*, Kathryn Sutherland asks if there is

a real danger that the scholar-worker, toiling for years in the remote  
regions of the library stacks in the hope of becoming expert in one  
small field, will be transformed by the computer into the  
technician, the nerdy navigator able to locate, transfer, and  
appropriate at an ever faster rate expert entries from a larger set  
of information that he/she no longer needs or desires to understand.  
(Sutherland 10)

Her inquiry is based on an issue that still plagues many scholars:  
with quick access to so much digitized information, how do  
we evaluate what we still need and desire to understand? Of  
course, her question implies that evaluating printed information  
is an evaluation based on less access and therefore a smaller  
set of information, and evaluating printed information is not  
an uncomplicated issue; it is one which scholars reconsider  
constantly. One such group--literary scholar-workers--may  
spend years 'toiling' over similar versions of a printed text in  
order to produce a single representative edition. In the case of  
Christopher Marlowe's *The Tragedie of Doctor Faustus*, for  
example, there is no extant manuscript, nine versions were  
printed between 1604 and 1631, and the first appeared almost  
nine years after Marlowe's death. Those that appeared in 1604,  
1609, and 1611 are similar and are collectively known as the  
A-text. The 1616, 1619, 1620, 1624, 1628, and 1631 versions

are also similar and known as the B-text. Which one should a  
reader or scholar consult?

Remarkably different, the A- and B-texts have inspired an  
extensive amount of critical commentary and scholarly editors  
since W.W. Greg appear to agree on one thing: neither the A-  
nor the B-text is considered wholly representative of Marlowe's  
original work. Still, scholars have attempted to represent what  
one most needs and desires to understand in an edition of  
*Doctor Faustus*. Evaluating a digital edition of *Doctor Faustus*  
can not—and should not—be based on exactly the same process  
even though it may be based on the same set of problems  
inherent to the *Doctor Faustus* work. Standards by which one  
may evaluate the digital *Doctor Faustus* are present in three  
very different digital versions of the work: The Perseus Digital  
Library edition ( <http://www.perseus.tufts.edu/> ),  
Early English Books Online (EEBO-TCP) collection ( <http://eebo.chadwyck.com/home> ),  
and the *Versioning Machine* ( <http://mith2.umd.edu/products/ver-mach/index.html> )  
electronic publishing environment.

To date, the two most critically important print editions of  
*Doctor Faustus* are W.W. Greg's *Marlowe's 'Doctor Faustus'  
1604-1616: Parallel Texts* (1950) and David Bevington and  
Eric Rasmussen's *Doctor Faustus A- and B-texts (1604, 1616)*  
published in 1993. Editors of print editions have sought to  
ameliorate the *Faustus* copy-text problem by printing both the  
A and B texts together in one volume. Greg lays out parallel  
versions on facing pages while Bevington and Rasmussen  
print the texts sequentially, A before B. Certainly, these print  
editions still pose some editorial complexities. In order to  
construct his facing-page arrangement, Greg had to  
"compromise" Marlowe's representation of the original plays.<sup>1</sup>  
In the original printings, a scene that appears early in the first  
act of the A text may not appear until much later in the B text,  
but since Greg was attempting to show parallel versions of the  
play, he moved those scenes to the same location in each play  
to facilitate that comparison (Greg 151). Likewise, Rasmussen  
and Bevington admit that while they "try to give the A- and  
B-texts straight," they "do, to be sure, adopt a few B-text  
readings in [their] A-text and vice-versa when corruption seems  
unmistakable" (Bevington and Rasmussen n. pag.). Both Greg  
and Bevington claim they seek to represent each text as it was  
"originally" intended but the print medium requires that these  
editors "compromise" their intentions.

Electronic versions of *Doctor Faustus* in EEBO-TCP and The  
Perseus Project allow for types of research that may have been  
unthinkable with traditional print resources. EEBO-TCP  
provides greater access to the first printings of *Doctor Faustus*  
with facsimiles and searchable text. Facsimiles of the original  
printings allow users to see bibliographical codes that a  
modernized printing or a digital transcription might otherwise

fail to present. The Perseus edition edited by Hillary Binda provides access to electronic tools for textual analysis. In Binda's edition of Marlowe's *Doctor Faustus* the user may compare Binda's modernized spellings with Greg's A- and B-texts. The user also has access to one of Marlowe's primary sources, *The English Faust Book* of 1592 by P. F. Gent. By providing the user simultaneous access to both versions, Binda fulfills her main objective, which is not to favor either version, an advantage that is replicated in the EEBO-TCP experience where the reader can choose to read any available printing in any order.

While the EEBO-TCP and the Perseus editions facilitate a new perspective of *Doctor Faustus*, they lack a level of editorial annotation that print scholarly editions of *Faustus* have usually included. Without annotations, unmediated facsimiles of Renaissance texts like those offered by EEBO-TCP may be considered misleading. As John Lavagnino points out, "a facsimile of an early edition may have more 'errors' in it than a modern reprint," but without the Renaissance reader's "awareness of the degree of uncertainty in the text; our corrected modern editions make it look like we're quite certain about what the text is supposed to say" (Lavagnino 67). The Perseus edition is problematic for other reasons. Binda chooses Greg's parallel edition "as a model for linking lines, passages and scenes between the two texts," but Binda also claims she does not want to indicate a preference for the A- or B-texts (Binda par. 14). Yet, as previously argued, Greg's paralleled version (which provides the basis of her encoded text) evinces a clear bias for B, making her claim untenable. Further, both projects use the TEI XML encoding standard — a standard that requires a substantial level of editorial decision-making, yet neither EEBO-TCP or Perseus discuss these choices, thereby abstracting another level of editorial influence. That basic aspects of textuality engendered by text selection or metadata encoding might not be declared by the editor of an electronic *Faustus* appears to elide an adequate level of accountability.

*The Versioning Machine* is an environment that facilitates displaying and comparing multiple versions of texts. The *VM* environment could be particularly productive in examining the *Doctor Faustus* text, because, as Schreibman notes, the text may be "freed from the spatial limitations of the codex" and could provide readers with the "reconstruction of [the] text's instantiations over time" (Schreibman "Computer-mediated texts" 291). Instead of simply providing access to facsimiles, this environment could provide access to a new perspective on the textuality of *Doctor Faustus* by including introductory material and traditional annotations plus "manipulatable images of the witness to be viewed alongside the diplomatic edition" (Schreibman *The Versioning Machine* 101). The *VM* environment could also provide access to sequential or parallel readings plus relevant images from first printings and annotations that discuss pertinent editing issues.

Of course, the *Versioning Machine*—like other electronic tools—is not without its limitations. For example, subtle variations that appear on the printed page (such as font or case changes, line numbers, act and scene numbers, etc.), must be 'hardcoded' or made explicit to the structure in the XML version, an encoding practice that is not encouraged by the TEI standard. Indeed, in order to elucidate editorial practices, the XML must account for all variations—even seemingly of the most diminutive significance—a cumbersome encoding process that may or may not yield a critically productive result for the user. Further, the encoding encouraged by the *Versioning Machine* documentation ( "parallel-segmentation" ) yields the same problems with *Faustus* that Greg faced; if a large section of "parallel" content appears in different locations in the text, the *Versioning Machine* cannot currently facilitate the HTML representation of that comparison (although changes in the XSLT and CSS could possibly provide alternatives).

It is apparent that our duty as computing humanists is not to evaluate which digital representation we may access or 'appropriate' most easily or which one might answer all our questions about what we most need or desire. Certainly, different scholars use different print versions as means for different ends. Likewise, a scholar may use EEBO-TCP to compare the bibliographic details of *Doctor Faustus* to one of approximately 125,000 other contemporary artifacts or use the very act of encoding an electronic version of *Doctor Faustus* in the *Versioning Machine* to analyze limitations in editing such different versions. The versions of *Doctor Faustus* that appear in EEBO-TCP and the Perseus edition and the relationship a scholar has with these texts in the *Versioning Machine* are different—from print editions and from each other—and they facilitate a different presentation of and relationship to the work. These electronic versions of *Doctor Faustus* should be evaluated on the goals their editors seek to achieve, the particular audiences for which they are intended, and the traditional modes of editorial accountability exemplified by Bevington, Boas, Breymann, Greg, Hunter, and Rasmussen, but they should also be evaluated in terms of the electronic medium.

In conclusion, for digital versions, scholars who do more than "locate, transfer, and appropriate at an ever faster rate expert entries from larger set of information" (Sutherland 10) in the digital environment must rely on the same scholarly pursuits as always: the desire to create new critical knowledge in the field. This knowledge, for a Renaissance text with the particular complexities inherent to *Doctor Faustus*, for example, may be dependent on some traditional editorial practices. After all, a tool like the *Versioning Machine* may provide access to sequential or parallel readings, relevant images from first printings, and annotations that discuss the editing issues at hand in various sections, but the *VM* environment, like print, like all electronic collections and editions, is limited in representing

the *Doctor Faustus* work. For scholars the question is not shall we listen to the good angel and lay that damned book aside or listen to the bad angel and be lord and commander of these elements. The choice, as we know the story goes, is not that simple.

---

1. To maintain the parallel representation, Greg admits that he must "compromise" the "typographical arrangement" between the quartos in order to facilitate "the detailed comparison of the texts" (Greg 151).

## Bibliography

Bevington, David, and Eric Rasmussen, eds. *The Federalist*. Cleveland: Meridian Books (The World Publishing Company), 1956.

Binda, Hillary. *An Overview of this Electronic Doctor Faustus*. Accessed 2005-03-21. <<http://www.perseus.tufts.edu/Texts/faustus2.html>>

Greg, W.W., ed. *Marlowe's Doctor Faustus 1604-1616: Parallel Texts*. Oxford: Clarendon Press, 1950.

Lavagnino, John. "Completeness and adequacy in text encoding." *The Literary Text in the Digital Age*. Ed. Richard J. Finneran. Ann Arbor: The University of Michigan Press, 1996.

Schreibman, Susan. "Computer-mediated Texts and Textuality: Theory and Practice." *A New Computer-Assisted Literary Criticism? Special edition of Computers and the Humanities* 36 (2002): 283-293.

Schreibman, Susan, Amit Kumar, and Jarom McDonald. "The Versioning Machine." *Literary and Linguistic Computing* 18.1 (2003): 101-107.

Sutherland, Kathryn, ed. *Electronic Text: Investigations in Method and Theory*. Oxford: Clarendon Press, 1997.

## Text Modeling and Visualization with Network Graphs

---

*Aaron Coburn (acoburn@middlebury.edu)*

*National Institute for Technology and Liberal Education (Middlebury College)*

---

As greater quantities of textual source material become available to Humanities scholars, it becomes ever more challenging to retrieve, explore and manage these text collections (Lyman). Large, heterogeneous collections present further difficulties in attempts to represent the text visually. Metadata, particularly XML, is highly effective in adding structure to these collections and therefore aids in the process of locating and visually rendering the relationship between documents. Nevertheless, the process of creating high-quality metadata is both time-consuming and expensive, and this luxury is not always available.

This presentation will describe a project that is investigating and implementing statistical and graph-theoretic techniques for identifying, classifying and representing the content of large document collections in the absence of metadata. Models of the collection can be displayed to show various types of relationships between documents and their content, including term-based concept maps across a set of texts, document similarity measures or visualizations of the interaction among the characters in a novel. Furthermore, this model allows users to query the collection, in most cases with better recall and precision than a full-text keyword search.

## Constructing a Network Graph

The basis of these investigations is in the representation of a document collection as a graph of interconnected nodes. Each node corresponds to either a document or a term appearing in the collection, and nodes are connected by edges according to the frequency of their co-occurrence. The list of term nodes is typically filtered to include only those grammatical constructs, such as nouns and noun phrases, which tend to carry more semantic information about the content of the document. Documents nodes connect only to term nodes and term nodes connect, likewise, only to document nodes. The connecting edges are weighted by several factors, including term frequency and a normalization for document size. Depending on the analysis being conducted, term nodes can represent single words, phrases, character names, locations, stylistic data or any

combination thereof. Likewise, document nodes may represent arbitrarily sized text blocks, from sentences to paragraphs to entire book-length works.

The index derived from the text collection is similar to the term-document matrix of a vector model, but it is interpreted differently. For instance, when the collection is being explored with natural language queries, it proceeds with a technique called spreading activation (Preese). Each term in the graph that appears in the search statement is initialized with an amount of activation energy. This value is dispersed along each edge according to the established weight of each connection, and additional nodes are activated with the remaining energy. This process repeats itself along the graph until the initial amount of energy has been completely dispersed. Those nodes with the highest energy levels at the end of this process are considered most relevant to the query, and the results can be sorted accordingly. The document nodes become the corresponding result set while the activated term list can be used as relevance feedback to the user: a guide to the semantic composition of the result set (Search::ContextGraph).

This process will not only find documents containing the terms in the query string, but also relevant documents in which the search terms do not appear. In full-text keyword searches, highly relevant documents that do not contain any of the search terms are necessarily excluded. With a graph-based technique, however, documents with many shared terms are considered more closely related, and therefore, the process of traversing the graph will mark those documents to be relevant even if they do not contain any of the original search terms. In this way, the query *arctic* would likely also return documents that contain words such as *polar*, *north*, *ice* and *tundra*, but which happen not to contain the term *arctic*. Relevance is determined based on the overall measure of connectedness between nodes, rather than whether a term exists in a document.

## Visualization

The graph-based model of the text collection is also useful in creating visualizations of document relationships. This requires, first, the creation of a distance matrix representing the degree of similarity between nodes, and second, the scaling of the matrix to two dimensions. The distance matrix is calculated by conducting traversals of the graph networks, either with a breadth-first algorithm or (in order to save processing time) a random walk of the graph. Then, in order to scale the graph structure to display on a screen, the system identifies smaller clusters of nodes, each of which is mapped to a reduced dimensional space. This technique of locally linear embedding tends to preserve the general structure of both local and global clusters of nodes from the original graph (Saul).

When the network of terms and documents are clustered in a reduced dimensional space, they can be rendered in a browser window, showing the relative similarity among nodes. This particular system will output a vector graphic map (SVG), and a user can both view and interact with the graph.

This technique has proved useful in several experiments to display the relationship and interaction among characters in literary texts. First, proper names are extracted with a part-of-speech tagger (Lingua::EN::Tagger); in the absence of metadata, variant forms can be combined during an interactive step. Term nodes, in this case, represent character names, while document nodes consist of either individual paragraphs or a window of adjacent paragraphs: the number of paragraphs in each 'window' typically ranges from one to three. Interaction between characters is based on the frequency in which individuals are named, and the strength of that connection is reflected in both the proximity of the respective terms and the strength of the connecting edge. In the attached example from *Master and Margarita* by Mikhail Bulgakov, the characters with greater amounts of interaction are clustered accordingly (see Figure 1). This type of visualization can also allow users not only to view the relationship among characters in a text, but also to provide clickable access to the very interactions to which the model refers. Resolving anaphora, however, remains a major challenge in assessing a truly accurate measure of character interaction.

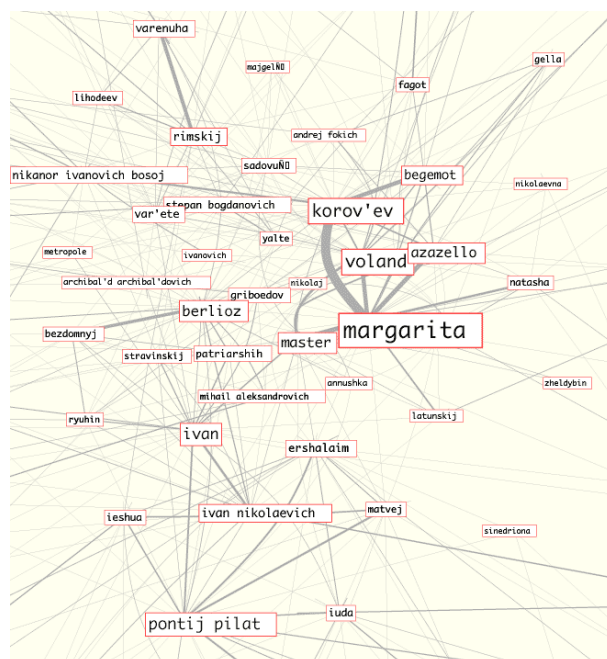


Figure 1

These models of text collections show promise in representing both content and stylistic similarity among texts. Furthermore, the ability to quickly and accurately search a collection of texts shows many advantages over a full-text search. In the absence

of metadata, this approach may provide a useful first step in navigating and managing any sufficiently large text collection.

## Bibliography

Bulgakov, Mikhail. *Master and Margarita*. n.p.: Penguin Classics, 2001.

*Lingua::EN::Tagger (Perl Module)*. Accessed 2005-04-07. <<http://search.cpan.org/~acoburn/Lingua-EN-Tagger-0.11/>>

Lyman, Peter, and Hal R. Varian. *How Much Information*. SIMS, University of California, Berkeley. Accessed 2005-04-07. <<http://www.sims.berkeley.edu/how-much-info-2003>>

Preese, Scott. *A Spreading Activation Model for Information Retrieval*. Ph.D. thesis, University of Illinois, 1981.

Saul, Lawrence, and Sam Roweis. *An Introduction to Locally Linear Embedding*. Accessed 2004-06-30. <<http://www.cs.toronto.edu/~roweis/lle/papers/lleintro4.pdf>>

*Search::ContextGraph (Perl Module)*. Accessed 2005-04-07. <<http://search.cpan.org/~mceglows/Search-ContextGraph-0.15/>>

---

## In the Philosophy Room: Australian Realism and the Digital Content Object

---

*Creagh Cole* ([c.cole@library.usyd.edu.au](mailto:c.cole@library.usyd.edu.au))

*University of Sydney Library*

*Paul Scifleet* ([p.scifleet@unsw.edu.au](mailto:p.scifleet@unsw.edu.au))

*University of New South Wales*

---

The recent digitisation of the papers and lecture notes of the Australian realist philosopher, Challis Professor of Philosophy John Anderson, has given us cause to reflect upon, on the one hand, the suitability of the TEI model for encoding digital documents and, on the other, the possibility that Anderson's philosophy itself may be relevant to some of the issues and debates in contemporary markup theory and practice.<sup>1</sup> The claim here is not that Anderson himself addressed, much less solved, the challenges we face in the construction of digital content, but that in our current situation reflection on ontological matters in this way may enlighten our thinking about the nature of the digital object and its descriptive encoding. We have come to think that the philosophical issues to be explored through this inquiry have a bearing on many of the more immediate empirical questions that we have previously raised. In this paper we seek to bring the philosophical debate to the fore. In an earlier presentation opening our case for the use of TEI in the description of digital library materials (Scifleet et al.), we argued that the collection and evaluation of information relating to actual markup practice from various institutions and research projects over time would further our understanding of difficult theoretical issues relating to the digital content object. The current paper extends this project to questions implicit in the TEI encoder's task of representing the text, which is increasingly seen to be a surprisingly problematic ambition. Our paper aims to contribute to philosophical debates of TEI encoding that have appeared in the work of a number of the leading theorists and practitioners in the field over the past decade.

Our study includes a brief review of debates on TEI ranging from McGann to Renear.<sup>2</sup> The trajectory of these debates suggests we may no longer be in the "progressive research program" that we had imagined. Many of the criticisms made of the ordered hierarchical model for encoding humanities texts have made an impression. On the other hand, criticisms from literary scholars of the descriptive encoding model are not



warranted in asserting the purely interpretive or constitutive nature of the encoder's task. Although the notion of representing the text in digital form is unclear, descriptive encoding is not interpretation all the way down. Textual features identified by our markup practices do have a reality independent of our thinking so and we do seem to be recording real and significant features in our assignment of tags to the digital content object. Nonetheless, there are real problems in practice in digitising materials such as the Anderson lecture notes and our project is driven by a desire to work some conceptual confusions through in a theoretically satisfying manner. Encoding is not a simple matter of reading off or copying textual features waiting to be recorded in digital form, for determinations about the nature of an object must be made. There seems to us to be a clear need for more information and guidance based on analysis and evaluation of actual markup practices over time. In place of this kind of guidance and engagement with real models, TEI proponents are left to gauge the extent to which they have failed to represent some ideal abstract object through resort to "tag abuse" and other coping mechanisms.

John Anderson is generally recognised as the most original and important philosopher to have worked in Australia. Between 1927 and 1958 he lectured in the Philosophy Room at the University of Sydney. His lecture notes in the Archives are acknowledged to be amongst our most important records of his philosophical thinking. Anderson developed a systematic realism which fostered a tradition of thinking about properties, qualities and relations which would seem to us to have some relevance to the encoder's world of elements, attributes and structural relations. Many of the recent debates on TEI and the descriptive encoding model have centred on our understanding of these seemingly intuitive concepts and the hierarchical structures they commit us to. We think that Anderson's insistence upon ontological seriousness and objective inquiry may help to illuminate many of the assertions about TEI's role in representing the text and that our problems may be clarified by establishing more clearly the ontological commitments of the various disputants. It would not surprise us to find that much of our conceptual thinking about these issues has been insufficiently critical in Anderson's sense.

In this paper we examine many of the current debates in the light of our understanding of Anderson's work: issues relating to the reality of the text; the descriptive and prescriptive distinction manifest in markup; whether identifying textual features is really an imposition upon the text and what this view might commit us to; whether TEI's acceptance of the markup language model leaves it unable to adequately represent imaginative, materially inscribed documents, as opposed to purely informational manuals and so forth. It is easy to take a distanced view of these issues and assume they don't directly affect the practice of encoding either in the digital library or the scholarly editing environments. However, the peculiar

nature of the TEI use of markup does seem to consistently raise issues we thought had been resolved, or which seemed to present no real constraint on practice. There is evidently some room for conceptual clarification here and it is possible that we have something to learn by adopting an attitude of "ontological seriousness" in relation to how we think about our markup practices. In that case, there may be some value in considering the lessons of an older philosophical tradition as practiced in John Anderson's Philosophy Room.

- 
1. The John Anderson Papers at the University of Sydney Library <<http://setis.library.usyd.edu.au/oztexts/anderson.html/>>
  2. Renear; Buzzetti; Caton; McGann; Huitfeldt; De Rose et al.; and Barwell et al.

## Bibliography

- Barwell, Grahame, Chris Tiffen, Phil Berry, and Paul Eggert. "The Authenticated Electronic Editions Project." *Computing Arts 2001: Digital Resources for Research in the Humanities*. Ed. Creagh Cole and Hugh Craig. Sydney: University of Sydney, 2003. 114-122.
- Buzzetti, Dino. "Digital Representation and the Text Model." *New Literary History* 33 (2002): 61-88.
- Caton, Paul. "Markup's Current Imbalance." *Markup Theory and Practice* 3.1 (2001): 1-13.
- DeRose, Steven J., David G. Durand, Elli Mylonas, and Allen Renear. "What is Text Really?" *Journal of Computing in Higher Education* 1.2 (1990): 3-26.
- Huitfeldt, Claus. "Multi Dimensional Texts in a One Dimensional Medium." *Computers and the Humanities* 28 (1995): 235-241.
- The John Anderson Papers at the University of Sydney Library*. Accessed 2005-03-21. <<http://setis.library.usyd.edu.au/oztexts/anderson.html/>>
- McGann, Jerome. *Radiant Textuality: Literature after the World Wide Web*. New York: Palgrave, 2001.
- Renear, Allen. "Out of Praxis: Three (Meta) Theories of Textuality." *Electronic Text: Investigations in Method and Theory*. Ed. Kathryn Sutherland. Oxford: Oxford University Press, 1997. 107-126.
- Scifleet, Paul, Creagh Cole, and Connie Wilson. "FRBR and Markup Analysis: A Common Denominator for Discourse." Paper delivered at Computing Arts 2004, Newcastle Australia. 2004.

## Developing the Humanities HyperMedia Centre @ Acadia University

---

**Richard Cunningham**

*(richard.cunningham@acadiau.ca)*

*Department of English, Acadia University*

**David Duke** *(david.duke@acadiau.ca)*

*Department of History & Classics, Acadia University*

**John Eustace** *(john.eustace@acadiau.ca)*

*Department of English, Acadia University*

**Anna Galway**

*Acadia student enrolled in HHC courses*

**Erin Patterson** *(erin.patterson@acadiau.ca)*

*Vaughan Memorial Library, Acadia University*

---

**W**e propose a panel session of five speakers, one of whom will act as chair also, to discuss the process of implementing a new initiative in humanities computing, the Humanities Hypermedia Centre, that has been under development at Acadia University since the fall of 2002.

At Acadia all faculty and students are issued with the same model of laptop computer outfitted with a common software template in order to enhance networking capabilities and ensure that all students have the opportunity to become proficient in the use of that software. Through funding from the McConnell Family Foundation, the Humanities HyperMedia Centre (HHC) was implemented to ensure that humanities students in particular were given opportunities to create their own new media projects and publish them in a dedicated database, the Acadia Humanities HyperMedia Archive, or AhHa!

AhHa! will be a database of digital objects and new media projects created primarily by Acadia students. It will give students access to each other's work and provide them with opportunities to contribute-and to know they are contributing-to the creation of a substantial, verified, body of work. It will also enable them to compile sophisticated examples of their own work for display when they leave Acadia. AhHa! is designed to allow faculty members to share teaching and research material in a more efficient manner than has been available previously.

It is the goal of the HHC to ensure that students graduate with a firm grounding in the Arts and Humanities, as well as a high level of information literacy and cutting edge skills in digital communication.

We have learned - sometimes the hard way - that developing and implementing a complex initiative involving six academic units (Classics, English, History, Philosophy, the university library system, and the Acadia Institute for Teaching Technology-an in-house technological development office) is a lengthy, challenging, and often surprising process. By the time of the ACH/ALLC Conference some of the components of the HHC will be up and running, whilst others will still be awaiting their launch. We therefore intend to offer the panel both to discuss the challenges we have faced, and the discoveries we have made, as we evolved the project from proposal stage to implementation, and also to solicit advice from other teams that have undergone or are undertaking the same process. Specifically we intend to address issues such as:

- The logistics of pedagogy - how do we assure program viability by gaining long-term financial and teaching commitments from teaching units, support units, and administration?
- The pitfalls of program development - unforeseen problems in proposing, developing, and setting up the administrative support for a new Multidisciplinary Minor in Hypermedia.
- Factors that contribute to the robustness of a program such as the HHC - and factors that contribute to its fragility. On the positive side these can be long term financial commitments from granting agencies; recognition from administrative planning sectors that new, multidisciplinary projects like HHC require a rethink of "faculty complement", especially on a departmental basis; ongoing technological support that ensures synergy between pedagogy and technological development, both hardware and software; and an administrative recognition that the development of innovative projects such as HHC need to be factored into the career review of those involved. On the negative side, bureaucratic neglect, problems of working within the short-term planning cycle common to most universities (and especially to most university departments), departmental parochialism and even outright competitiveness, and unanticipated events - such as labour actions or funding changes - can all affect the long-term viability of a program, however agile its implementation.
- The impact of working on such a project for a group of young, relatively junior and recently-appointed faculty members and technical personnel.

We therefore propose to offer a panel consisting of four members of the HHC team and a student who enrolled in both of the first two courses offered by the HHC. Each of the team members will discuss the process of development and

implementation from a different perspective, while our student will offer observations on the challenges and rewards of participating in the HHC.

**Richard Cunningham:** The team leader and point-person whose name is most closely associated with the program university-wide. His presentation will focus on the senior liaison aspects of the program, both with administration and with component departments and non-academic partners.

**David Duke:** A team member who has been active primarily on the pedagogical side of the program. His presentation will focus on the pressures of developing a "non-departmental" offering, and will discuss the opportunities inherent in this particular multidisciplinary project.

**John Eustace:** A team member who was particularly active in the early phase of the program, John left on a year-long sabbatical and has returned to continue his association with the project. His absence and return allows for a unique perspective from an "outsider-insider" who can comment on the nature of progress in a project such as this.

**Anna Galway:** A student in Sixteenth-century Literature and in Twentieth-century English Literature and Culture, the first two classes to be offered as HHC courses, will speak to the experience of learning to create hypertextual documents and submit them to a database still undergoing beta-testing.

**Erin Patterson:** A team member who has been active on both the pedagogical and technical sides of the project, whose background is not classroom-based but library science. Her presentation will cover the opportunities and pitfalls inherent in involving a university sector that has been traditionally ancillary to the direct pedagogical or curriculum development sectors of the university.

---

## Using Ancillary Text to Index Web-Based Multimedia Objects

---

*Lyne Da Sylva* ([lyne.da.sylva@umontreal.ca](mailto:lyne.da.sylva@umontreal.ca))

*EBSI, Université de Montréal*

*James Turner* ([james.turner@umontreal.ca](mailto:james.turner@umontreal.ca))

*EBSI, Université de Montréal*

---

*PériCulture* is the name of a research project at the Université de Montréal which is part of a larger project based at the Université de Sherbrooke. The parent project aimed to form a research network for managing Canadian digital cultural content. The project was financed by Canadian Heritage and was conducted during the fiscal year 2003-2004. *PériCulture* takes its name from *péritexte* and culture, *péritexte* being one of a number of terms used (in French, our working language) to mean ancillary text associated with images and sound. It is a sister project to *DigiCulture*, another part of the same larger research project which studied user behaviours in interactions with Canadian digital cultural content. The general research objective of *PériCulture* was to study indexing methods for Web-based nontextual cultural content, specifically still images, video, and sound. Specific objectives included:

1. identifying properties of ancillary text useful for indexing;
2. comparing various combinations of these properties in terms of performance in retrieval;
3. contributing to the development of bilingual and multilingual searching environments;
4. developing retrieval strategies using ancillary text and synonyms of useful terms found therein.

In computer science, research into indexing images and sound focuses on the low-level approach, performing statistical manipulations on primitives in order to identify semantic content. This approach is also referred to as the 'content-based approach' (e.g. Gupta and Jain, Lew). In information science, research into indexing images and sound focuses on associating textual information with the nontextual elements, and this often involves manipulating ancillary text. This approach is referred to as the 'high-level' or 'concept-based approach' (e.g. Rasmussen, O'Connor, O'Connor, and Abbas). A number of factors militate in favour of automating the high-level approach as much as possible. These include the very large volume of Web-based materials available, the disparity among cataloguing and indexing methods from one collection to another, and the high cost and relative inconsistency of human indexing.

Our work in this project focuses on text associated with Web-based still images, and builds on previous work in this area of information science (e.g. Goodrum and Spink, Jörgensen, Jörgensen et al., Turner and Hudon). We identified a number of Web sites that met our criteria, i.e., that contained multimedia objects, that had text associated with these objects that was broader than file names and captions, that were bilingual (English and French), and that housed Canadian digital cultural content. We identified keywords that were useful in indexing and studied their proximity to the object described. We looked at indexing information contained in the `Meta` and `Alt` tags, and whether other tags contained useful indexing terms. We studied whether standards such as the Dublin Core were used. We identified Web-based resources for gathering synonyms for the keywords.

Our study found that a large number of useful indexing terms are available in the ancillary text of many Web sites with cultural content. We evaluated various types of ancillary text as to their usefulness in retrieval. Our results suggest that these terms can be manipulated in a number of ways in automated retrieval systems to improve search results. Cross-language comparison of the results reinforces our previous research results, which suggest that indexing in other languages can be generated automatically from a single language using Web-based tools.

Rich information that can be used for retrieval is available in many places on Web sites with cultural content, from the file name to explicit information in captions to descriptive information in surrounding text to the contents of various HTML tags. Algorithms need to be developed to exploit this information in order to improve retrieval.

Finally, we feel that our work is useful because of the synergy created by the approaches we use. We are both interested in image indexing, but come from different fields. Lyne Da Sylva's expertise is in linguistics and James Turner's in information science. By working together, we are able to pool our knowledge and develop richer methods than would otherwise be available to either of us for approaching the question of automating indexing for images and other multimedia objects.

## Bibliography

Goodrum, A., and A. Spink. "Image searching on the Excite web search engine." *Information Processing and Management* 27.2 (2001): 295-312.

Gupta, A., and Ramesh C. Jain. " Visual information retrieval." *Communications of the ACM* 40.5 (71-79): 71-79.

Jörgensen, Corinne. *Image attributes: an investigation*. PhD thesis, Syracuse University, 1995.

Jörgensen, Corinne. "Image attributes in describing tasks: an investigation." *Information Processing and Management* 34.2/3 (1998): 161-174.

Jörgensen, Corinne, Alejandro Jaimes, Ana B. Benitez, and Shih-Fu Chang. "A conceptual framework and empirical research for classifying visual descriptors." *Journal of the American Society for Information Science and Technology (JASIST)* 52.11 (2001): 938-947.

Lew, Michael S. *Principles of visual information retrieval*. New York: Springer, 2001.

O'Connor, Brian C., Mary K. O'Connor, and June M. Abbas. "User reactions as access mechanism: an exploration based upon captions for images." *Journal of the American Society for Information Science* 50.8 (1999): 681-697.

Rasmussen, Edie M. "Indexing images." *Annual Review of Information Science and Technology* 32 (2004): 169-196.

Turner, James M., and Michèle Hudon. "Multilingual metadata for moving image databases: preliminary results." *L'avancement du savoir : élargir les horizons des sciences de l'information, Travaux du 30e congrès annuel de l'Association canadienne des sciences de l'information*. Ed. Lynne C. Howarth, Christopher Cronin and Anna T. Slawek. Toronto, 2002. 34-45.

# A la Carte Schema: A Case Study Comparison of the Application of DTDs and XML Schema to the Carte Calendar Project Template

*Ingrid Daneker* ([idaneker@ukonline.co.uk](mailto:idaneker@ukonline.co.uk))

*School of Library, Archive and Information  
Studies, UCL*

*Claire Warwick* ([c.warwick@ucl.ac.uk](mailto:c.warwick@ucl.ac.uk))

*School of Library, Archive and Information  
Studies, UCL*

## Introduction

This proposal describes research which compares the validation capabilities of conventional DTDs and XSD Schemas using the *Carte Calendar Project* template of the Oxford Digital Library (ODL) as a case study.

The *Carte Calendar* is a manuscript consisting of 75 volumes, cataloguing political papers on English and Irish history of the 17th century, originally collected and put together in chronological order by Thomas Carte, and converted into Library records by the Librarian Edward Edwards in the 19th century. The ODL is in the process of creating a digital record of Edwards' catalogue of Carte's collection, covering the period between 1660 and 1688, converting about 17,500 individual manuscript pages into XML-encoded transcripts using an EAD-DTD (1998) based transcription template.

The manuscript contains structured data which is repeated on almost every page, such as the shelf-mark, the unit-date, a document number, and a page number, as indicated by the keying sample below, showing the content of the <unittitle> element.

```
<!--shelfmark-->
<unitid type="shelfmark">MS. Carte 45,
fol(s). 67</unitid>
<!--document number-->
<unitid type="docno">34</unitid>
<!--pencil page number-->
<unitid type="page">451</unitid>
<!--red number-->
<unitid type="redno">?</unitid>
```

```
<unitdate> 22 May 1661</unitdate>
<physdesc>
<genreform>Holograph</genreform>
</physdesc>
```

These data-fields lend themselves well to the application of user-defined data-types, which Schema language can facilitate, thus making possible the automated verification of the data.

At any one time four freelance keying operators with varying levels of experience and three permanent staff might be working on the transcription of the project. Consistency of data entry is an ongoing concern, as it requires not only familiarity with the author's 19th century hand-writing and the subject matter, but also a good knowledge of the historical background.

The aim of this research was therefore to test whether the use of XML Schema could help to improve the accuracy of data entry by introducing stricter validation constraints than are mandated by the EAD-DTD. It will highlight some of the benefits of Schema definition language for the document type definition of the Carte template, compared to conventional DTD syntax. It will also consider whether the adoption of Schema based document type definitions is easily achievable for users of encoding standards such as EAD and TEI, who may have little prior knowledge of Schema.

This research is particularly timely, since in June 2004, Daniel Pitti announced that work was about to begin on developing one or more official EAD schemas (Pitti, 2004), as it provides a case study of how such a Schema might work in practice, and could inform work on the much broader EAD schema standard.

## Methodology

The researcher analysed the predefined element structure of the *Carte* project, keying template and the intellectual content provided by the Carte Manuscript to determine areas which would potentially benefit from the addition of Schema definitions, by restricting allowable (valid) data-input and element/attribute use. A template-specific DTD was then created based on SGML/EBNF syntax, containing specifications for only those elements/attributes of the EAD standard used in the template. Although significant changes to the original tag-set of the template were avoided in order to remain backward compatible with the original EAD-DTD, some alterations had to be made so that the addition of Schema's data-typing capabilities could later be facilitated.

The 'bespoke' DTD applied much stricter definitions than the original EAD-DTD with regards to the use of allowable elements/attributes in the template, the number of occurrences and their possible location within the document. This allows for stricter validation of the document structure and content,

providing more rigid constraints for keying staff, whose objective is to produce a valid document following the entry of data.

The new DTD was converted into XSD Schema language using XML editing software; *oXygen* XML editor 4.2 and *Altova's XMLSPY 2004 Enterprise Edition* (XMLSPY v2004 rel. 4 U). These specific tools were chosen because other packages such as the *XMetal Home Edition* (version 2004) provided limited support for XSD files which contain, for instance, multiple namespaces.

With the aim of applying even stricter rules to the allowable elements/attributes of the template's metadata structure, the researcher then introduced more rigid input data specifications to the resulting Schema file, facilitated by the additional capabilities of Schema definition 'language', which goes well beyond the potential of conventional DTDs. Further, in order to explore the benefits of the inbuilt namespace support Schema offers, two additional Schema files were created and associated with the main Schema using the `<xs:import>` element.

## Findings

### Use of the new DTD

The new project-specific DTD ensured a higher level of uniformity in structure and data entry verification. It did so by stipulating, where possible, the order and number of occurrences of EAD elements used in the template, as well as by specifying the requirement of all the attributes and possible attribute values. It also eliminated the very imprecise and inflexible definition structure of mixed content specifications, determined by the underlying syntax of conventional DTDs. This project-specific DTD should, however, only be used for the validation stage of the project. To assure interoperability and compatibility with other projects, such as its inclusion in the A2A (Access to Archives – <http://www.a2a.org.uk>) records, the instance document must remain backward compatible with the official EAD-DTD standard. Thus the more generic DTD file must be associated with the template again after the proofreading process.

### Benefits of Schema

The Schema allowed for the use of data-typing to increase the control over data entry by applying stricter validation parameters to the instance document. For example, when a user-defined global simple type specification such as 'content' (see the 'content' type definition below) is included in the element definition of, for example `<geogname>`, it ensures that an element must have content.

### The user-defined type definition 'content':

```
<xs:simpleType name="content">
  <xs:restriction base="xs:string">
    <xs:pattern
      value="([a-zA-Z?{$}]+(\.))*"></xs:pattern>
    </xs:restriction>
  </xs:simpleType>
```

This is not enforceable using DTD syntax, which allows the user either to define empty elements or to specify parsable character data (`#PCDATA`), which includes whitespace. The inclusion of the 'content' datatype in an element's content model will make sure that — should a transcriber forget to enter data for the element in question — validation against the Schema produces an error message pointing at the element's location in the instance document.

Further benefits were highlighted by the use of namespaces, which can help to avoid potential ambiguity problems, particularly where document exchange and integration is concerned. Schema allows the user to declare different namespaces, which can then be applied to individual element declarations, and thus used to differentiate between the varying element content models of the same element. In the *Carte* Template, for example, the `<num>` element has four different instances. This ambiguity was resolved by making use of additional namespaces for three of the four `<num>` elements.

Where different names describe the same set of data, Schema offers document authors the opportunity to define substitution groups which allow for increased compatibility of different instance documents. In the case of the *Carte* project this occurs with regard to `<genreform>` and `<physfacet>`, which are both allowable child elements of `<physdesc>` according to the original EAD-DTD.

On a more basic level, unlike DTD syntax, Schema does allow for sequencing of child elements, in a mixed content element model. Occurrence indicators for elements can be set much more specifically through the use of numerical values ranging from 0 to infinity, and global type definitions guarantee re-usability of individual type definitions and element declarations. All of these features proved useful in the *Carte* case study, and will be described in the proposed paper.

## Conclusion

The Schema-based keying template was tested by three experienced encoders at SERS (Systems and Electronic Resources Service). Their data-input would have validated against the original EAD-DTD. However, on average 3-4 errors were found by the *oXygen* 4.2 parser, pointing at entry format inconsistencies. For example if the unitdate was entered as *1*

*January 1661* instead of *01 January 1661* (or *10 January 1661*), the validator highlighted the missing digit as an error, in accordance with the pre-defined entry parameters provided by the underlying Schema. Similarly, missing content for the element `<geogname>` caused an error message (as the Schema's element declaration includes the global type 'content', discussed above).

These specifications were incorporated in the Schema to make sure that the encoder double checks the manuscript for such data fields, and testing proved it had been successful. Nevertheless a balance must be maintained between accuracy and speed of transcription in the case of the *Carte* project. Extensive data-typing of the `shelfmark`, `docno`, `pencilpage`, `redno` and `unitdate` fields might cause transcribers to spend more time trying to figure out which data format creates a valid document than is spent with the actual data input for each individual record.

The experimental DTD and resulting Schema have proved successful in helping transcribers to avoid errors and improve consistency. Yet despite the benefits of using Schema language over DTD syntax for XML document declarations, its complexity and to the untrained user's eye rather complicated element content modelling structure, underpinned by the W3C standard recommendation, might discourage potential users from trying to learn it. Nevertheless, this research has shown that software already available can help users to make the transition from DTD to Schema. While the needs of individual projects must always be considered carefully, the case study of the *Carte* project template shows that the benefits of Schema use should be taken seriously, despite its complexity.

## Bibliography

- Costello, Roger L. *XML Technologies Course, XML-Schemas: A downloadable schema tutorial*. xFront, 2003. Accessed 2005-02-27. <http://www.xfront.com/xml-schema.html>
- Deitel, H.M., et al. *XML: how to program*. Upper Saddle River, NJ: Prentice Hall, 2000.
- Encoded Archival Description. *EAD Tag Library for Version 1.0*. Accessed 2005-03-21. <http://www.loc.gov/ead/tglib1998/>
- LEADERS – Project (October 2001 to March 2004: Linking EAD to Electronically Retrievable Sources). Accessed 2005-03-21. <http://www.ucl.ac.uk/leaders-project/>
- Mertz, David. "TEI - the Text Encoding Initiative": An XML dialect for archival and complex documents. September 4, 2003. Accessed 2005-02-27. <http://www-106.ibm.com/developerworks/library/x-matters30.html>
- Pitti, Daniel. *EAD and W3C XML Schema*. Cover Pages. Accessed 2005-02-27. <http://xml.coverpages.org/ead.html>
- Pitti, Daniel. Encoded Archival Description List (EAD@LISTSERV.LOC.GOV), 4 June 2004. <http://listserv.loc.gov/cgi-bin/wa?A2=ind0406&L=ead&P=R605&D=0&I=-3>
- Rahitz, Sebastian. *Converting to schema: the TEI and Relax NG*. Text Encoding Initiative. Accessed 2005-03-21. <http://www.tei-c.org.uk/Talks/xmleurope2002/>
- Sperberg-McQueen, C.M., and Lou Burnard. *Tel P4: Guidelines for Electronic Text Encoding and Interchange*. Text Encoding Initiative, 2001. Accessed 2005-02-27. <http://www.tei-c.org/P4X>
- Tennant, Roy, ed. *XML in libraries*. New York: Neal-Schuman Publishers, Inc., 2002.
- Thompson, Henry S. "XML Schema types and equivalence classes." Paper presented at the XML Europe 2000 conference. 2000. Accessed 2005-02-27. <http://www.gca.org/papers/xmleurope2000/papers/s06-01.html>
- Van der Vlist, Eric. *Using W3C XML Schema*. O'Reilly xml.com. Accessed 2005-02-27. <http://www.xml.com/pub/a/2000/11/29/schemas/part1.html?page=1>
- Watt, Andrew, and R. Allen Wyke. *XML Schema Essentials (e-book)*. New York: John Wiley & Sons, Inc., 2002.
- Women Writers Project. "Encoding Practice." Rhode Island: Brown University. Accessed 2005-02-27. <http://www.wwp.brown.edu/encoding/research/NASSR/WWP.html#Heading3>
- Women Writers Project. "Methodological Issues." Rhode Island: Brown University. Accessed 2005-02-27. <http://www.wwp.brown.edu/encoding/research/NASSR/WWP.html#Heading4>

# Modelling Complex Multimedia Relationships in the Humanities Computing Context: Are Dublin Core and FRBR up to the Task?

---

**J. Stephen Downie** (*jdownie@uiuc.edu*)

University of Illinois at Urbana-Champaign

**Allen Renear** (*reinear@uiuc.edu*)

University of Illinois at Urbana-Champaign

**Adam Mathes** (*adam@adammathes.com*)

University of Illinois at Urbana-Champaign

**Karen Medina** (*kmedina@alexia.lis.uiuc.edu*)

University of Illinois at Urbana-Champaign

**David Dubin** (*ddubin@uiuc.edu*)

University of Illinois at Urbana-Champaign

**Jin Ha Lee** (*jinlee1@uiuc.edu*)

University of Illinois at Urbana-Champaign

---

## Introduction

It is now widely recognized that the creation, management, and analysis of content other than text is extremely important if the digital humanities are to deliver access to, and provide an analytical purchase on, the full range of human culture. However it is not clear to us whether the cataloguing and classification systems for digital content are up to the task. Difficulties in this area threaten to impede both the development of tools and techniques — and the production of sound theoretical results. In our paper we discuss some of these problems, focusing on *relationships* amongst the various cultural modes of expression. With the intention of convening a larger discussion of how these confusions might be remedied, we then propose directions for some clarification and improvement. However, the larger issues here are not merely terminological and resist any easy resolution.

## The Problem

Within the humanities computing community it has been a commonplace that while the emphasis on representing

and analyzing textual content may be understandable, it is important to support the other kinds of content as well. We agree. The 'digital humanities' must support the full range of human cultural products: text, music, images, dance, cinema, architecture, design, and so on. At present there are many different research communities looking into the organization of, and enhanced access to, these various modes of cultural expression. There is a text retrieval community (see Baez-Yates & Ribeiro-Neto), a growing music information retrieval community (see Futrelle & Downie), an image retrieval community (see Hsin-liang & Rasmussen), and so on. Notwithstanding the real progress being made by each of these, very astonishingly little work has yet been done to comprehensively address the issue that each of these individual modes of expression interact with each other in the ordinary course of production, management and use, as well as how formats at varying level of abstraction interact within a single modality.

First, to illustrate how the modes of expression interact with each other, let us consider the *Othello* corpus. An incomplete inventory of the *Othello* corpus includes the novella by Giraldi Cinthio (1565) "upon which Shakespeare based his play" (Hunt), Shakespeare's play (1604), the operas by Rossini (1816) and Verdi (1887), Dvorak's concert overture, Op. 93 (1892), and the ballet by Lubovitch (2002). If we are going to create a digital humanities repository worthy of use by humanities scholars and their students, it is incumbent on us to build a system that can 'collocate', or gather up, all extant digital representations of *Othello*: all recordings, all scores, all movies, all choreographies, all libretti, all scripts, all set and costume designs, all critiques, and so on. To aid in this collocation, we need to clearly express the relationships between each of these things at both the specific and generic levels. On the specific level, we need to indicate that, for example, *Othello* choreographic labanotation *W* is directly based on *Othello* score *X*, which was specifically used in *Othello* movie *Y*, and also released in *Othello* soundtrack recording *Z*. On the generic level, we need to indicate that all *Othello* scores have some generic relationship to all *Othello* recordings, to all *Othello* movies, etc. in such a way that explicates that the works are all members of the *Othello* corpus.

Second, to illustrate interactions between formats within a single mode, consider only the music mode of the *Othello* corpus. For each musical realization there usually exists a symbolic score and its individual parts. These symbolic representations can, in turn, be represented in a variety of digital formats: MusicXML, TIFF, *Finale*, etc. The aural aspect of the music is represented in another variety of digital formats: WAV, MP3, Ogg Vorbis, etc. Again, complex relationships exist between the 'symbolic' and 'aural' representations at both the specific (e.g., recording *X* used score *Y*) and generic levels (e.g., a 'fakebook' score used to generate different recordings of



improvised renditions). Other potentially complex relationships exist because many of these formats can be used to generate the others. For example, a TIFF scan of the 'original' score can be fed through an Optical Music Recognition (OMR) system to create a MusicXML score file which can generate a MIDI file which then can generate any of the audio file formats. Further complicating matters, research is also underway to 'backwards' create scores from audio recordings which would capture, symbolically, the nuances of a given performance (e.g., Plumbley et al.).

## Standards for Expressing Relationships Among and Within Modes

There is, of course, a body of work — standards and related research — within the cataloguing and classification communities that holds some promise for supporting the relationships described above. The Dublin Core (DC) is perhaps the most widely used within the digital humanities. IFLA's *Functional Requirements for Bibliographic Records* (FRBR) is becoming increasingly important. Work by organizations devoted to specific modalities such as the Federation Internationale des Archives du Film (FIAF)<sup>1</sup>, and the International Association of Sound and Audiovisual Archives (IASA)<sup>2</sup>, as well as work by such researchers as Martha M. Yee (moving pictures — see Yee), and Richard Smiraglia (music — see Smiraglia), etc., are also contributing insights and theory to this research domain.

## Are We There Yet?

We have reviewed results from projects and analyses that suggest there is still much work to do before the functionality envisaged above is a reality. Here we describe one such project that attempts to use FRBR and the DC to support inter- and intra-modal relationships. The DC does in fact hold the most promise for representing these relationships in a way that enables computer supported exploitation for retrieval, navigation, analysis, and so on.

Ayres describes a project at MusicAustralia to use FRBR and DC to create a digital repository that explicates the complex relationships between the works, expressions, manifestations and items of a collection of music and lyrics found that:

The DC .Relation element can be used to display and support navigation between items with flat, horizontal relationships [i.e., inter-modal relationships like those between some music and its text]. However, the kinds of relationships MusicAustralia wants to expose are a combination of vertical [i.e., intra-modal relationships like those between a score and its recording] and

horizontal relationships, and rely heavily on abstract but well understood and demonstrable concepts of the Work and the Expression or version. At this stage, DC does not offer support exposure of navigational pathways that explicitly acknowledge both vertical and horizontal relationships. [Bracketed injections are ours.]

Indeed, a close look at Dublin Core format and type elements suggests that the level of precision, and subtlety required is probably not yet available there. For instance the DC type vocabulary includes such disparate things as 'sound', 'text' and 'physical object', and examples for 'sound' include 'music playback file format' and 'an audio compact disc' (DCMI Usage Board).

## Next Steps: Exploring Ayres' Open Questions

Because the work of Ayres and her colleagues represents the most thorough examination of the combination of FRBR modelling and Dublin Core encoding to build a comprehensive multimodal repository, we are taking it as the starting point for our present work. The Ayres study uncovers a series of unresolved open questions associated with FRBR and the modelling of real-world multimodal information. In the Ayres case, the two modes are music (i.e., scores, recordings, etc.) and text (i.e., lyrics, poems, etc.). These two modes come together to create what we commonly consider to be 'songs'. To paraphrase Ayre's first open question:

1. Should we model as the primary work:
  - (a) the music;
  - (b) the text; or,
  - (c) the combination of text and music?

Ayres clearly illustrates that each modelling approach above clarifies a specific set of relationships between the music compositions and the texts while at the same time obscuring other relationships. The examination of this question has implications beyond the simpler music-text modelling case. For example, what are the implications when we attempt to model more complex cases (e.g., the *Othello* corpus, a Hollywood musical, etc.) with their exponentially growing relationships between text (novellas, plays, libretti, etc), music (i.e., notations, recordings, etc.), choreography (i.e., notations, video), and so on? Our paper examines this very question. We also explore the broader ramifications of Ayre's three related subsidiary open questions:

2. Should all notated and performed expressions of music [or dance, or text, etc.] be modelled as a single expression category?

3. Should expressions themselves be further modelled to include sub-categories for notated and performed expressions?
4. Should performed expressions based on particular notated expressions be modelled as expressions of expressions?

By examining these fundamental questions, we intend to encourage a long-overdue conversation within the humanities computing community. Unless our representation schemes do justice to the multidimensional complexity of cultural content in all its modes of expression, we will not realize the full potential of digital humanities repositories.

Smiraglia, Richard. *The Nature of "a work": implications for the organization of knowledge*. Lanham, MD: Scarecrow Press, 2001.

Yee, Martha M. "What is a Work?" *The Principles and Future of AACR: Proceedings of the International Conference on the Principles and Future Development of AACR, Toronto, Ontario, Canada, October 23-25, 1997*. Ed. Jean Weihs. Ottawa: Canadian Library Association, 1998. 62-104.

- 
1. <http://www.fiafnet.org/uk/>
  2. <http://www.iasa-web.org/index.htm>

## Bibliography

Ayres, Marie-Louise. "MusicAustralia: Experiments with DC.Relation." Presented at DC-ANZ (Dublin Core in Australia and New Zealand) Conference in Canberra. February 2003. Accessed 2004-11-17. <http://www.nla.gov.au/nla/staffpaper/2003/ayres1.html>

Baez-Yates, R., and B. Ribeiro-Neto. *Modern information retrieval*. 1st ed. Reading, MA: Addison-Wesley, 1999.

DCMI Usage Board. *DCMI Type Vocabulary*. 2004. Accessed 2004-11-27. <http://dublincore.org/documents/2004/06/14/dcmi-type-vocabulary/>

*Functional Requirements for Bibliographic Records (FRBR)*. UBCIM Publications, 19. Accessed 2004-11-27. <http://www.ifla.org/VII/s13/frbr/frbr.htm>

Futrelle, Joe, and J. Stephen Downie. "Interdisciplinary Research Issues in Music Information Retrieval: ISMIR 2000-2002." *Journal of New Music Research* 32.2 (2003): 121-131.

Hsin-liang, Chen, and Edie M. Rasmussen. "Intellectual access to images." *Library Trends* 48.2 (1999): 291-302.

Hunt, Mary Ellen. *Review of San Francisco Ballet, "Othello". War Memorial Opera House, San Francisco, CA*. criticaldance.com, 2002. Accessed 2004-11-27. [http://www.criticaldance.com/reviews/2002/sfb-othello\\_020301.html](http://www.criticaldance.com/reviews/2002/sfb-othello_020301.html)

Plumbley, M.D., S.A. Abdallah, J.P. Bello, M.E. Davies, G. Monti, and M.B. Sandler. "Automatic Music Transcription and Audio Source Separation." *Cybernetics & Systems* 33.6 (2002): 603-627.

# A Revolutionary Approach to Humanities Computing?: Tools Development and the *D2K* Data-Mining Framework

**J. Stephen Downie** ([jdownie@uiuc.edu](mailto:jdownie@uiuc.edu))

University of Illinois at Urbana-Champaign

**John Unsworth** ([unsworth@uiuc.edu](mailto:unsworth@uiuc.edu))

University of Illinois at Urbana-Champaign

**Bei Yu** ([beiyu@uiuc.edu](mailto:beiyu@uiuc.edu))

University of Illinois at Urbana-Champaign

**David Tcheng** ([dtcheng@ncsa.uiuc.edu](mailto:dtcheng@ncsa.uiuc.edu))

University of Illinois at Urbana-Champaign

**Geoffrey Rockwell** ([georock@mcmaster.ca](mailto:georock@mcmaster.ca))

McMaster University

**Stephen J. Ramsay** ([sramsay@uga.edu](mailto:sramsay@uga.edu))

University of Georgia

## Introduction

A new set of humanities computing (HC) research projects are underway that could revolutionize how the HC community works together to build, use, and share HC tools. The set of projects under consideration all play a role in the development work currently being done to extend the *D2K* (Data-to-Knowledge)<sup>1</sup> data-mining framework into the realm of HC. **John Unsworth** and **Stephen J. Ramsay** were recently awarded a significant Andrew W. Mellon Foundation grant<sup>2</sup> to develop a suite of HC data-mining tools using *D2K* and its child framework, *T2K* (Text-to-Knowledge). Drs. Unsworth and Ramsay, along with research assistant, **Bei Yu**, are working closely with **Geoffrey Rockwell**. Dr. Rockwell is the project leader for the *CFI* (Canada Foundation for Innovation) funded project, *TAPoR* (Text Analysis Portal for Research)<sup>3</sup>, which is developing a text tool portal for researchers who work with electronic texts. **J. Stephen Downie** and **David Tcheng**, through their work in creating the *International Music Information Retrieval Systems Evaluation Laboratory* (*IMIRSEL*)<sup>4</sup>, are leading an international researchers group to develop another *D2K* child system called *M2K* ("Music-to-Knowledge"). This panel session demonstrates how

all of these projects come together to form a comprehensive whole. The session has four major themes designed, through presentations and demonstrations, to highlight individual the project components being developed and their collective impact on the future of HC research. These themes are:

1. *D2K* as the overarching framework
2. *T2K* and its ties to traditional text-based HC techniques
3. *M2K* and its ties to multi-media-based HC techniques
4. The issues surrounding the HC community's development, validation, distribution, and re-use of *D2K/T2K/M2K* modules.

## Participants

**J. Stephen Downie**, Graduate School of Library and Information Science (GSLIS), University of Illinois at Urbana-Champaign (UIUC)

**John Unsworth**, GSLIS, UIUC

**Bei Yu**, GSLIS, UIUC

**David Tcheng**, National Center for Supercomputing Applications (NCSA), UIUC

**Geoffrey Rockwell**, School of the Arts, McMaster University

**Stephen J. Ramsay**, Department of English, University of Georgia

## Presentations, Demonstrations, and Discussions (in order)

### Overview of the NORA (No One Remembers Acronyms) project

**John Unsworth**

For decades, humanities computing researchers have been developing software tools and statistical techniques for text analysis, but those same researchers have not succeeded in producing tools of interest to the majority of humanities researchers, nor (with the exception of some very recent work in the Canadian *TAPoR* project) have they produced tools that work over the web. Meanwhile, large collections of web-accessible structured texts in the humanities have been created and collected by libraries over the last fifteen years. During that same time period, with improvements database and other information technologies, data-mining has become a practical tool, albeit one mostly used in business applications. We believe data-mining (or more specifically, text-mining) techniques can be applied to digital library collections to discover unanticipated patterns, for further exploration either through traditional criticism or through web-based text analysis.

Existing humanities e-text collections from Virginia, Michigan, Indiana, North Carolina, and other research universities form the corpus for the project. *NORA* brings NCSA's *D2K* data-mining architecture to bear on the challenges of text-mining in digital libraries, with special emphasis on leveraging markup, and on visualizations as interface and as part of an iterative process of exploration.

## Introduction to the D2K framework

### David Tcheng

Released in 1999, *D2K* was developed by the Automated Learning Group (ALG) at NCSA. *D2K* has been used to solve many problems for both industry (e.g., Sears, Caterpillar, etc.) and government agencies (e.g., NSF, NASA, NIH, etc.). Academic uses include bioinformatics, seismology, hydrology, and astronomy. *D2K* uses a data flow paradigm where a 'program' is a network (directed graph) of processing modules. Modules can be 'primitive', defined as a single piece of source code that implements a single well defined task, or can be 'nested' meaning it is defined as a network of previously defined *D2K* modules. Decomposition of programs into modules that implement a well defined input-output relationship promotes the creation of reusable code. Nesting modules into higher-level modules helps to manage complexity. *D2K* parallelizes across any number different computers by simply running a copy of "D2K Server" on each available machine. The *D2K* software distribution comes as a basic *D2K* package, with core modules capable of doing general purpose data-mining, as well as such task-specific add-on packages as text analysis (T2K), image analysis (I2K), and now music analysis (M2K).

## Introduction to T2K

### Bei Yu

Similar to many data-mining tools, *T2K* has implemented a number of automatic classification and clustering algorithms. Compared to the commercial text mining tools, for example SAS Text Miner, *T2K* has richer NLP preprocessing tools, especially after its integration with GATE. Tools include: stemmer, tokenizer, PoS-tagger, data cleaning and named-entity extraction tools. The clustering visualization is tailored for thematic analysis. On one hand, *T2K* provides a text mining platform for the HC community. On the other hand, *T2K* is also a platform to automate the HC research results and thus facilitate their applications to the text mining community in general. For example, most of the text mining tasks are still topicality oriented, but the affect analysis has emerged in the last couple of years. The affect of a document includes the subjectivity/objectivity, the positive/neutral/negative attitude, and the strength of emotions, etc. Some researchers have adapted stylistic analysis techniques from HC to analyze customer reviews. The found non-thematic features can also

be used as predictors for document genre, readability, clarity and many other document properties.

## The TAPoR portal and D2K

### Geoffrey Rockwell

*TAPoR* has released an alpha of the portal and will have the beta ready by June 2005. The portal is designed to allow researchers to run tools (which can be local or remote web services) on texts (which can be local or remote.) The *TAPoR* portal has been designed to work with other systems like *D2K* in three ways:

1. Particular tools or chains of tools can be 'published' so that they are available as post-process tool right in the interface of another system. Thus one can have a button that appears on the appropriate results screens of a *D2K* process that allows the user to pass results to *TAPoR* tools.
2. The portal has been released as open source and we are working on models for projects to run customized versions of the portal that work within their environment.
3. The portal can initiate queries to remote systems and then pass results to other *TAPoR* tools. Thus users can see tools like *D2K* (where they have permission) within their portal account.

## The Tamarind project and D2K

### Stephen J. Ramsay

Tamarind began with the observation that the most basic text analysis procedure of all — search — does not typically operate on the text archive itself. It operates, rather, on a specially designed data structure (typically an inverted file or pat trie index) that contains string locations and byte offsets. Tamarind's primary goal is to facilitate access to analytical data gleaned from large-scale full text archives. Our working prototype of Tamarind, for example, can quickly generate a relational database of graph properties in a text which can in turn be mined for structural information about the texts in question. Tamarind creates a generalized database schema for holding text properties and allows you to specify this structure as one that should be isolated and loaded into the database. Work is proceeding on a module that will allow the user to load a Tamarind database with millions of word frequency data points drawn from several gigabytes of encoded data. Unlike existing tools, this newest module includes information about where those counts occur within the tag structure of the document (something that is impossible to do without the raw XML). For the purposes of this project, we intend to use *D2K* and *T2K* as the primary clients for Tamarind data stores.

## The M2K project

### J. Stephen Downie

M2K is being developed to provide the Music Information Retrieval (MIR) community with a mechanism to access a secure store of copyright-sensitive music materials in symbolic, audio and graphic formats. *M2K* is a set of open-source, music-specific, *D2K* modules jointly developed by members of the *IMIRSEL* project and the wider MIR community. *M2K* modules include such classic signal processing functions as Fast Fourier Transforms, Spectral Flux, etc. In combination with *D2K*'s built-in classification functions (e.g., Bayesian Networks, Decision Trees, etc.), the *M2K* modules allow MIR researchers to quickly construct and evaluate prototype MIR systems that perform such sophisticated tasks as genre recognition, artist identification, audio transcription, score analysis, and similarity clustering.

**John Unsworth** and **J. Stephen Downie** will lead a wrap-up and future work open-forum discussion: For ambitious, multi-institutional projects like those presented in this panel many issues arise that can affect the sustainability and impact of the projects. In particular, the issues surrounding the HC community's development, validation, distribution, and re-use of *D2K/T2K/M2K* modules will be addressed.

1. See <http://alg.ncsa.uiuc.edu/do/tools/d2k> .
2. See <http://www.news.uiuc.edu/news/04/1025me1lon.html> .
3. <http://www.tapor.ca/>
4. See <http://music-ir.org/evaluation> .

## A Declarative Framework for Modeling Pronunciation and Rhyme

*David Dubin* ([ddubin@uiuc.edu](mailto:ddubin@uiuc.edu))

*University of Illinois*

*David J. Birnbaum* ([djbpitt+@pitt.edu](mailto:djbpitt+@pitt.edu))

*University of Pittsburgh*

Encoding standards such as TEI give scholars a great deal of flexibility in annotating texts to meet the particular needs of a study or project. Researchers necessarily make choices about which features of a text to highlight, what kinds of additional information to add, and what facts are left to be inferred from other sources of evidence apart from markup.

Among the factors to be considered in designing or adopting text encoding procedures are the prospects for:

- data reuse and generalization beyond the scope of the one's current project,
- investigating new questions about the texts that hadn't been anticipated, and
- planning for the integration of texts using other markup schemes.

This paper discusses considerations motivating the ongoing design of a software framework for analysis of rhyming schemes in 19th century Russian poetry. At its simplest, the application receives as input poems marked up like the example in Figure 1 and produces output like the following:

```
?- run('poem.xml').
Line 001 rhymes with Line 003
Line 002 rhymes with Line 004
Line 003 rhymes with Line 001
Line 004 rhymes with Line 002

<POEM OPID="S1.100" LINESPAN="1-4"
COLPAGE="1.261" YEAR="1817"
MASCRRHYME="2" FEMRRHYME="2"
OTHERRRHYME="0" MASCUNRRHYME="0"
FEMUNRRHYME="0" OTHERUNRRHYME="0"
NOENDWORD="0" COL13="2.75">
<TITLE>Надпись на стене больницы</TITLE>
<LINE LINENO="001">Вот здесь лежит
больной студ<STRESS>е</STRESS>нт;</LINE>
```

```
<LINE LINENO="002">Его судьба
неумол<STRESS>и</STRESS>ма.</LINE>
<LINE LINENO="003">Несите прочь
медикам<STRESS>е</STRESS>нт:</LINE>
<LINE LINENO="004">Болезнь любви
неизлеч<STRESS>и</STRESS>ма!</LINE>
</POEM>
```

Figure 1

Programs producing output such as the example above could be written in any of a variety of different programming languages. They might employ different strategies for integrating linguistic and orthographic processing rules with evidence encoded more directly using markup. For example, we discuss an earlier approach to the current project in Adams & Birnbaum. Our current implementation, however, is written in Prolog as an application of the BECHAMEL system for markup semantics analysis (Dubin et al.). The motivation for this choice was our wish to plan from the beginning for extensions to other encoding schemes and generalization to other kinds of analysis.

Prolog is a declarative language of rules and assertions, and BECHAMEL is a collection of predicates supporting the declaration of object classes, properties, relations among objects, and the execution of inference rules based on information extracted from XML documents. An example of a Prolog clause is shown below: it is part of our application's logic for determining that two sequences of phonemes at the ends of a pair of orthographic lines all agree with each other (i.e., that the sounds at the end of the lines rhyme with each other).

```
all_agree(P1,P2) :- agree(P1,P2),
/* P1 is written using C1 */
relation_applies(written,[P1,C1]),
relation_applies(written,[P2,C2]),
/* C3 follows C1 */
relation_applies(follows,[C1,C3]),
relation_applies(follows,[C2,C4]),
relation_applies(written,[P3,C3]),
relation_applies(written,[P4,C4]),
all_agree(P3,P4).
```

In the clause, P1, P2, P3, and P4 are variables representing phoneme objects, and C1, C2,C3, and C4 are variables representing character objects. The predicate all\_agree(P1,P2) will be satisfied if each of the predicates following the implication sign can be satisfied. The logic of the clause can be read as follows:

Phonemes P1 and P2 'all agree' if phonemes P1-P4 are written using characters C1-C4, respectively, if C3 follows C1 in the orthographic line, if C4 follows C2 in the line, if P1 and P2 'agree' and if P3 and P4 'all agree.'

This clause is one of several in a recursive rule that attempts to match up corresponding phonemes in rhyming lines, starting from the line's stressed vowel.

A major advantage of Prolog's declarative approach is the flexibility to define logic for separate cases in separate clauses for the same rule. For example, in the clause shown above, it is presupposed that in both lines there will be a simple one-to-one mapping from characters in sequence to phonemes. But accommodating more complex cases need not complicate the expression of the simple case: if the simpler clause cannot be satisfied then Prolog's inference engine will search for a different clause of the same rule that can be satisfied.

Reasoning about poems like the one in the example above requires that we model their contents at both a phonemic and orthographic level. The rules for Russian pronunciation include not only the way that particular vowels and consonants sound, but also how those phonemes are cued by the way the text is written (as with, for example, the palatalizing effect of the soft sign and the soft vowel letters). The BECHAMEL system's definition predicates for object classes, properties, and relations gives us the ability to model each of these levels with declarations such as the following:

```
/* There are things called phonemes */
declare_class(phoneme).
/* There are things called vowels */
declare_class(vowel).
declare_class(consonant).
/* vowels are a kind of phoneme */
declare_subclass(vowel,phoneme).
declare_subclass(consonant,phoneme).
/* vowels may be stressed or not */
declare_property(vowel,stress,atom).
declare_property(consonant,palatalized,atom),
declare_relation(written,[phoneme,
letter]),
declare_class(character),
declare_property(character,id,atom),
declare_property(character,palatalizing,atom),
```

In our approach, the phonemic properties of vowels and consonants are distinct from the orthographic properties of characters, written words, and lines. But it can occasionally be convenient for users to ignore these distinctions, particularly in comparing different proposed models for the same data. We therefore aim to let the models govern as much processing of our raw data as possible. For example, both phonemic and orthographic properties of letters are recorded in a data file using the same predicate as shown below:

```
alph_table(1083,id,u043B).
alph_table(1083,name,e1).
alph_table(1083,charclass,consonant).
```

```

alph_table(1083,case,lower).
alph_table(1083,voicing,voiced).
alph_table(1083,place,alveolar).
alph_table(1083,manner,liquid).

```

In this example, `id`, `name`, `case`, and `charclass` are all properties of characters, while `voicing`, `place`, and `manner` are phonemic properties. As individual characters and phonemes are instantiated, they acquire only those properties that are appropriate for their class. This is accomplished through general-purpose rules that match on the basis of our property declarations. So if we were to decide (for example) that `name` should be a property of the phoneme rather than the character, we need only change the declaration, and the property value recorded in the data file would be assigned to phoneme objects rather than character objects.

BECHAMEL supports `superclass` and `subclass` relations, which allows us to declare that vowels and consonants are subclasses of `phoneme`, and that letters, marks, and spaces are subclasses of `character`. Since the conventional superclass/subclass relation can prove awkward in some situations, BECHAMEL includes class declaration expressions similar to those found in ontology languages such as OWL (W3C).

For example, `place` and `manner` of articulation in consonants may be used to describe classes of phonemes, not merely features of them (the class of stops, the class of velar consonants, etc.). It would be awkward to declare each consonant as a subclass of both its `place` and `manner` of articulation. Instead we use a BECHAMEL predicate that permits us to declare membership in a class based on the value assigned to a particular property:

```

declare_propclass(alveolar,place,alveolar).
declare_propclass(velar,place,velar).
declare_propclass(glottal,place,glottal).
declare_propclass(glide,manner,glide).
declare_propclass(liquid,manner,liquid).
declare_propclass(nasal,manner,nasal).

```

An `alveolar`, therefore, is anything that takes the value `alveolar` on its `place` property, a `nasal` anything that takes `nasal` for its `manner` property, and so on. These class identities are in addition to the one that instantiated the object. A related feature of BECHAMEL is the ability to define class membership based on a Boolean expression. The following example declares that an `obstruent` is either a `stop`, a `fricative`, or an `affricate`:

```

declare_boolean(obstruent,or(stop,
or(fricative,affricate))).

```

All of these features are employed with the aim of making our understandings, models, and simplifications regarding the rules

of Russian pronunciation as clear and as explicit as possible. We express them in the form of declarative rules so as not to entangle implementation details of our code with aspects of our model that should be open to criticism, revision, and extension. For example, our rule governing the devoicing of word-final obstruents states that if an obstruent `O` is written with word-final letter `L`, then `O` should take a value of `voiceless` on its `voicing` property (unless it already has that property value):

```

mrule2 :- isa(O, obstruent),
relation_applies(written,[O,L]),
property_applies(L,wordfinal,true),
not(property_applies(O,voicing,voiceless)),
apply_property(O,voicing,voiceless),!.

```

Taking this approach, even a limited application, such as determining which lines of a poem rhyme, requires a large number of these declarations and rules; there is a very real sense in which we are doing it 'the hard way'. But as a result our application-specific code is only about seven percent of the size of the supporting declarations and rules. In addition to strengthening our confidence in our framework's potential for generalization, the demands of our approach have been a helpful modeling exercise in their own right.

## Bibliography

- Adams, L.D., and D. Birnbaum. "Perspectives on computer programming for the humanities." *Text Technology* 7.1 (1997): 1-17.
- Dubin, D., C.M. Sperberg-McQueen, A. Renear, and C. Huitfeldt. "A logic programming environment for document semantics and inference." *Literary and Linguistic Computing* 18.2 (2003): 225-233.
- W3C. *OWL Web Ontology Language Overview (W3C Recommendation)*. 10 February 2004. Accessed 2005-04-04. <<http://www.w3.org/TR/2004/REC-owl-features-20040210/>>

# Cardplay, a New Textual Instrument

---

David Durand (*dgd@acm.org*)

Brown University

Noah Wardrip-Fruin (*nwf@brown.edu*)

---

## Introduction: What is textual play?

We are exploring what it means to 'play' textual literature. We don't mean, by this, playing games that incorporate textual material within their structure but rather textual and literary structures for which play is a primary means of interaction. We are conducting our exploration both as creators and scholars of digital media. This paper discusses a number of related issues — including the notion of "instrumental texts" discussed by electronic literature authors, the critical games proposed by Jerome McGann and Johanna Drucker, and what Markku Eskelinen has characterized as the challenge of ludology (the study of games) to traditional literary study. This discussion then becomes the background for describing a system we are building — *Cardplay* — its design goals, our authoring work within it, and its relationship to prior work in the hypertext and artificial intelligence communities.

## Texts and instruments

In the electronic writing community there has been increasing talk, in the last few of years, about the idea of instrumental texts (Cayley, Moulthrop, Wardrip-Fruin). An instrumental text is meant to be played, and provides affordances for such play, much as folk musical instruments do (as the frets on a guitar invite the production of the notes of the scale). Such texts can and should provide opportunities for practice and reward mastery. What is practiced and mastered — again, the analogy is drawn with musical instruments — is often presented as a physical discipline. Instrumental texts also show close resemblances to computer games in these ways. Given that most works presented as examples of instrumental texts always use the same material for their play (always, so to speak, 'play the same tune') the analogy with games may be the more accurate of the two. However, the type of engagement that authors hope to produce with instrumental texts may be more musical than game-like.

A textual instrument, on the other hand, is a tool for textual performance which may be used to play a variety of compositions. In this sense it is evocative of Thalia Field's figure of the "language piano" — something that one learns to play, and which may produce a much wider variety of texts than is the case for those projects normally discussed as instrumental texts. However, a textual instrument need not be like a prepared piano. The direct selection of text, rather than the manipulation of a non-linguistic device, can be its interface. And the relationship between a textual instrument's interface affordances and the possible textual outcomes need not be one-to-one at all levels (as it is with a piano's keys, though they may be played in many combinations). Gaining an intuitive understanding of how a textual instrument will react for a given composition is part of learning to play that piece. Compositions, here, consist of a body of text (and/or a means of acquiring text) and a set of tunings for the instrument(s) used, where a tuning is a particular configuration of the interaction mechanisms and settings for the procedures (along with any instructions on how these change over time).

We have previously built two<sup>1</sup> textual instruments — one, for performing local pre-parsed texts (and for which currently there is one composition, *Regime Change*), the other for playing live network RSS feeds of current news. Both of these operate using the logic of n-gram statistical models of text (first used in textual play by Claude Shannon) and exhibited strengths and weaknesses that might be expected from such a purely statistical approach. An intuitive understanding of how to get 'good' results from both works is possible, however, despite the fact that there is no pre-set goal to the interaction. Because of the aleatoric nature of the automatic processes in both works, and the size and opaqueness of a statistical model of even a short text, these instruments are easy to compose for, but relatively resistant to precise control, by author or reader. *Cardplay*, on the other hand, is designed to operate more directly out of human authorship (of texts and rules), with interaction techniques and infrastructural motifs more typical of hypertexts or rule-based artificial intelligence systems. The aim is a blend of these parallel (non-intersecting) but closely related sets of techniques.

## Playing with games

Interest in games has a long history within the literary community. The work of the Oulipo, for example, could be seen as a relatively recent entry in a series of authoring games stretching back through literary history. (Amusingly, the Oulipo have referred to those employing difficult authoring constraints before them as "anticipatory plagiarists" — and characterized themselves as rats who design the mazes from which they propose to escape.) Critical interpretation has also been characterized as a game. Warren Motte has done important work on 'playtexts'. However there has been little attention to



games in a more formal sense — games involving rules, moves, and outcomes. Perhaps this is because few literal games have, before the last couple of decades, contained much of literary interest.

Now this is changing, and rapidly. Critics from literary backgrounds are among the most active in the formation of the rising field of 'game studies' (or 'ludology'). Meanwhile, other critics have proposed means by which the metaphorical game of literary interpretation can be literalized — via the introduction of rules, moves, and outcomes into public acts of interpretation carried out in a computational media environment.

Among those from a literary background who are now helping create the field of game studies, we will primarily focus on Markku Eskelinen's recent work. While we might also fruitfully consider the work of Espen Aarseth, Lisbeth Klastrup, Susana Tosca, and Torill Mortensen, it is Eskelinen who has most clearly brought concepts from ludology into contact with the notions of "instrumental text" and "textual instrument". He points out the importance of overcoming the "fear of variety" in order to understand instruments fashioned precisely so that each reading is different. We must find methods of reading not only textual outcomes (which vary) but the systems that produce them (which remain consistent).

In another sense, the creation of systems for 'playing literature' — but for critical purposes, rather than artistic ones — has been a focus of the Speculative Computing Laboratory at the University of Virginia. Best known of these projects is *The Ivanhoe Game* first proposed by Jerome McGann and Johanna Drucker. Here some types of literary interpretation are formalized in a manner that would be recognizable to ludologists, even if they do not fully satisfy all formal definitions of the term 'game'.

## Playing cards for drama

In *Cardplay*, we are trying to create a textual instrument whose center of gravity is clearly literary, focused on the creation of a work, a play, that is in some senses conventionally literary, and yet to make the process of playing the work simultaneously be the process of playing a game in the most literal sense. In *Cardplay*, players manipulate virtual cards (each associated with text that is not fully visible to the players), in an attempt to win the card game (Solitaire is also possible). However, a successful play wins points when the card played interacts with other cards played to advance the creation of the script of a play, whose transcript accumulates and may be saved. Copyright in the result may be automatically granted to the winner of the game, by the program, on the successful completion of the game. Players of the game are thus in competition with each other to advance the story. Unlike many

interactive fictions, however, neither player is identified with a character in the ongoing story, nor is the plot of the story necessarily determinative of victory in the game.

Significant aspects of *Cardplay* are inspired by the description of Mark Bernstein's systems *Thespis* and *Card Shark*. In Bernstein's *Card Shark*, players create texts by playing 'cards' each containing a fragment of narrative. Each card may have some named properties, which are active once the card is played. Cards may also have preconditions which must match the properties of active cards. *Thespis* extended this notion to a self-composing drama system in which a number of artificial agents try to play their own cards, with a similar condition system.

In neither of Bernstein's systems was the notion of a game used. In *Thespis*, a number of standard AI techniques (Blackboard systems, Agents) are used in a minimal way to create a reading experience. *Cardplay* cards can be divided into two types: Fundamental cards, which represent aspects of events, characters, places; and Master cards, which create textual content in the transcript. There is no procedural aspect to *Cardplay* cards, unlike Bernstein's *Thespis* agents, and the conditions by which cards are matched are more complex than those in *Card Shark*. A *Cardplay* card is more like a logical rule in an AI system, which has variables that it can match in the cards 'on deck', and results that it presents, to which other cards can match. When played, a Master card, and the cards that it has matched with, are all removed, and the transcript is augmented with the results.

We believe that the methods of symbolic AI provide a fertile area for exploration in the creation of literary systems and games. The thorny AI issues of how the knowledge in such systems is grounded are irrelevant to the creation of the experience of unreal worlds, which by definition are not so grounded. As authors, we find what has come to seem the naivete of early AI methods quite attractive, because it means that the resulting systems are relatively easy to understand and control. Finally, we find it interesting that in our system the reader will play a game whose issue will provide a soul for the machine that is our text.

- 
1. <http://www.turbulence.org/Works/twotxt/index.htm>

## Bibliography

Bernstein, Mark. "Card Shark and Thespis: Exotic tools for hypertext narrative." *Proceedings of the twelfth ACM*

conference on Hypertext and Hypermedia, Århus, Denmark.  
New York, 2001. 41-50.

Cayley, John. *From Byte to Inscription: An Interview with John Cayley*. Interviewed by Brian Kim Stefans. The Iowa Review Web, February, 2003. Accessed 2005-03-21. <<http://www.uiowa.edu/~iareview/tirweb/feature/cayley/>>

Eskelinen, Markku. "Six Problems in Search of a Solution: The challenge of cybertext theory and ludology to literary theory." *Dichtung Digital* (March 2004). <<http://www.dichtung-digital.com/2004/3/Eskelinen/index.htm>>

McGann, Jerome. *Radiant Textuality: Literature after the World Wide Web*. New York: Palgrave Macmillan, 2001.

Meehan, J.R. *The Metanovel: Writing Stories by Computer*. Yale Computer Science Research Report 74. New Haven: Yale, 1976.

Moulthrop, Stuart. *Interview with Stuart Moulthrop*. Interviewed by Noah Wardrip-Fruin. Accessed 2005-03-21. <<http://www.uiowa.edu/~iareview/tirweb/feature/moulthrop/>> The Iowa Review Web

Wardrip-Fruin, Noah. "From Instrumental Texts to Textual Instruments." *Proceedings of Digital Arts and Culture*. Melbourne, Australia, May 2003.

---

## User Generated Metadata: Creating Personalized Web Experiences

---

**Michael Fegan** ([mfegan@msu.edu](mailto:mfegan@msu.edu))

*Matrix, Michigan State University*

**Bill Hart-Davidson** ([hartdav2@msu.edu](mailto:hartdav2@msu.edu))

*WIDE Center, Michigan State University*

**Joy Palmer** ([palmerjo@msu.edu](mailto:palmerjo@msu.edu))

*Matrix, Michigan State University*

**Dean Rehberger** ([rehberge@msu.edu](mailto:rehberge@msu.edu))

*Matrix, Michigan State University*

---

### Abstract

This session will focus on the importance of measuring how communities of users interact with digital objects. By drawing on user-performance data and metadata generated for secondary repositories, we explore ways to enhance use and access of documents and digital libraries..

Speaker 1 proposes an approach to representing the structure of a document based on the way readers or users interact with it in the context of a deliberative task. This approach contrasts with other ways to model the structure of documents including approaches which map authorial intention and those which rely upon a well-known information model or genre. This presentation will highlight the benefits of understanding the structure of documents based on the rhetorical reading/using strategies of those who interact with them. These structures can be rendered as 'paths' through a given set of information resources, offering insight into the way that objects and relationships that make up a document can mediate, or complicate, activity. Speaker 1 will conclude by showing some examples which make use of user-performance data to create task-appropriate views of complex (multiscale) documents.

Speakers 2 and 3 will examine the role of secondary repositories can play in enhancing access and interaction for students and scholars in the humanities. The most entrenched *a priori* models for information structuring and delivery online are derived from library and archival cataloguing practices. In line with digital library best practices, digitized sources are typically cataloged to describe their bibliographic information, along with technical,

administrative, and rights metadata. While these practices are essential for preserving the digital object and making it available to users, unfortunately they do so in a language and guise often difficult to understand within the context of use. In addition, materials in digital libraries do not literally 'speak' for themselves and impart wisdom; they require interpretation and analysis within a context of use. Access and use of digital objects can no longer be thought of in terms of stand alone files or individual digital objects, but rather must directly impact the ways in which users reuse, repurpose, combine and build complex digital objects. This assumption relies on a more complex meaning for the term *access* that will be detailed and explained in this paper.

Following the examples in the first paper, speaker 4 will demonstrate an application that can be used to collect user generated metadata. Following the concepts developed in the second paper, speaker 4 will develop the argument in practice that one way we can enhance access to online digital objects is to facilitate the creation of secondary repositories. These repositories will provide discipline/community specific metadata and applications and will allow users to find, use, manipulate and analyze digital objects more easily. To this end, Speaker 4 has developed *Media Matrix 1.0* — an online, server-side suite of tools that allows users to locate specific media and streaming media files found in digital repositories and segment, annotate and organize this media online. This application provides users with an environment both to work with and personalize digital media, and also to share and discuss their findings with a community of users. This paper will explore if the creation of secondary repositories of usage statistics and user-generated materials/metadata (to supplement both traditional cataloging records and discipline-specific online indexes) can help scholars and students in the humanities gain better access to online materials.

### **Modeling Documents Based on User Performance: An Alternative to Author Intention and a priori Information Model Approaches**

**Bill Hart-Davidson, Ph.D.**

This paper proposes an approach to representing the structure of a document based on the way readers or users interact with it in the context of a deliberative task. This approach contrasts with other ways to model the structure of documents including approaches which map authorial intention and those which rely upon a well-known information model or genre. This presentation will highlight the benefits of understanding the structure of documents based on the rhetorical reading/using strategies of those who interact with them. These structures can be rendered as 'paths' through a given set of information resources, offering insight into the way that objects and relationships that make up a document can mediate, or

complicate, activity. The paper will conclude by showing some examples which make use of user-performance data to create task-appropriate views of complex (multiscale) documents.

There is a great deal of interest today in the idea of building models of texts. One reason is that, with the growth of the Web as a way to reach a wide and diverse audience, publishers of information of many types are now interested in building information structures that support multiple-audience adaptation. Another reason is to maximize the value of content by delivering information that is tailored to a particular task.

For example, imagine the day-to-day work of a claims processing agent for a large insurance company. The agent is responsible for making decisions based on information in documents of various types - claim forms, telephone call records, police and adjustor reports, medical records, even photographs - all stored in electronic policyholder files. There may be discernable patterns in these types of workflows which can be documented and used as the basis for information models that structure information at or below the level of an individual document. The model could allow information contained in all the documents associated with the policyholder file to transform to suit the decision-making needs of the specific users who interact with it.

Creating a system like the one described above would require analyzing and modeling fundamental patterns of document use, defining a basic modeling language for document-mediated interaction that can capture recurrent patterns of user performance. As Moser & Moore (1996) point out, most semantic modeling approaches construct formal text structures based on either author intention or, alternatively, on an information structure presumed to be instantiated in the text. Neither of these approaches is entirely appropriate for creating effective displays of information for potential users. Creating such displays requires a model of the text-mediated interaction between writers and readers which can then be used to define display conditions for a range of information "views" that a given document might support.

Performance-based text structure models differ from other types in that they are not primarily representations of a stable 'core' semantic structure that is assumed to be either domain independent (e.g. Mann and Thompson, 1987), or genre specific, as suggested by the work of Bazerman(1988) and others. Nor are these structures maps of author intentionality and/or struggles in creating intentional relationships similar to analyses by Van Wijk and Sanders (1999). Rather, the models emphasize structures that constitute the resources authors and the specific users or readers of a document share in order to come to some kind of agreement about an issue or question that all parties have a stake in. What gains status as a 'unit' or 'object' in user-performance based models depends upon the deliberative activity that the document is meant to support.

Relationships among objects are similarly defined by how the objects mediate a given decision. In this way, we can expect the model to account for both the regularities in text structures which correspond with similar texts doing similar mediational work, as well as quite specific and arbitrary text structures associated with any given deliberative activity as it unfolds in a social context. This modeling approach comes very close to a process described by Phelps (1985) who articulated an approach to structural analysis drawing on and responding to work by Faigley & Witte (1981) and Van de Kopple (1985) in composition studies, as well as Halliday & Hassan (1976) and Van Dijk (1976) in linguistics. The process, broadly, understands texts as objects with histories, requiring us to study them 'in process' if we are to understand how they shape the experiences of a reader.

## Bibliography

Bazerman, C. *Shaping written knowledge*. Madison: University of Wisconsin Press, 1988.

Faigley, L., and L.S. Witte. "Coherence, cohesion, and writing quality." *College Composition and Communication* 32 (1981): 189-204.

Halliday, M.A.K., and J.R. Hassan. *Cohesion in English*. London: Longman, 1976.

Mann, W.C., and S.A. Thompson. "Rhetorical structure theory: Toward a functional theory of text organization.." *Text* 8.3 (1988): 243-281.

Moser, M., and J.D. Moore. "Toward a synthesis of two accounts of discourse structure." *Computational Linguistics* 22.3 (1996): 409-419.

Phelps, L.W. "Dialectics of coherence: Toward an integrative theory." *College English* 47.1 (1985): 12-31.

Van Dijk, T. *Text and context: Explorations in semantics and pragmatics of discourse*. London: Longman, 1976.

Van Wijk, C., and T. Sanders. "Identify writing strategies through text analysis." *Written Communication* 16.1 (1999): 51-75.

Vande Kopple, W. "Given and new information and some aspects of the structures, semantics, and pragmatics of written texts." *Studying writing: Linguistic perspectives*. Ed. C.R. Cooper and S. Greenbaum. Beverly Hills, CA: Sage, 1986. 72-111.

## Enhancing Access to Online Digital Objects through Reciprocity between Primary and Secondary Repositories

Dean Rehberger and Joy Palmer

This paper will examine the role of secondary repositories can play in enhancing access and interaction for students and scholars in the humanities. While access to online resources has steadily improved in the last decade, online archives and digital libraries still remain difficult to use, particularly for students and novice users (Arms). In some cases, a good deal of resources have been put into massive digitization initiatives that have opened rich archives of sources to a wide range of users. Yet, the traditional cataloging and dissemination practices of libraries and archives make it difficult for these users to locate and use effectively these sources, especially within scholarly and educational contexts of the humanities. Many digital libraries around the country, large and small, have made admirable efforts toward creating user portals and galleries to enhance the usability of their holdings, but these results are often expensive and labor intensive, often speaking only directly to a small segment of users.

To address these problems, we begin with the assumption that access and preservation are mutually dependent concepts. Preservation and access can no longer be thought of in terms of stand alone files or individual digital objects, but rather must directly impact the ways in which users reuse, repurpose, combine and build complex digital objects. This assumption relies on a more complex meaning for the term *access*. Many scholars in the field have called for a definition of access that goes beyond search interfaces to the ability of users to retrieve information "in some form in which it can be read, viewed, or otherwise employed constructively" ((Borgman 57)). Access thus implies four related conditions that go beyond the ability to link to a network:

1. equity the ability of 'every citizen' and not simply technical specialists to use the resources;
2. usability the ability of users to easily locate, retrieve, use, and navigate resources;
3. context the conveyance of meaning from stored information to users, so that it makes sense to them;
4. interactivity the capacity for users to be both consumers and producers of information.

The keys to enhancing access for specific user groups, contexts, and disciplines are to build secondary repositories with resources and tools that allow users to enhance and augment materials (Shabajee), share their work with a community of users (Waller), and easily manipulate the media with simple and intuitive tools (or at least build interfaces that match existing, well-known applications). Users will also need portal

spaces that escape the genre of links indexes and become flexible work environments that allow users to become interactive producers (Miller).

Herbert Van de Sompel has proposed a successful system (OpenURL/SFX framework for context sensitive reference linking) for disaggregating reference linking services from e-publishing. In his framework, the service of providing links between references and across e-publisher's digital repositories is separated from the services provided by the e-publishers. In so doing, the service provides "seamless interconnectivity between ever-increasing collections of heterogeneous resources", freeing primary repositories from the difficult and expensive task of ensuring links to references while giving users greater access to resources and increasing the value of the digital object (Van de Sompel). Similarly, we propose the concept of secondary repositories that would be responsible for handling secondary metadata, extended materials and resources, interactive tools and application services. This information is cataloged, stored, and maintained in a repository outside of the primary repository that holds the digital object. The comments and observations generated by users in this context are usually highly specialized because such metadata is created from discipline-specific, scholarly perspectives (as an historian, social scientist, teacher, student, enthusiast, etc.) and for a specific purpose (research, publishing, teaching, etc.). Even though the information generated by a secondary repository directly relates to digital objects in primary repositories, secondary repositories remain distinctly separate from the traditional repository. The information gathered in secondary repositories would rarely be used in the primary cataloging and maintenance of the object, and primary repositories would continue to be responsible for preservation, management, and long-term access but would be freed from creating time-consuming and expensive materials, resources, services, and extended metadata for particular user groups.

In line with digital library best practices, digitized sources are typically cataloged to describe their bibliographic information, along with technical, administrative, and rights metadata. While these practices are essential for preserving the digital object and making it available to users, unfortunately they do so in a language and guise often difficult to understand within the context of use (Lynch 2003). Even though the author's name, the title of the work, and keywords are essential for describing and locating a digital object, this kind of information is not always the most utilized information for ascertaining the relevance of a digital object. For instance, K-12 teachers often do not have specific authors or titles in mind when searching for materials for their classes. Teachers more frequently search in terms of grade level, the state and national standards that form the basis of their teaching, or broad overarching topics derived from the required content and benchmark standards (e.g., core democratic values or textbook topics) that tend to

display too many search returns to make the information of value.

While cursory studies have indicated these access issues, still very little is known about archival use or how these users express their information needs (Duff, Duff & Johnson). For digital libraries to begin to fulfill their potential, much research is needed to better understand the processes by which primary repositories are accessed and how information needs are expressed. For example, research needs to address the ways in which teachers integrate content into their pedagogy so that bridges can be built from digital repositories to the educational process, bridges that greatly facilitate the ability of teachers and students to access specific information within the pedagogical process. Recent research strongly suggests that students need conceptual knowledge of information spaces that allow them to create mental models to do strategic and successful searches. As with any primary source, the materials in digital libraries do not literally 'speak' for themselves and impart wisdom; they require interpretation and analysis (Bowker & Star; Duff; Duff & Johnson). Allowing communities of users to enhance metadata and actively use, reuse, repurpose, combine and build complex digital objects can help users to contextualize the information they find, draw from deeper resources within the digital library, and find more meaningful relationships between digital objects and their needs. Thinking in terms of a distributed model (similar to the open source software community) that allows users both easier access to materials and a greater range of search criteria and also provides opportunity for active engagement in the generation of metadata and complex digital objects, promises to help us rethink our most basic assumptions about user access and long-term preservation.

Collections can also benefit by defining communities of users. For example, with the recent release of secret White House tapes (<http://millercenter.virginia.edu/>), the sheer number of tapes and hours make it impossible for adequate cataloging of content as well as the difficulty of determining the context and people involved (or even what is said given the poor quality of many tapes). Those historians and scholars (a more regulated and highly defined set of experts) allowed access to the collections could supply information about content and context as well as set terms for debates over more questionable areas of interpretation (e.g., when sound quality makes passages inaudible). While metadata gathered in these ways would need to be qualified (maintained in a secondary repository) because of lack of quality control, the processes could make large quantities of data that is key to many disciplines in the humanities more available and usable.

## Bibliography

Arms, William. *Digital Libraries*. Cambridge, MA: MIT Press, 2001.

Borgman, C. *From Gutenberg to the Global Information Infrastructure: Access to Information in the Networked World*. Cambridge, MA: MIT Press, 2000.

Bowker, Geoffrey C., and Susan Leigh Star. *From Gutenberg to the Global Information Infrastructure: Access to Information in the Networked World*. Cambridge, MA: MIT Press, 1999.

Cole, Charles. "Name Collection by Ph.D History Students: Inducing Expertise." *Journal of the American Society for Information Science* 51.5 (2000): 444-455.

Cooperstock, J.R. "Classroom of the Future: Enhancing Education through Augmented Reality." *Proceedings of the HCI International, Conference on Human-Computer Interaction, New Orleans*. 2001. Accessed 2005-04-15. <<http://www.cim.mcgill.ca/~jer/pub/hcii01.pdf>>

Duff, Wendy. "Evaluating Metadata at a Metalevel." *Archival Science* 1 (2001): 285-294.

Duff, Wendy, and Catherine A. Johnson. "A Virtual Expression of Need." *American Archivist* 64 (2001): 43-60.

Hedstrom, Margaret. "Research Challenges in Digital Archiving and Preservation." NSF Post Digital Libraries Futures Workshop. 15-17 June 2003.

Kornbluh, Mark Lawrence, Dean Rehberger, and Michael Fegan. "Media MATRIX: Creating Secondary Repositories in Research and Technology for Digital Libraries." *Proceedings of the 8th European Conference, ECDL2004*. Berlin: Springer, 2004. 329-340.

Lynch, Clifford. "Reflections toward the Development of a 'Post D-L Research Agenda." NSF Post Digital Libraries Futures Workshop. 15-17 June 2003.

Lynch, Clifford. "Interoperability: the standards challenge for the 90s." *Wilson Library Bulletin* March (1995): 38-42.

Lynch, Clifford. "Colliding with the Real World: Heresies and Unexplored Questions about Audience, Economics, and Control of Digital Libraries." *Digital Library Use: Social Practice in Design and Evaluation*. Ed. Ann Bishop, Barbara Butterfield and Nancy Van House. Cambridge, MA: MIT Press, 2001. 191-218.

Marshall, Catherine. "Annotation: from Paper Books to the Digital Library." *Proceedings of the ACM Digital Libraries '97 Conference, Philadelphia, PA*. July 23-26, 1997. Accessed 2005-04-15. <<http://www.csdl.tamu.edu/~marshall/dl97.pdf>>

Miller, Paul. "The Concept of the Portal." *Ariadne* 30 (20 December 2001). Accessed 2005-04-15. <<http://www.ariadne.ac.uk/issue30/portal/intro.html>>

Page, K.R., D. Cruickshank, and D.D. Roure. "Its About Time: Link Streams as Continuous Metadata." *Proceedings of the Twelfth ACM Conference on Hypertext and Hypermedia (Hypertext '01)*. 2001. 93-102.

Rydberg-Cox, Jeffrey. "Cultural Heritage Language Technologies: Building an Infrastructure for Collaborative Digital Libraries in the Humanities." *Ariadne* 34 (14 January 2003). Accessed 2005-04-15. <<http://www.ariadne.ac.uk/issue34/rydberg-cox/intro.html>>

Shabajee, Paul. "Primary Multimedia Objects and 'Educational Metadata': a Fundamental Dilemma for Developers of Multimedia Archives." *D-Lib Magazine* (March 2000). Accessed 2005-04-15. <<http://www.dlib.org/dlib/june02/shabajee/06shabajee.html>>

Van de Sompel, Herbert, and O. Oren Beit-Arie. "Open linking in the scholarly Information Environment Using the Open URL Framework." *D-Lib Magazine* (March 2001). Accessed 2005-04-15. <<http://www.dlib.org/dlib/march01/vandesompel/03vandesompel.html>>

Waller, Richard. "Functionality of Digital Annotation: Imitating and Supporting Real-World Annotation." *Ariadne* 35 (30 April 2003). Accessed 2005-04-15. <<http://www.ariadne.ac.uk/issue35/waller/>>

## Developing MediaMatrix: A Secondary Repository Tool

### Michael Fegan

Speaker Two will argue that one way we can enhance access to online digital objects (particularly in the humanities) is to facilitate the creation of secondary repositories. These repositories will provide discipline/community specific metadata and applications and will allow users to find, use, manipulate and analyze digital objects more easily.

Even though access by specialist scholars and educators to digital objects has grown at an exponential rate, tangible factors have prevented them from fully taking advantage of these resources in the classroom, where they could provide the conceptual and contextual knowledge of primary objects for their students. When educators do find the materials they need, using objects from various primary repositories to put together presentations and resources for their students and research can be challenging. Beyond merely creating lists of links to primary and secondary resources, assembling galleries of images, segmenting and annotating long audio and video files require far more technical expertise and time than can realistically be expected in the educational context. In addition, even though

scholars have a long history of researching archives and are comfortable sifting through records, locating items, and making annotations, comparisons, summaries, and quotations, these processes do not yet translate into online tools. Contemporary bibliographic tools have expanded to allow these users to catalogue and keep notes about media, but they do not allow users to mark specific passages and moments in multimedia, segment it, and return to specific places at a later time. Multimedia and digital repository collections thus remain underutilized in education and research because the tools to manipulate the various formats often 'frustrate would be users' and take too much cognitive effort and time to learn.

To this end, Speaker Two has developed *Media Matrix* 1.0 — an online, server-side suite of tools that allows users to locate specific media and streaming media files found in digital repositories and segment, annotate and organize this media online. The application has been developed as part of the *Spoken Word Project* funded by Digital Libraries Initiative II: Digital Libraries in the Classroom Program, National Science Foundation in conjunction with UK's Joint Information Systems Committee.

This application is an online tool that allows users to easily find, segment, annotate and organize text, image, and streaming media found in traditional online repositories. *MediaMatrix* works within a web browser, using the browser's bookmark feature, a familiar tool for most users. When users find a digital object at a digital library or repository, they simply click the *MediaMatrix* bookmark and it searches through the page, finds the appropriate digital media, and loads it into an editor. Once this object is loaded, portions of the media can be isolated for closer and more detailed work — portions of an audio or video clip may be edited into a time-segment, images may be cropped then enlarged to highlight specific details. *MediaMatrix* provides tools so that these media can be placed in juxtaposition, for instance, two related images, a segment of audio alongside related images and audio, and so forth. This can be particularly effective for students and researchers who need to fit images into a presentation or would like to demonstrate specific nuances and details about portions of images or artwork. Most importantly, textual annotations can be easily added to the media, and all this information is then submitted and stored on a personal portal page.

A portal page might be created by a scholar-educator who wishes to provide specific and contextualized resources for classroom use, and/or by a student creating a multimedia-rich essay for a class assignment. While these users have the immediate sense that they are working directly with primary objects, it is important to emphasize that primary repository objects are not actually being downloaded and manipulated. *MediaMatrix* does not store the digital object, rather, it stores a pointer to the digital object (URI) along with time or

dimension offsets the user specified for the particular object and the user's annotation for that particular object. This use of URI pointing as opposed to downloading is especially significant because it removes the possibility that items may be edited and critiqued in contexts divorced from their original repositories, which hold the primary and crucial metadata for such objects.

As long as primary repositories maintain persistent URIs for their holdings the pointer to the original digital object will always remain within the secondary repository, which acts as a portal to both the primary collection and contextualizing and interpretive information generated by individuals on items in those collections. This information can be stored in a relational database along with valuable information about the individual, who supplies a profile regarding their scholarly/educational background, and provides information of the specific purposes for this work and the user-group (a class, for example) accessing the materials. *Media Matrix* is a PHP based server side application that stores information in a *mySQL* database and exports that information into XML for display. The development of the tool and programming environment have been designed to keep it library and archive independent so that it can work with almost any site on the internet. It can also work easily with any of the standard courseware packages. The tool is also search independent because it relies on traditional internet search tools and a site's discovery tools to find an object. Once objects are found, *Media Matrix* is deployed by the user. Because *Media Matrix* does not actually copy the digital object from the site (it only stores a pointer to the object in the form of a URI and whatever time offsets are created by the user), it avoids some of the copyright and fair use pitfalls that often keep users from working with digital objects (although there are issues of deep linking to be addressed). The secondary repository can thus be searched and utilized in any number of ways.

Historians, for example, can browse the portals of other historians working specifically in their research areas or K-12 teachers can browse grade appropriate sections defined by specific grade levels and subjects to see what digital objects other teachers are using or, more important, for time challenged teachers, they can find specific presentations created around standard topics and curriculum frameworks. Users can also perform keyword searches over the annotations created by all users or specific groups of users. A teacher, for instance, can choose to search through only the information in eleventh-grade Civics groups in hopes of finding information that speaks directly to his/her needs. Because users have gathered content from across the Internet and from a variety of digital repositories, searching *Media Matrix* is equivalent to searching multiple repositories at once. Once users find an object from a particular digital library, they can jump to that repository to find what other objects are available.

Going beyond demonstration, this paper will also dive the latest findings and evaluations based on initial user testing in several classrooms as Tufts University and Michigan State University.

---

## Advanced Topics in TEI

---

**Julia Flanders** (*Julia\_Flanders@brown.edu*)

*Brown University*

**Syd Bauman** (*Syd\_Bauman@brown.edu*)

*Brown University*

**Laurent Romary** (*Laurent.Romary@loria.fr*)

*INRIA Laboratoire Loria*

**David J. Birnbaum** (*djbpitt+@pitt.edu*)

*University of Pittsburgh*

**Matthew Zimmerman**

(*Matthew.Zimmerman@nyu.edu*)

*New York University*

---

**I**n the decade since the 1994 publication of the *TEI Guidelines*, this important text encoding standard has seen widespread use in a variety of research and digitization environments. In some contexts, its application has become routine: digital libraries now publish huge volumes of lightly encoded *TEI* documents through mechanisms which are well understood and thoroughly documented. However, in other quarters intensive research on the *TEI* continues unabated. Not only are the *Guidelines* themselves now being revised (with the publication of *P5* planned for 2005), but applications of the *TEI* to specific research areas continue to emerge, and new tools are continually being developed to support a variety of analytic and publication functions.

This panel session brings together several short presentations on advanced topics in the *TEI* research landscape, which reflect the breadth and depth of work currently being done in this community. The presentations include advanced markup issues, the design of the language in which the *TEI* itself is written and documented, and current *TEI* tools development. The panel chair will open the panel by giving a very brief contextual description of the current development context for the *TEI*: the goals of *P5*, the user community, and current trends in analytical use of *TEI* markup. Following the four short papers by panelists (described below), the chair and panelists will lead a discussion of advanced use of the *TEI* and future research directions. The goal of the panel is twofold: first, to provide an update to the humanities computing community on some important research efforts within the *TEI*; and second, to provide an opportunity for a discussion of the impact and value of this research and its direction for the future.



The first paper will discuss the perennial problem of overlapping markup, and will describe a *TEI* implementation of the *CLIX* solution, which has emerged from the work of the *TEI Special Interest Group (SIG)* on *Overlap*. The *CLIX* approach involves using two empty elements to indicate where each element in a subordinate hierarchy (or at least, each element which overlaps an element in another hierarchy) begins and ends. These empty elements have the same name as would have been used for the equivalent 'normal' element which has content, and use special attributes, `sID=` & `eID=`, to indicate that an empty element indicates the beginning or the end of a pseudo-element (see <http://www.mulberrytech.com/Extreme/Proceedings/html/2004/DeRose01EML2004DeRose01.html#t6> ). *RelaxNG*, the schema language underlying *TEI P5*, is perfectly capable of representing some of the constraints that would be desired to validate this type of markup. However, *ODD*, the abstract literate encoding language in which *TEI P5* is written, cannot. A mechanism for permitting *TEI P5 RelaxNG* schemas to perform some *CLIX* validation without changing the *ODD* language itself, but rather by using a slightly more complex 'tangle' process to produce schemas from the *ODD* sources, will be presented.

The second paper will discuss analytical approaches to manuscript description and the use of this markup to support advanced research in quantitative codicology. Data-centric manuscript description has recently emerged as a topic of interest in light of the new opportunities provided by electronic text technology. While traditional printed manuscript descriptions have been substantially prose-like (a tendency reflected in more document-centric encoding approaches), the more analytical approach presented here (which will be adopted as part of the new *TEI* chapter on manuscript description) treats manuscript description as structured databases rendered in XML. Highly structured descriptions with rich markup of all descriptive details (using controlled vocabularies wherever possible) permit users to conduct much more advanced research, for instance on the correlation between specific watermarks and specific orthographic norms, or on the resemblance between manuscripts in a given set of features. These kinds of questions go well beyond the tradition of consulting indices or searching for access points, and enable scholars to envision manuscript transmission in ways that would otherwise be impossible. This presentation will illustrate both the provisions of the *TEI MS* description module and its application to these advanced research topics.

The third paper will focus on designing and extending document models with the *TEI*. It will present the main characteristics of the new *TEI* specification platform, which is being used to describe both the documentation and technical characteristics of the next edition of the *TEI guidelines (P5)*. The specification platform (also known as *ODD* for "One Document Does it all") allows one to describe elements and their attributes, through a

combination of prose and formal descriptions. It also allows document model designers to refer to classes of elements, when similarity of behaviour or semantics have to be taken into account. The presentation will illustrate the new *TEI* architecture by presenting the online environment (*Roma*) that allows anyone to design his or her own *TEI* subset and possibly extend the *TEI* capacities by adding or modifying elements and attributes. We will exemplify these mechanisms in the light of the new terminology chapter that is to appear in the *TEI P5* edition.

The final paper in this panel will present the current landscape of *TEI* tools development, and in particular the work of the *TEI Tools Special Interest Group (SIG)*. It will discuss the current challenges faced by developers of *TEI* tools, the genres of tools which are currently of greatest interest, the ways in which the *TEI* community can most effectively assist tool developers (for instance, by contributing to a library of sample documents for testing), and the support framework provided by the *SIG*.

## Hybrid Cyber-Librarians: The CLIR Post-Doctoral Fellowship in Scholarly Information Resources for Humanists

---

**Amanda French** ([amanda\\_french@ncsu.edu](mailto:amanda_french@ncsu.edu))

North Carolina State University

**John Unsworth** ([unsworth@uiuc.edu](mailto:unsworth@uiuc.edu))

University of Illinois

**Susan Nutter**

North Carolina State University Libraries

**Sarah Michalak**

North Carolina State University Libraries

**Patricia Hswe** ([phswe@uiuc.edu](mailto:phswe@uiuc.edu))

University of Illinois Urbana-Champaign

**Daphnée Rentfrow** ([daphnee.rentfrow@yale.edu](mailto:daphnee.rentfrow@yale.edu))

Yale University

---

### Panelists

**Amanda French**, CLIR Post-Doctoral Fellow, North Carolina State University (organizer)

**John Unsworth**, Dean and Professor, Graduate School of Library and Information Science, University of Illinois Urbana-Champaign (panelist)

**Susan Nutter**, Vice Provost and Director of Libraries, North Carolina State University (panelist)

**Sarah Michalak**, University of North Carolina Library (panelist)

**Patricia Hswe**, University of Illinois Urbana-Champaign (panelist)

**Daphnée Rentfrow**, CLIR Post-Doctoral Fellow, Yale University (panelist)

The Council on Library and Information Resources ( <http://www.clir.org/> ) has developed a unique new program of crucial interest to the membership of the Association for Computers and the Humanities: the *CLIR Post-Doctoral Fellowship in Scholarly Information Resources for Humanists*. The stated purpose of this program, now in its first year at ten U.S. academic libraries, is "to establish a new kind of scholarly

information professional" by offering individuals with a recent Ph.D. in a humanities field a chance to acquire the experience of the academic librarian in the digital age (CLIR , <http://www.clir.org/fellowships/postdoc/postdoc.html> , "Information"). ACH is one of the few professional associations in which scholars, librarians, and technology specialists come together as a coherent community; every one of its members is no doubt aware that communication in other venues between these three professions is highly problematic. Moreover, people who undertake humanities computing projects are frequently frustrated by the fact that their professions lack a structure that allows new-model collaborative work to be recognized and rewarded.

Successful humanities computing initiatives such as the *Institute for Advanced Technology in the Humanities* ( <http://www.iath.virginia.edu/> ) at the University of Virginia could not have been possible without the enthusiastic collaboration of scholars, librarians, and technology specialists. Those involved in these initiatives understand that the nature of the work often goes beyond the collaborative or even the interdisciplinary to become what we might call the 'interprofessional'. On humanities computing projects, it is rare to find individuals rigidly adhering to their own professional skills and duties as though on an assembly line. Usually, those who participate develop into information professionals who possess various proportions of expertise in scholarship, library science, and technology.

Such interprofessional work is not credentialed, however, nor is it generally practiced outside the humanities computing community. This creates serious problems not only for individuals whose work has been interprofessional but also for the whole enterprise of digital scholarship. Scholars who have been involved in humanities computing projects are far more likely than other scholars to understand the challenges facing academic librarians in the digital age — but all scholars are beginning to expect high-level digital services from their libraries. CLIR frames this problem thus:

Libraries digitize primary resources to respond to the demands of individual scholars, but scholars don't recognize the complexity of carrying out this task nor do they understand the demands placed on librarians who need to improve access and ensure preservation... Scholars are increasingly accepting of digital publication, while librarians are finding that the contract law that controls access to such publications makes preservation impossible and access problematic.

(CLIR , <http://www.clir.org/fellowships/postdoc/detail.html> , "Program")

But is the creation of a new hybrid professional the answer to problems such as these, or does a program such as the CLIR post-doctoral fellowship threaten to undermine the valuable discrete professional knowledge of librarians, scholars, and technology specialists? A recent *Library Journal* article on the

CLIR Post-Doctoral Fellowship takes the latter position, declaring that "It weakens our profession when we open it to Ph.D.'s without established library credentials" (<http://www.libraryjournal.com/article/CA474993>), Crowley) This perspective emphasizes the unique skills belonging to a particular profession, a perspective that might well be shared by some scholarly professionals and some technology professionals. Is the specialized knowledge of these professions endangered or diluted by the collaborations that are so common on humanities computing projects? Should interprofessional credentialization be encouraged?

We suspect that the membership of ACH has much to contribute to this discussion, and we believe that the first year of the CLIR Post-Doctoral Fellowship is the ideal time to discuss the professionalization issues faced by information professionals engaged in humanities computing projects. We therefore propose a ninety-minute session in which panelists will engage in debate about the CLIR Post-Doctoral Fellowship and related professionalization and specialization issues. We invite the members of ACH to engage in this discussion of the principles behind—and future of—this attempt to credential a new species of digital scholar-librarian.

## Bibliography

Council on Library and Information Resources. *Post-Doctoral Fellowship in Scholarly Information Resources for Humanists: Information for Applicants*. Accessed 2005-03-23. <http://www.clir.org/fellowships/postdoc/postdoc.html>

Council on Library and Information Resources. *Post-Doctoral Fellowship in Scholarly Information Resources for Humanists: Program Background*. Accessed 2005-03-23. <http://www.clir.org/fellowships/postdoc/detail.html>

Crowley, Bill. "Just Another Field?" *Library Journal* 129.18 (2004): 44-6. Accessed 2005-03-23. <http://www.libraryjournal.com/article/CA474993>

Oder, Norman. "New Movement for Ph.D.'s To Work in Academic Libraries." *Library Journal* 128.11 (2003): 16-17. Accessed 2005-03-23. <http://www.libraryjournal.com/article/CA302413>

## The Canadian Century Research Infrastructure Project and Computing in the Humanities

**Chad Gaffield** ([gaffield@uottawa.ca](mailto:gaffield@uottawa.ca))

*Department of History, University of Ottawa*

**Marc St-Hilaire** ([Marc-St-Hilaire@ggr.ulaval.ca](mailto:Marc-St-Hilaire@ggr.ulaval.ca))

*Centre interuniversitaire d'études québécoises, Université Laval*

**Claude Bellavance** ([Claude\\_Bellavance@uqtr.ca](mailto:Claude_Bellavance@uqtr.ca))

*Centre d'études québécoises (CÉDEQ)*

**Gordon Darroch** ([darroch@yorku.ca](mailto:darroch@yorku.ca))

*Sociology, York University*

**Peter Baskerville** ([pab@uvic.ca](mailto:pab@uvic.ca))

*History, Univ of Victoria*

## Introduction

One of the most comprehensive humanities and social science research projects in Canadian history, the *Canadian Century Research Infrastructure* (CCRI) is a five-year, pan-Canadian initiative to develop a set of interrelated databases centered on census records for the 1911-1951 period.

The databases being developed from manuscript census records for the period 1911 to 1951 form the core of a much larger research infrastructure, the objective of which is to construct an evidentiary foundation for research on the transformation of Canadian society from the late 19th century to the later 20th century. To construct this evidentiary foundation, the *CCRI* will have two major components: primary sources and secondary sources.

Census microdata from the 1911-1951 enumerations form the first and the core of the four primary data sources. Other primary data sources include Statistics Canada documentary sources concerning the enumeration process; newspaper evidence about the enumerations at the time of each enumeration; and House of Commons and Senate debates related to the enumerations. The goal of this component of the *CCRI* is to provide researchers with the contextual evidence necessary to undertake appropriate analysis of the census microdata.

The secondary data sources are intended to facilitate research on the primary sources, and are equally varied in nature. They range from introductory descriptive statements about the census enumeration process, to highly technical discussions of data-entry and coding issues, and bibliographies of census-research publications.

Integral to the entire project is the construction of a geographic framework for the historical census data, using a *Geographic Information System (GIS)*. *GIS* map layers are being created to enable geographic location, selection, aggregation and analysis of sample data, as well as some mapping of generalized census data. This will allow researchers to ask questions of the database which are much more geographically specific than in the past. Interface tools to make these geographic queries and analysis as user-friendly as possible are also being developed.

The *CCRI* will be structured in terms of five distinct articulations, each devoted to one of the enumeration years (1911-1951). The *CCRI* will include cross-census harmonization bridges (or crosswalks) that connect each of the five articulations, to enable comparative analysis. A variety of user guides will be developed to aid researchers. In addition to a general introduction to each census enumeration, there will be user guides for each census variable, as well as a separate guide detailing the coding scheme for that variable. As it is expected that some variables (such as occupation) may be coded according to more than one scheme, each scheme will be discussed in the guides.

The *CCRI* databases will be made available through Research Data Centres across Canada; versions will also be available through the *Data Liberation Initiative* at Canadian universities. Once completed, the *CCRI* databases will be joined to other databases that cover the periods from 1871 to 1901 and from 1961 to 2001. The result will be a new foundation for the study of social, economic, cultural, and political change, as the *Canadian Century Research Infrastructure* will include an extraordinary range of data about the twentieth century.

The proposed panel discussion will focus on the following related issues:

- Integrating Words and Numbers in Historical Databases;
- Mapping Time: Using *GIS* to enhance historical understanding;
- Meta-data, Contextual Data, and User Guides for Historical Evidence: How much is enough?
- Integrating Primary and Secondary Sources in Historical Research Infrastructures.

---

## **METS in Action: Standardization and Interoperability in the Digital Library**

---

**Richard Gartner** (*richard.gartner@sers.ox.ac.uk*)

*Oxford University Library Services*

**Rick Beaubian** (*rbeaubie@library.berkeley.edu*)

*University of California, Berkeley*

**Jerome McDonough** (*jerome@nyu.edu*)

*New York University*

**Susan Dahl** (*Susan.Dahl@ualberta.ca*)

*University of Alberta*

**Brian Tingle** (*brian.tingle@ucop.edu*)

*California Digital Library*

---

**T**he use of SGML/XML for describing collections and digital objects has been in place since the mid-1990s (for instance, in the UK's *Internet Library of Early Journals (ILEJ)* (<http://www.bodley.ox.ac.uk/ilej/>)). By the end of that decade, the EAD ("Encoded Archival Description") had been developed to give the archivist community an encoding standard for describing archival collections digitally and for helping scholars and researchers identify and locate relevant materials in these collections. However the EAD did not address the problem of how electronic versions of the individual items comprising an archival collection could themselves be described digitally in a standard way.

Such an encoding standard needed to provide a means for inventorying the individual content files (image files, structured text files, etc.) comprising an electronic version of an archival item, applying a structure or structures to these content files, and associating relevant descriptive and administrative metadata with both structure and content. Beginning in 1997, a Digital Library Federation initiative called *Making of America II* sought to address this need for a digital object encoding standard, and out of this initiative came the MOA2.DTD; from these origins, the current METS ("Metadata Encoding and Transmission Standard") has developed.

METS is now firmly established within the digital library community, although the number of projects employing it is

still relatively small. It is intended to act as a "MARC standard" for digital objects, by providing a standardized framework within which the metadata detailed above may be contained. This standardization will allow the degree of interoperability that has prevailed in the cataloguing world to become possible in the more complex environment of digital objects, and hopefully facilitate the pooling of digital resources in similar ways to the union catalogues that MARC has allowed.

This panel session will provide a short introduction to METS and its history, and show how it is currently implemented, demonstrating how it offers solutions of wide applicability to some difficulties presented by the older standards and techniques. The participants are all members of the METS Editorial Board, responsible for the maintenance of the standard. It is intended that each participant will give a short (10-15 minute presentation) and half an hour will be available for general discussion.

**Rick Beaubian**, Software Engineer, Digital Library Projects, University of California, Berkeley, will trace the evolution of METS from its origins in the *Making of America II* initiative to the present, and examine its progress from a relatively narrow standard primarily targeting archival and traditional library materials to one that is now also being applied to encoding audio and video content as well as to archiving websites. It will also look briefly at METS' position and application relative to other emerging content packaging standards: IMS-CP/SCORM and MPEG-21.

**Richard Gartner**, Pearson New Media Librarian at Oxford University Library Services, will discuss ways of integrating METS and TEI: METS is designed to allow easy integration with any XML-based documents, either by reference or by direct embedding within the METS file itself. This talk will demonstrate how TEI documents can be integrated within METS files and how the two in tandem can overcome some of the difficulties experienced when using the TEI on its own. In particular, he will cover how METS and TEI can handle images in complex objects more simply than in TEI alone, and how overlapping hierarchies can be handled neatly using multiple METS structural maps.

**Jerome McDonough**, Digital Library Team Leader at New York University, will discuss the application of METS to video, specifically a series of videos of performances gathered by the Hemispheric Institute. He will focus on some of the problem issues in trying to collect descriptive and structural metadata on a multi-institution project such as this, and how METS may help alleviate them.

**Susan Dahl**, Metadata and Cataloguing Librarian at the University of Alberta, will discuss how METS can be incorporated in a project that uses OCLC's Olive Software to digitize textual materials. Using the Peel's Prairie Provinces

Project at the University of Alberta as an example ( <http://peel.library.ualberta.ca/> ), the presentation will demonstrate how the XML supplied in Olive's format can be incorporated into METS documents and the benefits this offers. Also, it will feature how other descriptive, structural and administrative metadata is included, to make a complete METS document.

**Brian Tingle**, Content Management Designer at the California Digital Library, will discuss the Local History Digital Resources Project, which explores a model to support the creation of, and permanent public access to, standardized digital objects with associated collection guides through a single point of access. The project aims in particular to develop requirements for, and helping to select, a common digital project tool that all contributors will be required to use.

The METS home page is available at <http://www.loc.gov/standards/mets> .

## ***Clotel*: An Electronic Scholarly Edition**

---

**Matthew Gibson** ([mgibson@virginia.edu](mailto:mgibson@virginia.edu))  
University of Virginia

---

### **Background**

In the fall of 2001, Christopher Mulvey, Professor of English at the University of Winchester, came to the Electronic Text Center (Etext) at the University of Virginia Library to collaborate on creating a scholarly electronic edition of William Wells Brown's *Clotel*. In the spring of 2005, Adam Matthew Publishers (UK) and the University of Virginia Electronic Imprint (US) will jointly publish the product of this collaboration: *Clotel: An Electronic Scholarly Edition*. This collaborative opportunity proved intriguing for a number of cultural, technical, and theoretical reasons.

### **The Technical Challenges of History**

Of cultural and historical interest, *Clotel* was the first African-American novel ever published and its content proves particularly germane to the University of Virginia since the institution's "father", Thomas Jefferson, is the father of *Clotel*, the mulatto fugitive slave and heroine of Brown's novel. In addition, the novel's publication history spans an incredibly dynamic period of United States history (slavery, Civil War, emancipation, and Reconstruction) and the substantive changes among the four editions of the novel between 1853 and 1867 are certainly reflective of that political environment.

Technically, the *Clotel Electronic Edition* pushes the bar on past electronic scholarship that the Electronic Text Center has engaged specifically, as well as developments in the creation of electronic scholarly editions generally. The innovative visualizations that *Clotel* forced Etext to explore are due, in large part, to the milieu of cultural events in which the novel was published. The historical events, which directly affected William Wells Brown as he went from a fugitive slave to a free man, engendered a large number of substantive variations between the different editions of *Clotel*. Because the editions are so different from one another, because, in most cases, they are different artistic "works", the task of aggregating them into a single project for scholarly comparison, analysis, and discovery provided a sizeable challenge.

Using Early American Fiction's<sup>1</sup> first American edition of William Wells Brown's novel, *Clotel: A Tale of the Southern States* (1864), as a starting point, Mulvey outsourced the digitization of three more editions: the 1853 first edition published in London, the 1860 version serialized in the *Weekly Anglo-African*, and the 1867 second American edition. Mulvey's idea was to take each of the editions, mark up regions of contextual similarity, and then provide tools for the user to approach the electronic edition in a number of ways without necessarily privileging any one version. Utilizing the Text Encoding Initiative (TEI) as the markup standard for the project, Etext had to take Mulvey's ideal requirements and visually try to understand how that markup should function in the context of publication.

Emphasis upon visualizing the changes between the editions was particularly important for Mulvey. He wanted to compare the texts in different visual environments: one, for instance, that was heavy on exposing emendation and another that privileged the act of reading with optional tools for comparison. A dominant conceptual view that he maintained actually came from Microsoft. Having often used the "track" tool in Microsoft Word to compare version changes in his own documents, Mulvey wanted to develop a similar tool (or use a tool that already existed) to emulate that functionality so that users could see the changes that occur between each edition at any given time. Privileging that act of uninterrupted reading but giving the option for comparison, under Mulvey's direction Etext placed the four "witness" versions of *Clotel* in a parallel reading view in which a user could link to comparable passages in any of the other three editions at any point in any version of the text. The edition also gives users attendant "reading" copies of each text along with links to annotations.

### **Theoretical Implications**

Within the larger environment of scholarly publishing and the framework of debate over its future, the *Clotel* project brings to light several speculative issues that may offer an option to that environment and the role that libraries could have and are beginning to have in what has historically been unfamiliar territory. While the Electronic Text Center has always theoretically aligned itself as being an electronic "publisher", it has never had the ancillary apparatuses of the traditional publishing environment: it has no "publicity office" and, more important, no real mechanism for peer review. However, the stakeholders that the *Clotel* project has brought together—an ambitious scholar, two scholarly presses (Adam Matthew in the UK and the University of Virginia's Electronic Imprint) and a digital library unit with expertise in content creation—have made themselves into an ad-hoc publishing force. While the crisis in scholarly publishing continues, a collaboration of this sort which brings together these types of

players with typically varying interests and priorities is certainly a stab in a direction, albeit experimental and unknown.

Although the Electronic Text Center has been, for lack of a better word, "publishing" content for well over a decade using the web and other non-web electronic formats, its line-straddling between library and publishing culture has never been as apparent or important as it is now. One might posit that this position is not one of happenchance or exception, but one necessarily born by the realities, limits, and failures in the domain of print culture.

- 
1. Funded by two grants from the Andrew W. Mellon Foundation, *Early American Fiction (1789-1875)* is a recently completed collection of digitized early first American editions and primary source materials. The on-line collection includes 886 volumes from 136 different authors. There are 199 transcribed manuscript items (525 pages of drafts, letters, and miscellaneous items) and 124 non-text items (photos, engravings, etc.) included in the collection. See <http://etext.lib.virginia.edu/ea> . [Ed. note: *Clotel* is not among the online texts available for external access. PL]

---

## Words as Data? Estimating the Fiscal Conservativeness of Provincial Premiers Using the Wordscore Procedure of Content Analysis

---

*André S. Gosciniak*

*(Andre.Gosciniak@capp.ulaval.ca)*

*Laval University*

---

**W**hat is the usefulness of the new computer wordscore method developed by Laver, Benoit and Garry (2003) for research in political science, and particularly for estimating the policy position of key political actors? This is the central question I address in this paper. Using the computer wordscore methodology I analyze the speeches given in the legislative assemblies of Quebec and Ontario by provincial premiers to assess their budgetary policy position. I present the empirical results of the wordscore content analysis and discuss the validity and the effectiveness of this new technique.

Manual content analysis, but also traditional computer content analysis (either using a dictionary approach or another one) has involved large amounts of highly skilled labor required for developing and testing the coding dictionaries that are central to these approaches. In addition, important human involvement brings the risk of potentially biased human coders. These two shortcomings can be avoided by using a new probabilistic word-scoring method developed recently by Laver, Benoit and Garry (2002 & 2003) which extracts policy positions from political texts by treating them as data in the form of words.

The technique using the software *Wordscore* estimates policy positions by comparing a set of texts whose positions are known or can be either estimated with confidence from independent sources or assumed uncontroversially as representative for a policy position (reference texts) and a set of texts whose policy position one must search to uncover (virgin texts).

Using the computer wordscore methodology I analyze the policy position of provincial premiers in the budgeting process in Quebec and in Ontario in the years 1968-2004 using the 'guardian-spender' budgetary framework (Wildavsky 1964 & 1984), which supposes the existence of two main roles attached to the institutional positions held by the participants in the budgetary negotiations: they are either 'guardians' of the treasury

(participants from central agencies controlling the budget: Minister of Finance, President of the Treasury board) or advocates of program spending coming from program ministries (minister of Health, minister of Education, etc).

Since we can not predict the fiscal behaviour of the government on the base of a simple partisan left-right dichotomy (Imbeau), the traditional models of budgetary theories employing politicians' [here called 'actors'] preferences do not give us any appropriate measure of the budgetary position of key political 'actors'. Thus, a way to uncover policy position of a premier is to assess his or her preferences on an additional dimension, differentiating between budgetary actors with a total vision of budget (guardians) and those with a partial vision of it (spenders). The vision is total if the budget balance is more important than a party's taxing or spending preferences. On the other hand, the vision of the budget is partial if the party program (more or less tax and spending) is more important than the budget balance.

Which texts are representative of the policy position of budgetary 'actors' we are interested in? We certainly can assume that the Speech from the Throne delivered by the Lieutenant-Governor at the opening of each legislative session expresses the Premier's policy preferences. Its delivery constitutes one of the central moments in a session, thus it is carefully crafted by the Premier's Office and its content is reviewed many times before delivery. The Throne speeches are compared to two reference texts: the budget speech representing the guardians' expression of a total view of the budget, and the preliminary remarks by the Ministers of Health and of Education at budget hearings representing the spenders' view of a partial vision of the budget. Although all three texts are usually drafted so as to represent the position of the government, they should be differentiated because of the policy role that each assumes.

Based on these assumptions and using the wordscore technique of content analysis, I develop a fiscal conservatism index of budgetary 'actors'. I then assess the internal validity of my results using a variety of validity tests.

## Bibliography

Imbeau, L.M. "Deficits and Surpluses in Federated States: A Review of the Public Choice Empirical Literature." Paper presented at the annual conference of the Canadian Political Science Association, Winnipeg, 2004.

Laver, M., K. Benoit, and J. Garry. *Placing Political Parties in Policy Spaces*. Dublin: Trinity College, 2002.

Laver, M., K. Benoit, and J. Garry. "Extracting Policy Positions from Political Texts Using Words as Data." *American Political Science Review* 97 (2003): 311-331.

Wildavsky, A.B. *The Politics of the Budgetary Process*. Boston: Little, Brown and Company, 1964.

Wildavsky, A.B. *The Politics of the Budgetary Process (4th Edition)*. Boston: Little, Brown and Company, 1984.



---

# An Examination of the Authorship Attributions of Two Major Roman Authors

---

*Lyman W. Gurney (lwgurney@telus.net)*

*Themis Research (ret.)*

*Penelope J. Gurney (pgurney@uottawa.ca)*

*University of Ottawa (ret.)*

---

## Introduction

This paper describes a stylometric analysis of 16 texts attributed by the manuscript tradition to two Roman authors of the first century BCE: Gaius Julius Caesar (100-44; all dates BCE, unless noted), and Gaius Sallustius Crispus, commonly referred to as Sallust (86-34). The 'control' consists of the *Lives of the Twelve Caesars* of Gaius Suetonius Tranquillus (77-121 CE at the earliest). All twelve works are attributed by the tradition to Suetonius, and our analysis, given in earlier publications, has corroborated this tradition of single authorship.

The texts analysed include the fourteen books attributed to Caesar, and two books by Sallust. The texts of Caesar comprise: the eight books on the war in which he conquered Gaul in the years from 58 to 50; the three books on the following Civil Wars (49-48), in which Caesar eliminated his great political rival Pompey and his Eastern army; and the three books on his final battles against the surviving generals: in Alexandria (48-47), Africa (47-46), and Spain (46-45). This study then completes a similar analysis of the works of Sallust: the war in Africa against the Numidian king Jugurtha (111-106); and the failed rebellion of the Roman noble Lucius Sergius Catilina, known as Catiline, in 63.

Of the eight books of the Gallic Wars, the question arises as to exactly when and by whom they were created. Some have argued that the first seven were written as a group, but that the eighth must be attributed to his general Hirtius, who claims in the work that he was responsible for both that and the later text on the war in Alexandria.

Of the further works, the three on the Civil Wars, from his crossing of the Rubicon in 49, to the battle of Pharsalus in Thessaly in 48, are generally accepted to have been authored by Caesar. There is considerable disagreement, however,

concerning the three wars in Egypt, Africa and Spain: the Alexandrian War, as noted, is claimed by Hirtius; the origin of the African War is uncertain; and the internal character of the text of the Spanish War clearly defines it to be the creation of an unknown person. There is little disagreement on this last score, since the text on the Spanish War comprises some of the worst Latin extant, and was probably intended to be the raw material for a more structured history.

## Statistical Routines

The Stylometric Analysis of the 28 texts has been conducted by use of the SPSS routines Hierarchical Cluster Analysis, and Principal Component Analysis. Discriminant Analysis has then been used to test the group memberships suggested by the first two.

## Data

The data for the statistical routines have been provided by a matrix of 9,000 by 58 real-valued elements that represent the normalized frequencies of occurrence of unique lemmas (dictionary head-words) in 58 texts. This matrix has been generated from the fully disambiguated texts of the 58 works of Caesar, Sallust, Suetonius, and the *Scriptores Historiae Augustae*, and involves a reduction from 329,000 to 305,000 numeric values to be handled after the removal of all proper nouns. This matrix has been sorted in decreasing order on the frequencies of lemmas, and lists also the number of texts in which each individual lemma is not found. It has therefore been easy to choose for analysis the most frequent function words, verbs, nouns & pronouns, and adjectives in the texts under consideration.

The main thrust of the research has been conducted on the data set of function words, but, because the most frequent verbs, nouns, and adjectives can be identified so accurately in a disambiguated and tagged text, it has been possible to compare the statistical results of these three parts of speech with those from function words, which themselves are considered in the literature to be standard as data in stylometric analysis. The first thing noticed, however, has been the necessity of comparing the results from a full set of lemmas, with a set from which the most frequent 2 or 3 have been removed. These few very frequent lemmas can apparently overwhelm the effects of the other lemmas, and skew the results slightly, but noticeably.

## Analysis

In a test of 37 function lemmas (with the removal of the three most frequent: *et*, *in*, and the separable suffix *que*),

the twelve works of Suetonius, the 'control' works, demonstrate that the lives remain closely grouped, as found in our earlier research, with only the life of Titus being slightly removed from all others. When all 40 of the top function lemmas are involved, however, there are slight changes in the distances of most lives, and the life of Otho joins that of Titus at the further remove, although this removal does not involve any question of a change of authorship.

In an analysis of Sallust's *Catiline* and the *Jugurthan War*, the results from an analysis of the 37 most common function lemmas reinforce the manuscript tradition of authorship, with a very close association; with all 40 of the most frequent lemmas, however, the relative spacing of the two works approaches even a possible difference in authorship.

The analysis of 37 of the 40 most common function lemmas in the works of Caesar provides a greater complexity. There is, first of all, a very close association between *Gallic VIII* and the *Alexandrian War*, with both being clearly separated from most of the other works attributed: an apparently clear vindication of the claims by Hirtius to be the author of both. The *Spanish War*, that work of execrable Latin, is obviously of totally separate authorship. The most remarkable result, however, is the distance between Book I of the *Gallic Wars* (that book with the famous beginning: "Gallia est omnis divisa in partes tres" - "All Gaul is divided into three parts" ), and the other works attributed to Caesar (other than the *Spanish War*).

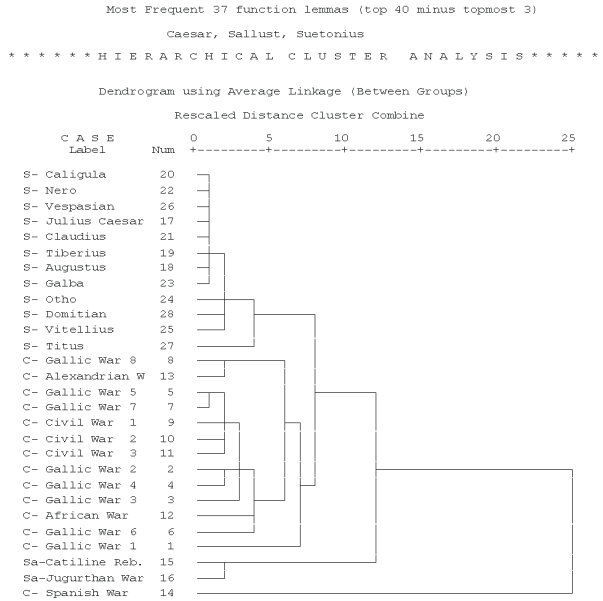


Figure 2

Principal Component Analysis

Caesar, Sallust, Suetonius

37 Function Lemmas

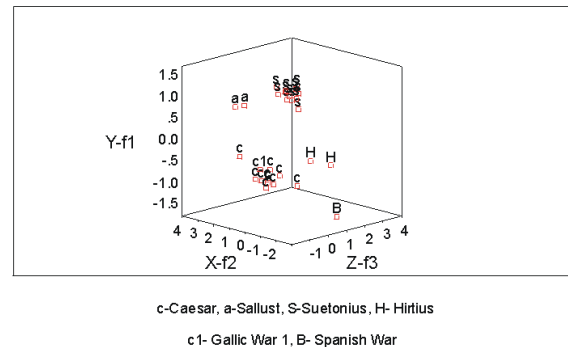


Figure 3

When all 40 top function lemmas are employed, most separations between the works attributed to Caesar increase. *Gallic I* and *Gallic IV* are now both at a further remove from the other works, and are hardly within reasonable attribution to Caesar. The *Alexandrian War*, however, is now very close to other texts, and far from *Gallic VIII*; and both the *African War* and *Gallic VIII* are now possibly beyond any reasonable attribution to Caesar, although the attribution of *Gallic VIII* to Hirtius remains possible. The *Spanish War* continues to be far distant from any other work in the 28 studied.

The analyses of verbs, nouns, and adjectives, all demonstrate considerable differences amongst the works. For example, *Gallic VIII* and the *Alexandrian War* become relatively distant in the analysis of nouns (less the top three), although they are

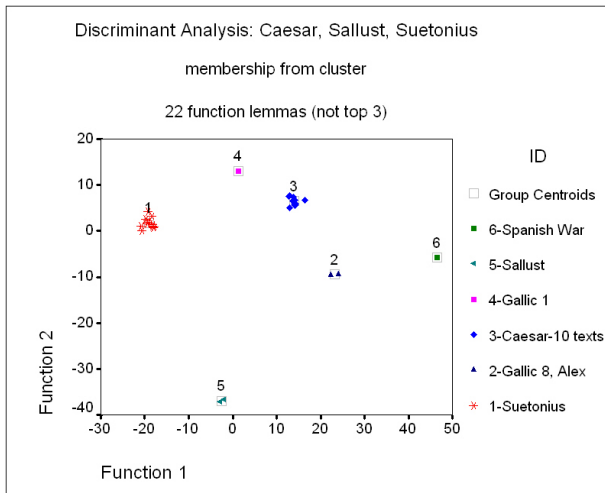


Figure 1

still closer to one another than to any other works. In the analysis of adjectives, however, they appear no longer to be related, and in addition, *Gallic VIII* becomes quite close to several other books of the *Gallic Wars* and *Civil Wars*.

## Conclusions

It appears clear that the arguments in the stylistic literature, describing the necessity of using function lemmas, remain valid. Nonetheless, verbs, nouns, and adjectives must not be discarded as being inferior to function lemmas in the identification of authorship, since they provide valuable insights to the Latinist on the individual differences in word usage by the various authors.

The conclusion to be drawn from the differentiation between the use of all most common lemmas and those lemmas with the top 2 or 3 removed appears to be that a blind use of the most frequent lemmas can skew the results, and demonstrates that close cooperation between statistician and Latinist is required.

The overall conclusion on the authorship attributions to Caesar and Sallust are clear. The two works of Sallust appear to be correctly attributed. The attributions to Caesar remain complex, however: the text on the *Spanish War* is undeniably not that of any literate Roman; and *Gallic VIII* and the *Alexandrian War* appear definitely to be of separate authorship, and quite possibly that of Hirtius, as the tradition claims. The most striking fact, however, is that the first book on the *Gallic Wars* is of a high literary quality, yet is undeniably different from the other works of Caesar. Hence it now lies in the realm of the Latinist for a full analysis of the author's uses of all parts of speech, and the manner in which he apparently poured so much more effort into this first book that brought his conquests in a new land to the attention of the Roman People who would later be voting on his bitterly fought candidature for the Consulship.

## Bibliography

- Gurney, L.W., and P. Gurney. "The Scriptorum Historiae Augustae: History and Controversy." *Literary and Linguistic Computing* 13.3 (1998): 105-109.
- Gurney, L.W., and P. Gurney. "Authorship Attribution of the Scriptorum Historiae Augustae." *Literary and Linguistic Computing* 13.3 (1998): 119-131.
- Gurney, L.W., and P. Gurney. "Subsets and Homogeneity: Authorship Attribution in the Scriptorum Historiae Augustae." *Literary and Linguistic Computing* 13.3 (1998): 133-140.

# Gottlob Frege's *Grundgesetze der Arithmetik*: Computational Linguistics Meets the Founder of Logicism

*Felicitas Haas* ([fha@ikp.uni-bonn.de](mailto:fha@ikp.uni-bonn.de))

*IKP, Univ. Bonn*

*Bernhard Schröder* ([bsh@ikp.uni-bonn.de](mailto:bsh@ikp.uni-bonn.de))

*IKP, Univ. Bonn*

Our paper deals with the text technological challenges which arise from the retrodigitalization of Gottlob Frege's *Basic Laws of Arithmetic*. The digitalization is the basis for a hypermedia presentation with various views about the text.

The German mathematician, logician, and philosopher Gottlob Frege is regarded as the founder of logicism, i.e., the view that mathematics can be completely derived from pure logic. The two volumes of the *Basic Laws of Arithmetic* are Frege's major work (1893/1903). Here he develops his logicist program in detail, i.e., he deduces central arithmetic laws from logic. Frege's work is seen as the basis of modern logic. Therefore it is of interest not only for mathematicians but also for scholars of various arts. Apart from its metamathematical considerations the *Basic Laws* are relevant especially for philosophers of language and for linguists, because Frege's investigations on sense and reference, functions and concepts are accumulated in these books.

Frege uses in his work a peculiar two-dimensional notation for formulas of predicate logic. His notation was not taken up by other logicians and is therefore regarded as hardly readable. The structure of a formula is primarily symbolized by horizontal and vertical branching lines, e.g., a little vertical line symbolizing negation. A line which runs vertically downwards and then parallel to another horizontal line represents a conditional clause. Formula parts may be written linearly. Furthermore Frege uses a lot of special symbols, which he defines in the *Basic Laws*. Like his two-dimensional notation, they were not taken up by others. For these reasons Frege's notation deviates quite a lot from contemporary logical formalisms. Compared to his smaller writings the *Basic Laws* are therefore studied less frequently.

Our project has the goal to make Frege's work accessible to a broader scientific community. We will offer various views with

modern notations. As a precondition we have digitalized the *Basic Laws* using an XML structure according to a specific DTD. The pure text part was scanned and converted into plain text by OCR. This part has posed only few problems due to the high quality of the original typesetting. The digitalization of the formula parts has cost much more effort. The logical structure of formulas had to be entered completely by hand. Automated methods failed due to the peculiarity of the formalism. Special attention had to be given to the encoding scheme of the formulas. From the XML code the original notation should be derivable as well as various modern notations, among these heuristically simplified notations. Frege refers from the text to parts and single symbols of the formulas. Changes of the form of formulas can therefore have impact on the surrounding text. For example, he represents a universal quantifier by a bow in a horizontal line which is superscripted by a Gothic character. He refers to this symbol as the *cavity* (*Höhlung*). In the text all occurrences of *cavity* must be replaced by some other description in views with modern formula notations and deviating symbols for universal quantifiers.

The XML coding of the formula trees has another scientific application apart from providing the basis for alternate views: it allows for an automated verification and analysis of Frege's proofs (formula forests). The extremely compact derivations in the Fregean calculus practically exclude a manual verification. Here automated procedures can provide essential help for the reader by 'unfolding the proofs', i.e., by making intermediate implicit proof steps fully explicit.

The immediate goal of this project is the derivation of presentations for two media. We will produce a print and web version. The target format of the print version is valid *LaTeX* documents. For the webpages we use XHTML in order to arrive at a maximally browser independent representation. We use exclusively XSLT for transformations. The print version with the original Fregean formula notation is produced primarily for control purposes. A print version with modern notational variants is applied in an ongoing commentary project.

The web version will support a couple of notational variants as well as proof unfolding. Links will be generated between text and proof references. A full text search and search facilities for formulas, definitions and lemmas will be provided.

The forms of presentation and the views sketched above will considerably enhance the accessibility of Gottlob Frege's *Basic Laws of Arithmetic* for many recipients. It provides new ways of reception for a relevant but formally peculiar text.

## Bibliography

Frege, Gottlob. *Grundgesetze der Arithmetik, Band I/II*. Jena: Verlag Herman Pohle, 1893/1903.

---

## Delta, Delta Prime, and Modern American Poetry: Authorship Attribution Theory and Method

---

*David Hoover* (*david.hoover@nyu.edu*)

*New York University*

---

In the three years since John F. Burrows presented Delta, his new measure of authorial difference, in his Busa Award lecture (2001), there has been a flurry of activity in the authorship attribution community and beyond. Delta measures the difference between test texts and a set of texts by possible authors in an elegantly simple way: the frequencies of the most frequent words in the test text and in each of the primary texts are compared with their mean frequencies in the primary set. The difference between the test text and the mean is then compared with the difference between the texts by each author in the primary set and the mean. Then the absolute values of the differences between the z-scores for all the words are summed and the mean is calculated, producing Delta, "the mean of the absolute differences between the z-scores for a set of word-variables in a given text-group and the z-scores for the same set of word-variables in a target text" (Burrows 2002a, 271). The primary author whose texts show the smallest Delta, the smallest mean difference, from the test text has the best claim to being the author of the test text.

Burrows has published two articles demonstrating the effectiveness of Delta on Restoration poetry, even for small texts (2002a, 2003), and has applied the technique to the interplay between translation and authorship in "The Englishing of Juvenal: Computational Stylistics and Translated Texts" (2002b). David L. Hoover has just published two studies involving Delta (2004a, 2004b) that automate the process of calculating and evaluating the results of Delta in an *Excel* spreadsheet with macros. Hoover's first article demonstrates Delta's effectiveness on early 20th century novels, and shows that increasing the number of frequent words to be analyzed far beyond the 150 most frequent that Burrows uses—to the 700 or 800 most frequent—substantially improves the results, as does the removal of personal pronouns and words that are frequent in the entire corpus only because they are extremely frequent in a single text. It also shows that large drops in Delta from the first to the second likeliest author are strongly associated with correct attributions. The second article shows that it is possible to improve the accuracy of attribution by Delta by selecting subsets of the word frequency list for analysis

and by changing the formula of Delta itself, and also extends the testing of the measures to contemporary literary criticism, where they continue to perform very well. These new methods recapture information about whether a word is more or less frequent than the mean, about how different the test text is from the mean, about the size of the absolute difference between the test text and each primary text, and about the direction of the difference between the test text and the primary text.

In spite of the fact that Burrows's Delta is simple and intuitively reasonable, it, like previous statistical authorship attribution techniques, and like Hoover's alterations, lacks any compelling theoretical justification. Nonetheless, it and some of the variations upon it are manifestly and surprisingly effective, even in difficult open authorship attribution situations in which the claimants cannot be limited to a small number by traditional means. Other ongoing studies that are not ready for public discussion are underway by several researchers, involving a 'real life' attribution problem on 19th century prose, another on a Middle English saint's life, and an application of the technique and its variants to biology.

In this paper I investigate the effectiveness of Delta and Hoover's various Delta Prime candidates on a corpus of 1,430,000 words of Modern American Poetry by poets born between 1902 and 1943. This investigation returns to poetry but brings the techniques forward to the 20th century. Although it is well known that changes in language and style across long spans of time are very considerable, and that many authorship attribution techniques are sensitive to these differences, preliminary results show that Delta and the various Delta Primes are even more accurate on the corpus investigated here than on the restoration poetry that Burrows investigated. They are so accurate, in fact, that the differences between the original Delta and the alternatives are relatively small (it is difficult to improve much on 100% accuracy). These results may be related to a greater individuality in poetic styles in modern poetry, with some poets using rhyme and meter and others working in much looser forms, and to the presence of dialect. Whatever the cause, however, they further demonstrate the robustness of the techniques, which have now been tested on two corpora of poetry written nearly 300 years apart, on novels from 1900, and contemporary literary criticism. Further tests on contemporary prose and on texts tagged for part of speech are ongoing, not so much in an attempt to further confirm the effectiveness and reliability of Delta and Delta Prime, which now seem very solidly validated, but rather in the hope of more fully understanding why these relatively simple techniques work so well, and in continuing to improve their already impressive power.

## The Delta Spreadsheet

---

### Bibliography

Burrows, J.F. "Questions of Authorship: Attribution and Beyond." Presented at the Association for Computers and the Humanities and Association for Literary and Linguistic Computing, Joint International Conference, New York, June 14, 2001. 2001.

Burrows, J.F. "'Delta': a measure of stylistic difference and a guide to likely authorship." *Literary and Linguistic Computing* 17 (2002a): 267-287.

Burrows, J.F. "The Englishing of Juvenal: computational stylistics and translated texts." *Style* 36 (2002b): 677-99.

Burrows, J.F. "Questions of Authorship: Attribution and Beyond." *Computers and the Humanities* 37.1 (2003): 5-32.

Hoover, David L. "Testing Burrows's Delta." *Literary and Linguistic Computing* 19.4 (2004a): 453-475.

Hoover, David L. "Delta Prime?" *Literary and Linguistic Computing* 19.4 (2004b): 477-495.

---

*David Hoover* ([david.hoover@nyu.edu](mailto:david.hoover@nyu.edu))  
*New York University*

---

John F. Burrows introduced Delta, a simple measure of authorial difference in his Busa Award lecture (2001), and further elaborated upon it in three articles (2002a, 2002b, 2003). In all of these discussions Burrows relies on an Excel spreadsheet that helps to simplify and partially automate the calculation of Delta. At the ALLC/ACH conference in Gothenburg, David L. Hoover presented the results of further tests of Delta on prose and discussed a more complex version of Burrows's spreadsheet that takes the automation of the calculation and the analysis of results much further (2004a), and he has just published two articles that rely on such spreadsheets (2004b, 2004c).

Given the burst of activity in authorship attribution circles following the introduction of Delta, many researchers are interested in using it on various projects. Unfortunately, even Hoover's 2004 versions of the spreadsheet are rather daunting in their complexity, and their macros are difficult to understand because they do not include comments. Further, the researcher must do substantial analytical work on raw word frequency lists before they can be inserted in the spreadsheet for Delta testing. Once the lists are produced, the frequencies must be transformed into text percentages and a zero frequency record must be inserted in the list for each text if any of the most frequent words does not occur in that text. This is not a significant problem for analyses using only a small number of the most frequent words because nearly all of them will occur in each text, but, as Hoover has shown (2004b, 2004c), increasing the word list to the 700-800 most frequent often improves the accuracy of the analysis, and many of the 800 most frequent words will normally fail to appear in one or more of the texts. Manually adding zero records may be an acceptable method in small analyses, but it would be an extremely time-consuming and error-prone process if the 800 most frequent words in a set of fifty or more texts were to be analyzed.

Hoover's analyses also show that removing personal pronouns and words for which a single text provides nearly all the occurrences significantly improves Delta (and other kinds of statistical analyses of authorship), and these are non-trivial processes that are difficult enough to prevent some researchers from trying out these techniques. The addition of the various possibilities for *Delta Prime* introduced in Hoover's second

article (2004c) makes for still greater complication, and seems likely to prevent the interested humanist who is not an Excel maven from further testing these innovative measures on new corpora and from using them in real authorship attribution problems.

My current project involves further elaboration of Hoover's spreadsheets to automate more of the necessary processes. Beginning with a version provided by Hoover that includes explanatory comments on the macros by Marc LeBlanc of Wheaton College (MA), I hope to produce a spreadsheet that can accept as input a list of the authors and texts, the raw word frequencies from the corpus as a whole and from the individual primary and test texts. The complete analysis will be performed within the spreadsheet itself. This will allow anyone who has access to any of the myriad of software tools that can produce ranked frequency lists to try out Delta and the various Delta Primes without needing to have expertise in text analysis, Excel, or Visual Basic. The project is currently under way, with the various formulas for calculating Delta and the various Delta Primes already added and the analytic work planned out and in progress. Initial testing has begun to determine whether or not the macros will operate with acceptable speed, and whether the limitations of Excel will impact the number of frequent words that can be analyzed. If performance proves too poor, I intend to use other methods than Visual Basic and link them as seamlessly as possible with the spreadsheet. By the time of the conference, I expect to have a fully operational version to demonstrate and distribute to anyone who is interested.

A secondary benefit of the current project is more wide ranging, having to do with the question of how to balance using the good tools for performing the analysis and manipulation of the word frequency lists (certainly Visual Basic is not one of them!), and providing a tool that is usable by the largest possible number of users, even if those users are not particularly computer literate. This has long been a question of serious interest to software developers, and the relatively small scale of this project may allow it to come to the fore in interesting ways. I hope to benefit from the expertise of conference attendees in continuing to develop and improve The Delta Spreadsheet.

## Bibliography

Burrows, J.F. "Delta': a measure of stylistic difference and a guide to likely authorship." *Literary and Linguistic Computing* 17 (2002a): 267-287.

Burrows, J.F. "The Englishing of Juvenal: computational stylistics and translated texts." *Style* 36 (2002b): 677-99.

Burrows, J.F. "Questions of Authorship: Attribution and Beyond." *Computers and the Humanities* 37.1 (2003): 5-32.

Burrows, J.F. "Questions of Authorship: Attribution and Beyond." Paper delivered at the Association for Computers and the Humanities and Association for Literary and Linguistic Computing, Joint International Conference. June 14, 2001.

Hoover, D.L. "Testing Burrows's Delta." Paper delivered at the Association for Literary and Linguistic Computing and Association for Computers and the Humanities, Joint International Conference, Göteborg, Sweden. 2004a.

Hoover, D.L. "Testing Burrows's Delta." *Literary and Linguistic Computing* 19.4 (2004b): 453-475.

Hoover, D.L. "Delta Prime?" *Literary and Linguistic Computing* 19.4 (2004c): 477-495.

# Concurrent Markup Hierarchies: a Computer Science Approach

---

*Ionut Emil Iacob* (*eiaco0@csr.uky.edu*)

*University of Kentucky*

*Alex Dekhtyar* (*dekhtyar@cs.uky.edu*)

*University of Kentucky*

---

## Abstract

It is known that text has not, in general, a regular structure. However, since its invention and despite the fact that it represents hierarchical structures, XML has gained a lot of popularity among humanities researchers: XML is easy to use and it comes with a handful of free processing tools. A variety of solutions were proposed to represent overlapping structures in XML. More or less easy to maintain from the point of view of data management, none of these solutions provides full support for two of the most demanded processing tasks: querying and presentation (XSL-like transformation).

We propose a processing framework for complex document-centric XML which generalizes the traditional way of XML data management to support overlapping markup processing. Our framework provides support for overlapping structures representation in XML, querying, authoring, and presentation of overlapping hierarchies.

## 1. Introduction

The newborn TEI Overlapping Markup Special Interest Group comes to support the fact that overlapping XML structures are of great interest for the text encoding community. Why is XML so popular? First at all, XML is the legitimate descendant of SGML, which was also popular among humanists. Then there is the fact that XML comes with a handful of processing (free) tools. This fact is clearly expressed by TEI's "Strategic Considerations in Migration of TEI Documents from SGML to XML". More specifically, DOM, SAX, XPath, and XSL and the companion software are very attractive for humanities computing. In addition, XML is flexible, intuitive, and readable: it is text, isn't it? However, there is an annoying detail about XML that does not fit into the same picture with text encodings: XML allows only properly nested markup structures. However, overlapping structures

(concurrent hierarchies) often occur in applications. Czymiel points out that the proposed solutions for the overlapping markup problem fall in three categories: XML based workaround (milestones and fragmentation suggested by TEI in Sperberg-McQueen & Burnard), new markup languages such as LMNL (Tennison et al.), MECS (Huitfeldt), and TexMecs (Huitfeldt & Sperberg-McQueen), or content and structure separation (standoff markup, JITTs (Durusau & O'Donnell)). None of the solutions previously presented contains complete answers for the problems of management of concurrent XML data.

Part of the problem is the correct identification of what is the "problem of overlapping markup". For some researchers in humanities, the problem lies in determining the markup elements that can overlap each other, and in devising a way to apply one of the TEI-based suggestions, such as milestone elements. For others, the problem lies in specifying the correct order in which elements overlap, e.g., determining whether paragraph markup must commence inside or outside the page markup.

Such considerations are of importance because they are dictated by the nature of the documents and encoding under consideration. Yet, successful resolution of such issues for specific projects does not constitute solving the overarching problem of representation, storage, management and querying of overlapping markup. Approaching the general solution of such a problem requires a change of a viewpoint from that of a humanities scholar working on document encoding, to the viewpoint of a computer scientist striving to provide generally applicable methodology, algorithms and software tools.

In this paper, we address the issues related to overlapping markup in document-centric XML documents from the perspective of computer scientists. We show that, in general, the framework for processing multihierarchical markup is the same as the framework for processing single-hierarchy XML documents. The difference comes not in the laundry list of tasks, and not in overall organization of the framework, but rather in the approaches to solving individual subproblems/tasks within it.

The processing framework we describe below covers all the core XML processing tasks: representation and parsing, data structure, querying, and presentation.

## 2. A Framework for Management of Concurrent Markup Hierarchies

The framework we propose (Figure 1) generalizes the traditional XML processing framework: parsing an XML document into a DOM data structure (or, alternatively,



constructing DOM from an XML database), then using the DOM API support for editing, querying, and transforming the XML document.

The core of our processing framework is the data structure for storing concurrent XML markup: the GODDAG data structure first introduced by Sperberg-McQueen and Huitfeldt. We enhanced the GODDAG with API and we have designed and implemented a parser for building a GODDAG (Iacob, Dekhtyar & Kaneko) from separate XML files, one file per hierarchy (this would present the advantage of a basic concurrency control over authoring the document encodings). In general, a GODDAG data structure can be built using specialized drivers for different concurrent markup representations.

In Iacob, Dekhtyar & Zhao, we present an extension of the XPath query language for querying concurrent markup represented as a GODDAG. As GODDAG represents a "multidimensional" generalization of DOM, our extension of XPath generalizes XPath to deal with concurrent hierarchies. In the absence of multiple hierarchies, GODDAG reduces to DOM whereas extended XPath reduces to XPath (the extended XPath semantics is given at: <http://dblab.csr.uky.edu/~eiaco0/docs/expath/>). With the parser and the query language we provide answers to two of the open problems in Sperberg-McQueen & Huitfeldt. Moreover, the presentation issue (XSL) is implicitly solved as we employ patterns expressed in the extended XPath language we propose. The XML editorial tools are based on the GODDAG API: text (PCDATA) update, markup insertion and deletion, and searching (using the XPath extension).

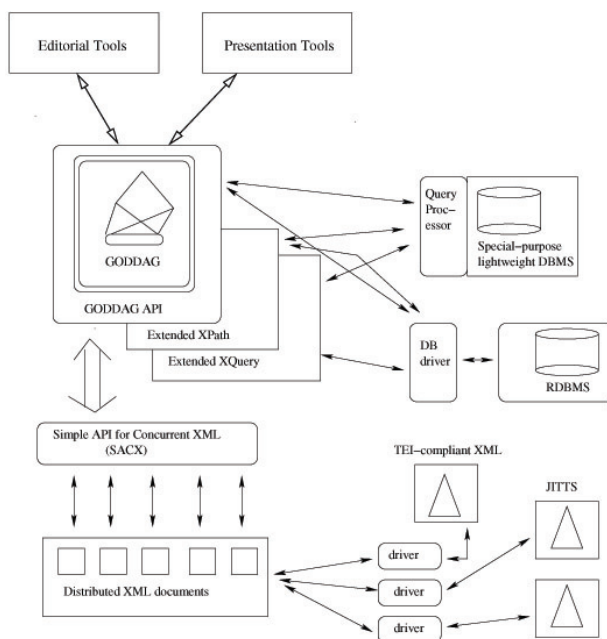


Figure 1: A framework for management of Concurrent XML Hierarchies

For representing and storing concurrent XML markup we defined the notion of *distributed XML document* (Dekhtyar & Iacob): a virtual collection of XML documents, one document per hierarchy. The distributed XML document is obtained via drives from various representations: BUVH and JITTS introduced by Durusau and O'Donnell, XML documents with fragmentation and/or milestones (as in TEI).

Finally, we are currently working on implementing persistent storage support for concurrent XML hierarchies: a specialized database for storing XML with overlapping structures. Our plans include providing support for storing XML with overlapping structures in a relational database.

The framework for processing concurrent XML markup is successfully implemented in the *ARCHway* and *Electronic Boethius* projects (<http://www.rch.uky.edu/>) at the University of Kentucky. The APIs and (part of) the software programs are available at: <http://dblab.csr.uky.edu/~eiaco0/research/cmh/>.

## Bibliography

Bauman, S., A. Bia, L. Burnard, T. Erjavec, J. Hekman, T. Rischer, C. Powell, C. Ruotolo, S. Schreiber, N. Smith, J. Walsh, S. Wells, and F. Wiering (TEI Task Force on SGML to XML Migration). *Strategic Considerations in Migration of TEI Documents from SGML to XML*. Text Encoding Initiative, 2004. Accessed 2005-04-11. <http://www.tei-c.org/Activities/MI/miw02.html>

Czmiel, A. "XML for Overlapping Structures (XfOS) using a non XML Data Mode." Paper delivered at the Joint International Conference of the Association for Humanities Computing and the Association for Literary and Linguistic Computing, June, Göteborg, Sweden 2004. 2004. Accessed 2005-05-25. <http://www.hum.gu.se/allcach2004/AP/html/propl04.html>

Dekhtyar, A., and I.E. Iacob. "A Framework for Management of Concurrent XML Markup." *Special Issue Data and Knowledge Engineering* 52 (2005): 185–208.

Durusau, P., and M.B. O'Donnell. "Concurrent Markup for XML Documents." *Proceedings of XML Europe*. Atlanta, Georgia, 2002. Accessed 2005-04-11. [http://www.idealliance.org/papers/xml02/dx\\_xml02/papers/03-03-07/03-03-07.html](http://www.idealliance.org/papers/xml02/dx_xml02/papers/03-03-07/03-03-07.html)

Huitfeldt, C. "MECS - A Multi-Element Code System ." 1998. Accessed 2005-04-11. <http://helmer.hit.uib.no/clauss/mecs/mecs.htm>. Forthcoming in Working Papers from the Wittgenstein Archives at the University of Bergen, No 3.

Huitfeldt, C., and C.M. Sperberg-McQueen. "TexMECS: An experimental markup meta-language for complex documents." February 2001.

Iacob, I.E., A. Dekhtyar, and W. Zhao. "XPath Extension for Querying Concurrent XML Markup." *Technical Report TR 394-04*. University of Kentucky Department of Computer Science, February 2004. <<http://www.cs.uky.edu/~dekhtyar/publications/TR394-04.pdf>>

Iacob, I.E., A. Dekhtyar, and K. Kaneko. "Parsing Concurrent XML." *Proceedings of the 6th ACM International Workshop on Web Information and Data Management (WIDM 2004)*, Washington, DC. November 2004.

Sperberg-McQueen, C.M., and Lou Burnard, eds. *Guidelines for Electronic Text Encoding and Interchange*. Chicago and Oxford: TEI P4, 2001.

Sperberg-McQueen, C.M., and C. Huitfeldt. September 2000. Early draft presented at the ACH-ALLC Conference in Charlottesville, June 1999.

Tennison, J., G.T. Nicol, and W. Piez. *Layered Markup and Annotation Language (LMNL)*. Dissertation, University of Wisconsin, 2002. First introduced at the Extreme Markup Languages Conference 2002, Montreal.

---

## ***Edition Production Technology: an Eclipse-Based Platform for Building Image-Based Electronic Editions***

---

***Ionut Emil Iacob*** ([ionut@ms.uky.edu](mailto:ionut@ms.uky.edu))

*Department of Computer Science, University of  
Kentucky*

***Kevin Kiernan*** ([kiernan@uky.edu](mailto:kiernan@uky.edu))

*Department of English, University of Kentucky*

***Alex Dekhtyar*** ([dekhtyar@cs.uky.edu](mailto:dekhtyar@cs.uky.edu))

*Department of Computer Science, University of  
Kentucky*

---

We are developing the *Edition Production Technology (EPT)*, an integrated development environment for building image-based Electronic Editions (IBEE) (Kiernan 2005), through the *Electronic Boethius* (Kiernan and Porter 2005) and ARCHway Projects (Kiernan et al. 2004; Kiernan et al. 2005) at the University of Kentucky. We built the EPT using *Java*, and it operates through the *Eclipse* platform, benefiting from *Eclipse's* open architecture and portability. Currently the *EPT* runs on *Windows XP*, *Linux*, and *Mac OS X*.

The goal of the *EPT* is to provide software support for building image-based electronic editions of cultural manuscripts. Starting with images and text, the *EPT* enables the editor to create an electronic edition with complex, pervasive XML encodings, search the electronic edition, link text and images, and deploy the completed electronic edition using filters and XSLT.

A fully functional demo version of the *EPT* software suite for PC, including sample projects, is available for download at <<http://rch01.rch.uky.edu/~ept/download>>.

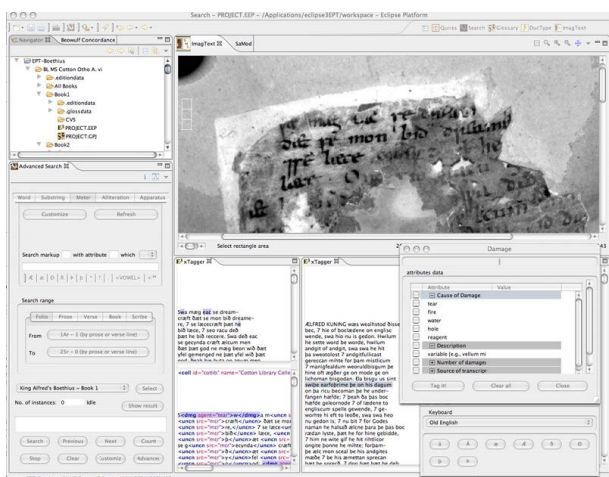


Figure 1. A snapshot of EPT illustrating image-based encoding through *ImagText*, *xMarkup*, and *xTagger* (including an XML view). The figure also shows the Keyboard panel and Search Tool.

## Editorial Tools on EPT Platform

- Project wizard initializes an electronic edition project. The input data consists of image files, text content (or partially encoded text) and one or more DTDs (*EPT* provides support for concurrent markup using multiple DTDs).
- *xMarkup*, *xTagger* and *ImagText* form the core component for encoding image-based, document-centric XML, working together to link text and image. Through *xMarkup*, the editor selects edition markup through a series of simple, configurable templates. *xTagger* introduces markup into the text and provides filtered XML views, while *ImagText* associates that text section with the corresponding image region, selected by the editor. In summary, the tagging works as follows: the editor selects text and image (in any order), describes the manuscript or textual feature, choosing tags and attribute values through *xMarkup*, and inserts the markup. See Figure 1 for an illustration of the cooperation between *xMarkup*, *xTagger*, and *ImagText*. The *xTagger* ensures that the new markup is well-formed and potentially valid (Iacob, Dekhtyar, and Dekhtyar).
- *DucType* provides a specialized interface for describing individual manuscript letters. The editor configures *DucType* through the Letter Template, which also creates and maintains a repository of letter images for a manuscript. The letter images are then used by the *DucType* tool as base of comparison with any letter image in manuscript image.
- *Overlay* provides image manipulation support. An editor may find that multiple images of the same folio are required for a complete view of the manuscript. Using this tool the editor lays one image over another image of the same folio (using, for instance, ultraviolet and normal lightening conditions) and changes the transparency of the upper layer, enabling a useful comparison of the two images.
- *Samod* is a specialized tool for creating manuscript text collations with text from multiple sources (for example, the same text found in different manuscripts). This tool recognizes differences between transcripts and marks up these differences as variants of the text the editor identifies as the base text.
- *StaTend*: Using a transcript marked with basic navigational markup – folio and folio line tags – this tool calculates manuscript statistical tendencies (number of folios, lines per folio, characters per line, etc.). Based on these statistical tendencies the tool reconstructs missing folios for which we can supply the text from another source, based on these statistics. The *StaTend* tool also includes functionality, called *RamSome*, for taking these textual 'virtual folios' and translating the text into image, built character by character using letters taken from the manuscript.
- *Quires* is a specialized interface for the edition and visualization of codicological markup. It allows the editor to build a virtual map of the physical object. We used this tool in the *Electronic Boethius* project to reconstruct the gatherings of a manuscript whose binding was destroyed by fire.
- The Search GUI is an interface for searching the edition. The editor can configure it to search any combination of XML markup, while hiding the intricacies of the query language (an extension of *XPath* that supports multiple hierarchies).
- Datalayer is the API for data access in *EPT*. Tools request and deliver edition data (image and text files, DTDs, etc.) through the Datalayer API, which can interface with a variety of data storage devices, whether a database, file system, or remote server.
- *Glossary* is a data-centric XML editor for creating a glossary including each word from the edition text. It automatically generates a complete word list from a transcript file encoded with basic formatting information (folio and folio line markup). The glossary links its entries to the text through the <word> tag – changes made within the edition text are automatically reflected in the glossary. It provides customizable templates for parts of speech and tools for saving the information in XML format (used later on for searching purposes) and HTML format (used for display glossed information).
- The HTML browser provides HTML display and general browser support in *EPT*. Having a browser integrated in the platform enables the *EPT* to direct XSL transformations dynamically to the browser.

- The Keyboard panel enables the editor to configure keyboards containing special characters (Old English *æ,ð*, and *þ*, Greek characters, etc.).

In addition to editorial tools, the *EPT* provides support for project management such as: Project properties editor is a GUI for various settings related to the project, such as fonts, encoding, title, etc. It provides support for adding and removing project images and for customizing markup tags, grouping tags in meaningful use categories, assigning aliases to tags and attributes, and adding and removing DTDs from a project. XML filter allows the editor to create encoding filters for viewing different combinations of elements from the entire set. The output of a filter can be used for visualization, XSL transformation, or data interchange. Extended XPath search is a search GUI using extended XPath language (an extension of XPath that applies to concurrent markup structures).

From the *Eclipse* platform, *EPT* inherits three important features for project development: versioning control (CVS), automatic updates, and help content support. The editing team uses CVS to share project work-in-progress and as projects repository. Updates are useful for providing tools updates as well as bug fixes: an *EPT* user need only check for updates and download them if available. Finally, the open help architecture enables the editor to create and use help files in such a way that the application help information is added independently of the application program.

## Demo overview

Our demonstration will begin with examples of the most basic *EPT* functionality, and depending on time we will demonstrate any tool or function. We will begin by creating a project and going through the usual operations for preparing an image-based electronic edition: content markup (using only text projections or filtered XML views), automatic linking of images and text, and text updates. We will demonstrate that our document-centric XML editor (*xTagger*) can significantly simplify and speed up the encoding process. The editor can search for the information, visualize the encodings using customizable filters, or change project properties at any point in the editorial process. We will demonstrate the support for overlapping markup structures by adding/removing DTDs and markup encodings from external files. Depending on the interests of the audience, we can also show how a project can be customized, starting with user interfaces (toolbox, fonts, encodings, etc.) and ending with markup customization: associating aliases to tag elements and attributes, grouping tag elements by functionality, and displaying status bar information based on XPath queries. We will also be prepared to demonstrate *Quires*, *Overlay*, and *DucType*, and show how to customize *DucType*. Statistical information for the project encodings can be obtained

dynamically and we can show how this information can be used in folio reconstruction (text and image) for missing manuscript part. We can also demonstrate *SaMod*, showing how it collates several different texts.

The demo may also include automatic generation of HTML content from edition data (glossaries, manuscript edition and manuscript transcription).

We emphasize during the demonstration how the *Eclipse*'s open architecture is an excellent platform choice for implementing the *EPT*.

## Bibliography

- Iacob, Ionut Emil, Alex Dekhtyar, and Michael I. Dekhtyar. "Checking Potential Validity of XML Documents." *Proceedings, Seventh International Workshop on the Web and Databases, WebDB@SIGMOD/PODS*. 2004. 91-96.
- Kiernan, Kevin S. "Digital Facsimiles in Editing: Some Guidelines for Editors of Image-based Scholarly Editions." *Electronic Textual Editing*. Forthcoming. A volume of essays jointly sponsored by the Modern Language Association and the TEI Consortium, funded by the Mellon Foundation, and co-edited by John Unsworth, Katherine O'Brien O'Keefe, and Lou Burnard, 2005.
- Kiernan, Kevin S., Alex Dekhtyar, Jurek Jaromczyk, Dorothy C. Porter, and Ionut Emil Iacob. "Edition Production Technology (EPT) and the ARCHway Project." *DigiCULT.Info* (August 2004): 36-38.
- Kiernan, Kevin S., Jurek Jaromczyk, Alex Dekhtyar, Dorothy C. Porter, Kenneth Hawley, Sandeep Bodapati, and Ionut Emil Iacob. "The ARCHway Project: Architecture for Research in Computing for Humanities through Research, Teaching, and Learning." *Literary and Linguistic Computing* (Forthcoming). Special issue, papers from Joint International Conference of the Association for Computers and the Humanities and the Association for Literary and Linguistic Computing, 2003
- Kiernan, Kevin S., and Dorothy C. Porter. "Edition Production Technology (EPT) and the Electronic Boethius Project." *DigiCULT* (Forthcoming).

---

# Animated Dynamic Highlighting

---

**Bill Janssen** (*janssen@parc.com*)

*Palo Alto Research Center*

**Olga Gurevich** (*olya@socrates.berkeley.edu*)

*University of California, Berkeley*

**Lauri Karttunen** (*karttunen@parc.com*)

*Palo Alto Research Center*

---

## 1. Introduction

The recent years have seen an exponential increase in the amount of information available through the Internet on any given topic. Information retrieval techniques have been steadily improving and can provide a mass of relevant results, but those results still have to be processed and digested by a human reader. Information professionals need technology that helps people absorb large amounts of text quickly. We introduce *Animated Dynamic Highlighting (ADH)*, an interactive, user-controlled technology to improve presentational aspects of the reading task. We present the research underlying the ideas of *ADH*, the *ADH* technology itself, and some results from an initial user study evaluating its effectiveness and usability.

## 2. Background

The study described in this paper is part of a larger effort at PARC called *Productive Reading*. We are looking at ways in which computation can be applied to the reading process, in two major ways: to enhance document content, and to enhance the user experience of reading.

The current model of the reading interface is heavily based on the static experience of words imaged on paper. This model has been carried over directly to the presentation of text to the computer screen. Some attention has been given to using computation to modify the presentation structure of documents (Beveret al. 75-87; Walker et al.), but with certain exceptions (Chang et al.). These presentations are inherently static.

The major exception to this is the presentation technique commonly known as *rapid serial visual presentation (RSVP)*. The overview of studies in *RSVP* given in Slicheritz suggest that a dynamically altered presentation of text may be able to enhance comprehension without negatively affecting reading

speeds. However, *RSVP* is often found to suffer from some serious disadvantages, notably eyestrain, usually attributed to the fact that the user's eyes do not move from a fixed position, and user anxiety, due to the inability to look back at previously-read text. Other studies such as Castelhana et al. have demonstrated ways to alleviate some of these issues.

## 3. ADH

### 3.1. What ADH does

The goal of *ADH* is to preserve the apparent advantages of *RSVP*, while mitigating the apparent disadvantages. It paces the user through an electronic document, sequentially highlighting parts of the text, each a few words long, without modifying the spatial layout of the original page, so that the reader's eyes move in a normal reading fashion. The speed with which the highlighting moves depends on properties of the chunks and on a base speed set by the user. The reader can adjust the speed, and also restart *ADH* from any point in the document. The reading speed may be at a speed somewhat faster than the user's habitual reading speed.

### 3.2. The viewing technology

The *ADH* presentation system is part of a larger system at PARC for archiving and reading documents, called *UpLib* (Chang; Mackinlay; Zellweger). The *UpLib* system includes a document reader, called *ReadUp*, which normally supports a conventional page-oriented document display. *ReadUp* was modified to present documents in both *RSVP* and *ADH* mode.



Figure 1: A document page shown with ADH highlighting

### 3.3. Phrase-breaking technology

The text of a document is first annotated with part-of-speech tags using the *Inxight* tagger. In contrast to most taggers, the *Inxight* tool has a large inventory of labels to distinguish between different types of determiners, adverbs, and pronouns. While the information is less detailed than a syntactic parser could produce, the markup makes it possible to divide the text into semantically coherent pieces. We have defined a large set of phrasal patterns and compiled them into finite-state transducers (Beesley; Karttunen). The transducers are applied in a cascade taking the output of one pattern matching step as input to the next one. This process splits the input text into phrases proceeding from larger constituents (sentences and clauses) to smaller constituents (NPs, VPs, PPs) and their components. Each phrase should contain between 2 and 4 content words (such as nouns, verbs, adjectives, and adverbs); the boundaries of syntactic constituents are in most cases preserved. An example of a partitioned sentence is below:

```
<phrase>The Marine Corps band</phrase>
<phrase>played the national
anthem</phrase> <phrase>as Dailey
unveiled a space-suited Glenn</phrase>
<phrase>in his new place of
honor,</phrase> <phrase>suspended 40
feet above the floor</phrase> <phrase>of
```

```
the museum's breathtaking Gallery
100.</phrase>
```

Finally, the established phrase boundaries are projected back to the original source text to enable the dynamic highlighting in presenting the text to the user.

### 3.4. Display timing

Each phrase is allocated an initial display time based on the user-selected speed. This base span is then modified in a number of ways: shorter phrases get somewhat less time, longer ones more time. The timespan is further modified to reflect the findings in Just; Carpenter: phrases ending a line, at the end of a page, at the beginning of a new line, or ending a sentence all receive varying amounts of extra time, reflecting the extra time human subjects tend to take with these kinds of phrases. Finally, the occurrence of linguistic constructs in the phrase, such as pronouns and compound nouns, is used to modify the timespan in additional ways.

## 4. User Study

### 4.1. Method

The goal of the user study was to assess the effectiveness of *ADH* and to compare it to *RSVP* (Sicheritz); the same phrase-breaking and timing were used for *ADH* and *RSVP*. Eighteen test subjects, mostly researchers and interns, were given three alternative modes of presenting documents: plain (not modified in any way), *ADH*, and *RSVP*. The texts contained simple factual information and were followed by questions testing the recall accuracy. The first stage of the experiment used documents with automatic phrase breaking, the second one used manual phrase breaking.

The subjects were also asked about their reactions to the *ADH* and *RSVP* technologies.

### 4.2. Results

Although there were too few subjects for significant results, some interesting trends emerged. Overall, *ADH* was found to be faster than either plain or *RSVP* mode; it was also somewhat less accurate. In general, there was a tradeoff between speed and accuracy in *ADH*: the faster a document was read, the less accurate was the recall. However, both the speed and accuracy results were better with manual phrase-breaking than with automatic phrase-breaking. Users found both *ADH* and *RSVP* to be somewhat annoying, but rated *RSVP* worse than *ADH*. However, most said they would use *ADH* again for skimming through short articles, especially with improved phrase-breaking

and timing algorithms. On the other hand, most users rejected future uses of *RSVP*. The lower user ratings and reading speeds may be the result of novelty shock. The results are nevertheless encouraging: younger subjects in particular were very enthusiastic about *ADH*, and the user study produced many suggestions for future improvements and well as possible applications of *ADH*.

## 5. Conclusion

*ADH* is one of the many possibilities inherent in the idea of actively presented text. Interfaces that attempt to work with the user in understanding the underlying text would seem to have wide applicability for reading text of all kinds, from technical papers to email to biography, particularly in overview reading, such as Adler's *systematic skimming* and *superficial reading* (van Doren; Adler). They may offer special advantages to those with reading disabilities, or for specific tasks, such as proofreading. Our initial investigations into this technique seem promising, and a number of improvements in both phrase analysis and presentation timing are already being investigated.

## Bibliography

Beesley, Kenneth, and Lauri Karttunen. *Finite State Morphology*. Stanford: CSLI Publications, 2003.

Bever, Thomas G., Rebecca Burwell, Steven Jandreau, Ronald M. Kaplan, and Annie Zaenen. "Spacing printed text to isolate major phrases improves readability." *Visible Language* 25 (1990): 75-87.

Castelhano, Monica S., and Paul Muter. "Optimizing the reading of electronic text using rapid serial visual presentation." *Behaviour & Information Technology* 20.4 (2001): 237-247.

Chang, Bay-Wei, Jock Mackinlay, and Polle T. Zellweger. "Fluidly revealing information in Fluid Documents." *Proceedings of Smart Graphics 2000 AAAI Spring Symposium*. Stanford University, 2000.

Janssen, William C., and Kris Popat. "UpLib: a universal personal digital library system." *Proceedings of the 2003 ACM symposium on Document Engineering*. Grenoble, France, 2003. 234-242.

Just, Marcel Adam, and Patricia A. Carpenter. "A theory of reading: From eye fixations to comprehension." *Psychological Review* 87 (1980): 329-354.

Sicheritz, Karen. *Applying the Rapid Serial Presentation Technique to Personal Digital Assistants*. Master's Thesis, Department of Linguistics, Uppsala University, Sweden, 2000.

van Doren, Charles, and Mortimer Adler. *How to Read a Book*. New York: Simon & Schuster, 1972.

Walker, Randall C., and Stan D. Walker. *An Introduction to Live Ink Technology*. Rochester, MN.: Walker Reading Technologies, Inc., 2001.

# The ARCHway Software Infrastructure: a Demo of a Platform and Utilities for Building Applications for Electronic Editions

---

**Jerzy W. Jaromczyk** ([jurek@cs.uky.edu](mailto:jurek@cs.uky.edu))

*University of Kentucky, Computer Science*

**Neil Moore** ([neil@s-z.org](mailto:neil@s-z.org))

*University of Kentucky, Computer Science*

---

As advocated in our paper (Jaromczyk & Bodapati), the *Eclipse* platform (*Eclipse Platform Technical Overview*) is a suitable choice for building complex systems in environments that involve interdisciplinary collaboration and the involvement of researchers and students at different levels. *Eclipse* provides a *universal tool platform* (*Eclipse*) which allows different software tools to work together using a plugin system to form an integrated whole. This architecture underlies the *Edition Production Technology* (Kiernan et al.), an integrated set of tools for creating, managing, and viewing image-based electronic editions (IBEEs) (Dekhtyar et al.). In order to provide more comprehensive and integrated support for the needs of editors and students of electronic editions, we provide an additional, task-oriented, infrastructure which provides functionality for extending *Eclipse* to support tasks for editing electronic editions.

In this demonstration, we will present several pieces of this infrastructure. The infrastructure includes support for management of resources essential for humanities scholars interested in electronic editions, and a uniform structure for accessing and combining transcripts, concurrent XML documents, and images. It is designed to accommodate the varying needs of editors and allows for specializations ranging from simple adaptation through the graphical user interface (GUI), to medium-level changes based on XML configuration files and wizards, and finally to advanced customization by developing new plugins in Java. As such, it is intended to meet needs of individual scholars or large groups.

The infrastructure we present consists of a number of parts, organized into *layers*. The lowest layer, sitting directly atop *Eclipse*, is the Data Layer, a framework for uniform access to resources in different physical locations such as filesystems,

databases, and web servers. Above the Data Layer is the Project Explorer, a tool for modelling resources and projects for IBEEs and navigating those models. Making use of both the Data Layer and the Project Explorer is the Resource Registry, which organizes collections of resources according to user-defined criteria. Finally, higher-level components of the *EPT* use these tools as a framework for managing and organizing the data in an image-based electronic edition.

This demonstration will present the infrastructure and show how it fits into the general architecture of the *EPT*. We will furthermore give examples illustrating how the tools can be customized. We demonstrate the use of the Project Explorer to create, manage, and navigate electronic edition projects, accessing both local and remote data sources through the Data Layer. We will also show how the Project Explorer's model-based approach to projects allows for highly customizable views of the structure of an electronic edition. In particular, the Resource Registry contributes a model of the contents of a project that can be changed on the fly at runtime.

In addition to this infrastructure, we will present tools which make use of the infrastructure to help editors produce electronic editions. The Line Tracer allows editors to annotate (mark up) images in an electronic edition, using a model based on the segment tree data structure (Dekhtyar et al.; Jaromczyk & Moore) that provides a natural support for the concurrent XML based on image tagging. Another tool, the Image Morpher, works with the Line Tracer and Resource Registry to *re-morph* manuscript pages, correcting deformations of the lines of text introduced by centuries of wear and tear.

The demo will focus on illustrating how the implemented infrastructure can be customized to meet a broad range of needs related to individual and collaborative work on electronic editions.

## Bibliography

Dekhtyar, Alex, et al. "Database Support for Image-based Electronic Editions." *Proceedings, 10th International Workshop on Multimedia Information Systems (MIS 2004)*. College Park, MD, August 25-27, 2004. 147-156.

*Eclipse*. Accessed 2004-10-11. <<http://www.eclipse.org/>>

*Eclipse Platform, Technical Overview*. Object Technology International, Inc. Last modified Feb. 2003. Accessed 2005-04-12. <<http://www.eclipse.org/whitepapers/eclipse-overview.pdf>>

Jaromczyk, Jerzy W., and Sandeep Bodapati. "An Architecture Promoting Collaborative Research, Teaching, and Learning."



*Proceedings, Joint International Conference of the Association for Computers and the Humanities and the Association for Literary and Linguistic Computing*. Athens, GA, May 29-June 2, 2003. 10.

Jaromczyk, Jerzy W., and Neil Moore. "Geometric data structures for multihierarchical XML tagging of manuscripts." *Proceedings, 20th European Workshop on Computational Geometry*. Seville, Spain, 2004. Accessed 2005-04-14. <<http://www.us.es/ewcg04/Articulos/jaromczyk.ps>>

Kiernan, Kevin, Jerzy W. Jaromczyk, Dekhtyar, Alex et al., Dorothy Carr Porter, Kenneth Hawley, Sandeep Bodapati, and Ionut Emil Iacob. "The ARCHway Project: Architecture for Research in Computing for Humanities through Research, Teaching, and Learning." *Literary and Linguistic Computing, 2005* (Forthcoming). Special issue, papers from Joint International Conference of the Association for Computers and the Humanities and the Association for Literary and Linguistic Computing, 2003.

---

## In Search of Humanities Computing in Teaching, Learning and Research

---

*Martyn Jessop* ([Martyn.jessop@kcl.ac.uk](mailto:Martyn.jessop@kcl.ac.uk))

*King's College London*

---

### Introduction

Humanities computing as a field of study has many facets which can be used to define it. John Unsworth (2000) suggested seven 'Scholarly Primitives' of discovering, annotation, comparing, referring, sampling, illustrating and representing. These low-level research methods combine and feed into each other to form the basis for higher-level scholarly activity not just in the humanities but throughout the academy. A primary task of humanities computing is to provide the technological tools to allow academics to apply these primitives to the range of digital data and resources available across computer networks and to ensure the viability of these resources into the future.

Willard McCarty and Harold Short (2002) have produced a rough intellectual map of humanities computing. At the centre of the map is a large 'methodological commons' of computational techniques shared among the disciplines of the humanities and closely related social sciences, e.g., database design, text analysis, numerical analysis, imaging, music information retrieval and communications. Each disciplinary group contributes techniques to the methodological commons. As new applications of these techniques are demonstrated in other disciplines they in turn are exported from the commons into new disciplinary groups. Humanities computing is the agency that oversees this development process taking methods from one discipline, developing them and then applying them in other disciplines. Part of this process is the identification or creation of the tools to fulfil the roles of the scholarly primitives described earlier. The tools do not exist in isolation; they must be developed and used in ways that satisfy the scholarly criteria of all the disciplines involved in their production, which is again a role of humanities computing specialists.

The nature of humanities computing can also be explored by looking at the teaching and learning taking place in the courses that are intended to prepare the next generation of practitioners. Researchers who are active in the field design and implement the curriculum of these courses. The content therefore reflects

what they believe students need to know but do these courses reveal more about humanities computing than is written in the course handbook?

## Courses at King's College London

The Centre for Computing in the Humanities (CCH) at King's College London offers undergraduate and postgraduate degree programmes in Humanities Computing and Digital Culture and Technology. This paper focuses on the teaching and learning that takes place primarily in the final year of the B.A. Minor Programme in Humanities with Applied Computing. The final year is taken up with a practical applied computing project. This offers an opportunity to examine how effective the rest of the programme has been in equipping the student to tackle the ill-defined, open-ended style of question asked by researchers in the humanities. Although the emphasis here is on the final year of the undergraduate programme many of the issues it raises apply equally to the other programmes of study and some examples may be drawn from them and indeed from some of the research projects at King's.

## Final Year Projects

The School of Humanities at King's College London has shown a high level of commitment to developing the effective use of applied computing in research, teaching and learning in the Humanities. Their support has allowed the Centre for Computing in the Humanities to play a central role in developing humanities computing at King's and in the wider academic and cultural heritage communities. The students on the humanities computing courses at King's are drawn from a broad range of humanities disciplines and have opportunities to examine an extensive set of humanities computing projects first hand at King's. Because of this, applications of humanities computing chosen by the students for their projects are varied and extensive. The projects vary considerably in content and scope but many involve the creation of a digital resource; examples from recent years have included:

- a computer assisted learning module for learning verbs with common roots in Modern Greek;
- investigating the use of computer animation to analyse Naval Battles;
- a study of the representation of women in three French novels from the nineteenth century using text analysis tools;
- using a database to investigate patterns of involvement by individuals and institutions in corruption scandals in France during the 1990s;

- exploring the effects of the Mexican Revolution on the demography of Mexico using a Geographical Information System;
- an Investigation into how global warming is portrayed by Online Resources;
- an XML Mark-up Scheme for Texts in a Virtual Museum of Latin American iconography.

The applications of computing techniques in the humanities are evolving rapidly, as is the technology being used by the students. Many challenges are posed by this rapid change when supervising and assessing the projects.

## Conclusion

Humanities computing does not exist in isolation. It integrates a large body of knowledge from the humanities disciplines and many facets of computing and information science into a single discipline. This integrated body of knowledge has to be applied in a way that satisfies the scholarly criteria of each of the original source disciplines. The level of integration means that the teaching of humanities computing should affect teaching and curriculum development elsewhere. This raises issues surrounding the institutional role of humanities computing and new media within the contemporary academy, including curriculum development and collegial support for activities in the fields with which it exchanges knowledge.

This paper reflects on the use of project work as a means of teaching humanities computing. Pedagogic, and more pragmatic issues are discussed from the viewpoints of both the teacher and learner. The experiences of staff and students on the undergraduate and postgraduate courses in humanities computing at King's College London are used to explore the nature of humanities computing. The project work performed in the Centre for Computing in the Humanities at King's will also be drawn on to illustrate key issues where appropriate.

## Bibliography

- Biggs, J. *Teaching for quality learning at university*. Buckingham: Open University Press, 1999.
- Booth, Wayne C., Gregory G. Colomb, and Joseph M. Williams. *The Craft of Research*. Chicago: University of Chicago Press, 2003.
- Borger, R., and A. Seabourne. *The Psychology of Learning*. London: Penguin, 1966.

Botkin, J., M. Elmandjra, and M. Malitza. *No limits to learning: Bridging the human gap. A report to the Club of Rome*. Oxford, UK: Pergamon Press, 1979.

Freire, P. *The Pedagogy of the Oppressed*. Harmondsworth: Penguin, 1972.

Jarvis, P., J. Holford, and C. Griffin. *The Theory and Practice of Learning*. Routledge Falmer, 2003.

Jessop, M. "Humanities or Computing? The Growth and Development of Humanities Computing." *Association of Computer Machinery* 41.5 (December 2004).

Jessop, M. "Teaching, Learning and Research in Final Year Humanities Computing Student Projects." *Literary and Linguistic Computing* (due Summer 2005).

Kolb, D. *Experiential Learning: Experience at the Source of Learning and Development*. Kogan Page, 1984.

Marton, F., D. Hounsell, and N. Entwistle, eds. *The Experience of Learning*. 2nd ed. Edinburgh: Scottish Academic Press, 1997.

McCarty, Willard, and H. Short. *A Roadmap for Humanities Computing*. . Accessed 2005-02-15. <<http://www.kcl.ac.uk/cch/allc/reports/map>>

Moon, Jennifer A. *A Handbook of Reflective and Experiential Learning: Theory and Practice*. London: Routledge Falmer, 2004.

Prosser, M., and K. Trigwell. *Understanding Learning and Teaching*. Buckingham, UK: Open University Press, 1999.

Unsworth, J. *Scholarly Primitives: what methods do humanities researchers have in common, and how might our tools reflect this?* . Accessed 2005-15-02. <<http://jefferson.village.virginia.edu/~jmu2m/Kings.5-00/primitives.html>>

## Are Targeted User-Centred Interfaces the Key to Facilitating the Conversion of the Traditional Non-User to a User of Archives?

*Andrea Johnson* ([aj1@csmail.ucc.ie](mailto:aj1@csmail.ucc.ie))

*University College Cork*

Archival digitisation brings limitless opportunities, the advantages of which are obvious; however in the race to exploit this new media fundamental user-centred design principles have been ignored.

Based on body of research undertaken recently in the UK (Sabin et al.), the Internet is now viewed as a key point of entry to archival services. This poster examines the issues surrounding traditional non-user groups and the challenges that the designer faces when designing an interface to facilitate their needs.

The poster session will afford the opportunity of presenting the first in a series of prototype user-centred designed interfaces that aim to actively support the traditional non-user thereby facilitating the conversion process to that of an archive user. It will also examine the need for a cohesive national digitisation strategy that sustains targeted, well designed interfaces.

Based on a MORI poll (2003) nine out of ten adults in the UK aged fifteen plus are classified as non-archive users. This figure indicates that there is vast scope to identify and target groups within the non-user classification with the aim to persuade them to use an archive service, whether that be locally, nationally or via the Internet.

The MORI poll found that one in four people stated that making information available on the Internet was especially important to them with over a third of respondents stating that they are most likely to use the Internet to research their family tree within the next two years.

These figures are promising and suggest that the intellectual barrier that exists on accessing archive material seems to be dissipating. This is supported by research undertaken by the LEADERS project who found that the majority of users in their survey, 60%, could be categorised as 'personal leisure' users as opposed to 22% categorised as 'professional or occupational' users (Sexton et al.).

In response to this shift within the traditional user profile a small body among the archival profession has begun to address

the need for a user-centred approach when developing new ways to support and develop archival access by online users. LEADERS is one such project that has examined different types of users and their requirements in order to develop a set of open-source tools that can be used by archivists to create online content (Sexton et al.).

Wendy Duff has called for the establishment of a global research network of archival user studies. Her work and that of her colleagues Joan Cherry, Catherine Johnson and Barbara Craig has focused on the information seeking activities of the user. She believes that a detailed understanding of the information seeking behaviour of different types of users is a pre-requisite to providing the archival services of the future (Duff).

This growing interest in end-users and how they seek and use information is a departure from the traditional archivist's rationale. Wendy Duff offers an explanation as to why end-users have traditionally been marginalised in the design process. "At the heart of archival theory is the record, not its secondary use nor the various types of researchers who visit archives seeking information" (Duff).

This marginalisation of the user manifests itself in the inadequate consideration given to them and their specific needs in many of the digitisation best practise guidelines; the focus being strongly on technical standards and material issues.

Therefore consideration must now be given to all types of users in order to encourage the next generation of archive users. The challenge that remote users bring is the inherent difficulty in their identification and classification of use, this information is vital when producing comprehensive user requirements. Without a detailed understanding of the user and their requirements, the fulfilment of the later is unlikely.

Amanda Hill acknowledges this in her research on the characteristics of users of online archival resources.

Users of our online services are just as important as users who enter our record offices and, if we are to form a clear picture of the overall use of our services, we need to ensure that there are processes in place to count those users

(Hill)

In order to engage key groups amongst the traditional non-users it is essential to undertake a comprehensive evaluation of what has gone before. With the advent of large scale digitisation projects it has become possible to undertake summative evaluations of the user's experience of archival digitisation.

My own research into a digitisation project highlighted the pressures of a publicly funded scheme and the ensuing problems of designing for the widest possible audience. As a consequence the user was insufficiently defined which resulted in a poor understanding of the user's requirements (Johnson).

The current process for bidding for funds for digitisation compounds this approach. In order to attract new archive users and offer previously 'unavailable' access of archival material to the individual, the user must be placed at the centre of the design process. One significant influence in this pursuit of converting the non-user has to be a cohesive digitisation strategy, with local projects uniting under an agreed national strategy. This could generate a co-ordinated system where the user could enter at any point and navigate according to their individual needs.

The need for a targeted approach is the focus of this poster presentation, the focus of my own research is targeted user interfaces supporting all levels of users. These prototype interfaces encompass all that user-centred design can convey and aid in actively supporting the user in sharing in the vast opportunities that archival digitisation affords.

A multi-method research strategy which reflects real life has been utilised to produce a variety of data that provides a rich picture of the current problem. A number of groups selected from the non-user classification have been identified as suitable subjects for prototype users.

Using both summative evaluations of existing digitisation projects and analytic analysis of prototypes, the design process is both iterative and ensures the empirical measurement of prototype usage.

With an early focus on user requirements and tasks, the prototype interfaces are to undergo a rigorous testing programme. Understanding the behaviour of users when seeking information coupled with the amount and type of information requires further investigation with consideration given to the presentation and the interpretation of archive resources.

The interfaces presented at the session will include many features that the evaluation of current projects have shown to be lacking or poorly designed. The features include a single directory to help potential users identify what information may be useful of interest to them, simple navigation tools, a comprehensive help system and appropriate information of the archive resources available to the user both locally, nationally and via the Internet.

All prototype interfaces conform to ISO Standards 9241 and 13 407 and W3C guidelines.

This poster session marks the beginning of my research with the prototypes being the first in a series, each placing the user at the centre of the design process, which I believe is fundamental in securing new audiences for archive services via the Internet.

This targeted designed process hopes to overcome the restrictions that funding can often apply to these areas.

More research needs to be undertaken to examine at what point users stop using archival web resources and for what reason i.e. information overload, poor navigation etc.

Further research, in addition to what has already been produced by LEADERS (Sexton et al.), Duff and Hill is required into the types and categories of records users want to have digitised as most of the current projects have been driven by funding considerations.

"Review is vital" (MORI) therefore in an effort to ensure that archival digitisation delivers consequential accessibility, computer science has a key role in providing innovative solutions that actively sustain and promote the use by traditional non-user groups thereby encouraging new audiences to access archival material via this exciting media.

## Bibliography

Duff, W. "Understanding the information-seeking behaviour of archival researchers in a digital age: paths, processes and preferences." *Proceedings of the DLM Forum 2002*. Luxembourg, 2002. Accessed 2005-02-17. <[http://euro.pa.eu.int/historical\\_archives/dlm\\_forum/doc/dlm-proceed2002.pdf](http://euro.pa.eu.int/historical_archives/dlm_forum/doc/dlm-proceed2002.pdf)>

Hill, A. "Serving the invisible researcher: meeting the needs of online users." *Journal of the Society of Archivists* 25.2 (2004).

Johnson, A.C. *"The Mersey Gateway project: How was it for you?" A User Centred Evaluation of a Digitisation Project*. Dissertation, University of Lancaster, 2004, 1966.

MORI. *Listening to the Past, Speaking to the Future. Annex D: Non-Archive Users Survey :Omnibus Study*. MLA Publications, 2003. Accessed 2004-11-03. <[http://www.mla.gov.uk/documents/atf\\_annex\\_d.pdf](http://www.mla.gov.uk/documents/atf_annex_d.pdf)>

Sabin, R.W., and L. Samuels. *Listening to the Past, Speaking to the Future. Annex E: Towards a better Understanding of Non-Users*. MLA Publications, 2003. Accessed 2004-11-03. <[http://www.mla.gov.uk/documents/atf\\_annex\\_e.pdf](http://www.mla.gov.uk/documents/atf_annex_e.pdf)>

Sexton, A., C. Turner, G. Yeo, and S. Hockey. "Understanding Users: a prerequisite for developing new technologies." *Journal of the Society of Archivists* 25.1 (2004).

## Towards an Automatic Index Generation Tool

*Patrick Juola (juola@mathcs.duq.edu)*  
*Duquesne University*

**A**lmost every non-fiction author has been faced from time to time with the generation of an index. Most novice authors (myself included) are taken aback by the magnitude of the task and the limited amount of computational and software support available.

The current state of the art is significantly improved from the days of 3-by-5 'index cards', (a telling term?), but only in mechanical, not intellectual terms. Modern publishing practice typically involves the author delivering a machine-readable 'manuscript', written in a document-processing system such as *LaTeX*. Index entries are defined as specific term/location pairs by the author. For example, an index entry written in *LaTeX*, might look as follows

```
The \index{Pittsburgh!University of}
University of Pittsburgh was established
in \index{Pittsburgh!city of}
Pittsburgh, Pennsylvania, in the
year....
```

This will create an index entry on the 'current page', under the heading "Pittsburgh, University of" (as opposed to "Pittsburgh, city of," which would be the second entry, a related but separate subentry). Although guidelines for a good index (Northrup; University of Chicago Press Staff) are commonly available, the process of producing a good index is still largely unsupported, even by major and relatively sophisticated publishing companies such as Prentice-Hall.

What differentiates an index from a mere concordance?

There are at least six cognitive tasks (Maislin; Saranchuk) related to the production of a good index, as follows. Current standard support covers only the last.

- Identification of terms to index;
- Location of all informative references in the text;
- Identification/location of synonymous terms (e.g. "University of Pittsburgh" / "Pitt" );
- Splitting of index terms to split into subterms;
- Development of cross-references within the index itself;
- Compilation of page numbers,

I will present a framework for the development of a 'machine-aided index generation system'. This bears the same relationship to an automatic indexer that machine-aided translation (MAT) does to machine translation (MT), in that it provides suggestions and reduces the overall workload for the human, but post-editing will still be necessary. Specifically, recent results in corpus linguistics (Charniak; Manning & Schuetze), including the development of taggers for part of speech (Cutting et al.; Schmid) the availability of ontologies and semantics networks, plus the light semantic analysis capabilities of latent semantic analysis (Landauer et al.), can be combined in a multi-phased iterative framework and implemented as user-level software. This paper presents some aspects of "good" indices (Northrup; University of Chicago Press Staff) and illustrates how they can be achieved computationally.

In general, following the University of Chicago's dictum that "it is always easier to drop entries than to add them, err on the side of inclusiveness," (rule 18.120) we start by assuming that every term is a potential index entry and look for criteria by which to eliminate enough terms to produce a reasonably-sized index. (5-15 references/page, between 2% and 5% of the length of the final work, according to rule 18.120.) For example, rule 18.8 states that "the main heading of an index entry is normally a noun or noun phrase---the name of a person, a place, and object, or an abstraction." A first pass, then, can use the results of a part-of-speech tagger and eliminate all terms that do not appear as a noun in the document. Within this set of nouns, I suggest two possible heuristics for further pruning; first, common nouns that are too common or too rare are unlikely to be useful index terms, and second, words that are too uniformly distributed are unlikely to be useful index terms. On the other hand, a case can be made that all proper nouns should be included. Other suggested heuristic will be discussed.

Within a single index term, "an entry that requires more than five or six locators is usually broken up into subentries" (rule 18.9). This can be treated as an example of word-sense disambiguation, for example, between *Pittsburgh* (University of) and *Pittsburgh* (city of). Again, I conjecture (and present supporting evidence) that existing technology can provide a useful and helpful basis for later human editing. Specifically, existing semantic representation techniques can model the context, and therefore the meaning, of each index token. For truly polysemous terms, cluster analysis of the set of token representations should yield a set of clusters equivalent to the degree of polysemy; by setting the separation threshold to an appropriate level, the analysis can be forced to produce clusters of maximum size at most 5-6. At the same time, passing and uninformative references can be expected to produce isolated 'clusters' containing a single outlier — a strong candidate for omission. Once a list of index terms is collected, tokens not on that list can be compared in their semantic representation for

similarity with existing index terms; any word with near-identical meaning is a potential synonym and a candidate for a cross-reference.

Unfortunately, the evidence to be presented is largely heuristic and exploratory in nature. We are currently developing a prototype system, using LSA (Landauer et al.) and elementary corpus statistics such as TF-IDF to identify index terms. We also have a well-developed and intuitive GUI wizard for ease of use by a non-technical user. At present, the planned heuristics may or may not be sufficiently reliable to use without a human post-editor. However, if they can be shown to substantially reduce the work load on the human author, the resulting tool may still be of interest. I present the results of some prototype-scale experiments, plus some ideas about usability testing and the directions of future development.

## Bibliography

Charniak, E. *Statistical Language Learning*. Cambridge, MA: MIT Press, 1993.

Cutting, D., J. Kupiec, J. Pedersen., and P. Sibun. "A practical part-of-speech tagger." *Proceedings of the Third Conference on Applied Natural Language Processing*. Trento, Italy, 1992. 42-46. Association for Computational Linguistics. Also available as Xerox PARC technical report SSL-92-01.

Landauer, T., P. Foltz, and D. Laham. "Introduction to latent semantic analysis." *Discourse Processes* 25 (1998): 259-284.

Maislin, S. "The cognitive half of indexing." *Proceedings of Massachusetts Society of Indexers Fall Conference*. Massachusetts Society of Indexers, 1996. n. pag. Association for Computational Linguistics. Also available as Xerox PARC technical report SSL-92-01.

Manning, C., and H. Schuetze. *Foundations of Statistical Natural Language Processing*. Cambridge, MA: MIT Press, 1999.

Northrup, M.J. "The role of indexing in technical communication." *Proceedings of SIGDOC-90*. Association for Computing Machinery, 1990. n. pag.

Saranchuk, G.R.Z. *A new index for the graduate program manual of the Faculty of Graduate Studies and Research at the University of Alberta*. University of Alberta, 1996. Accessed 2005-04-11. <[http://www.slis.ualberta.ca/cap03/georgina/lis600fgsr\\_introduction.htm](http://www.slis.ualberta.ca/cap03/georgina/lis600fgsr_introduction.htm)>

Schmid, H. "Part-of-speech tagging with neural networks." *Proceedings of SIGDOC-90*. Proceedings of COLING-94, 1994. n. pag.

---

University of Chicago Press Staff. *The Chicago Manual of Style*. 15th ed. Chicago: University of Chicago Press, 2003.

## A Prototype for Authorship Attribution Software

---

*Patrick Juola* ([juola@mathcs.duq.edu](mailto:juola@mathcs.duq.edu))  
*Duquesne University*

---

**T**he task of computationally inferring the author of a document based on its internal statistics — sometimes called *sylometrics*, *authorship attribution*, or (for the completists) *non-traditional authorship attribution* is an active and vibrant research area, but at present largely without use. For example, unearthing the author of the anonymously-written *Primary Colors* became a substantial issue in 1996. In 2004, *anonymous* published *Imperial Hubris*, a followup to his (her?) earlier work *Through Our Enemies' Eyes*. Who wrote these books? Does the author actually have the expertise claimed on the dust cover ('a senior U.S. intelligence official with nearly two decades of experience')? And, why haven't our computers already given us the answer?

Part of this lack of use can be attributed to simple unfamiliarity on the part of the relevant communities, combined with a perceived history of inaccuracy<sup>1</sup>. Since 1996, however, the popularity of corpus linguistics as a field of study and vast increase in the amount of data available on the Web (Nerbonne) have made it practical to use much larger sets of data for inference. During the same period, new and increasingly sophisticated techniques have improved the quality (and accuracy) of judgements the computers make.

As a recent example, in June 2004, *ALLC/ACH* hosted an *Ad-hoc Authorship Attribution Competition* (Juola 2004b). Specifically, by providing a standardized test corpus for authorship attribution, not only could the mere ability of statistical methods to determine authors be demonstrated, but methods could further be distinguished between the merely 'successful' and 'very successful', and analyzed in particular into possible areas of individual success.

The contest (and results) were surprising at many levels; some researchers initially refused to participate given the admittedly difficult tasks included among the corpora. For example, Problem F consisted of a set of letters extracted from the Paston letters. Aside from the very real issue of applying methods designed/tested for the most part for modern English on documents in Middle English, the size of these documents (very few letters, today or in centuries past, exceed 1000 words) makes statistical inference difficult. Similarly, problem A was a realistic exercise in the analysis of student essays (gathered

in a freshman writing class during the fall of 2003) — as is typical, no essay exceeded 1200 words. Despite this extreme paucity of data, results could be stunningly accurate. The highest scoring participant was the research group of Vlado Keselj, with an average success rate of approximately 69%. (Juola's solutions, in the interests of fairness, averaged 65% correct.) In particular, Keselj's methods achieved 85% accuracy on problem A and 90% accuracy on problem F, both acknowledged to be difficult and considered by many to be unsolvably so.

However, the increased accuracy has come at the price of decreased clarity; the statistics used<sup>2</sup> can be hard to understand, and perhaps more importantly, difficult to implement or to use by a non-technical scholar. At the same time, the sheer number of techniques proposed (and therefore, the number of possibilities available to confuse) has exploded. This limits the pool of available users, making it less likely that a casual scholar — let alone a journalist, lawyer, or interested layman — would be able to apply these new methods to a problem of real interest.

I present here a prototype and framework for a user-friendly software system (Juola & Sofko) allowing the casual user to apply authorship attribution technologies to her own purposes. It combines a generalized theoretical model (Juola, 2004b) built on an inference task over *event* sequences with an extensible, object-oriented inference engine that makes the system easily updatable to incorporate new technologies or to mix-and-match combinations of existing ones. The model treats linguistic (or paralinguistic) data as a sequence of separable user-defined *events*, for instance, as a sequence of letters, phonemes, morphemes, or words. These sequences are treated to a three-phase process:

- **Canonicization** — No two physical realizations of events will ever be exactly identical. We choose to treat similar realizations as identical to restrict the event space to a finite set.
- **Determination of the event set** — The input stream is partitioned into individual non-overlapping *events*. At the same time, uninformative events can be eliminated from the event stream.
- **Statistical inference** — The remaining events can be subjected to a variety of inferential statistics, ranging from simple analysis of event distributions through complex pattern-based analysis. The results of this inference determine the results (and confidence) in the final report.

As an illustration, the implementation of these phases for the Burrows method would involve, first, *canonicization* by norming the documents of interest. For example, words with variant capitalization (*the*, *The*, *THE*) would be treated as a single type. More sophisticated canonicization procedures could regularize spelling, eliminate extraneous material such as

chapter headings, or even "de-edit" (Rudman) the invisible hand of the editor. During the second phase, the appropriate set of function words would be determined and presented as a sequence of events, eliminating words not in the set of interest. Finally, the appropriate function words are tabulated (without regard to ordering) and the appropriate inferential statistics (principle component analysis) performed. However, replacement of the third stage (and only the third stage) by a linear discriminant analysis would produce a different technique (Baayen et al.).

This framework fits well into the now-standard modular software design paradigm. In particular, the software to be demonstrated uses the Java programming language and object-oriented design to separate the generic functions of the three phases as individual classes, to be implemented as individual subclasses.

The user can select from a variety of options at each phase, and the system as a whole is easily extensible to allow for new developments. For example, the result of event processing is simply a Vector (Java class) of events. Similarly, similarity judgement is a function of the Processor class, which can be instantiated in a variety of different ways. At present, the Processor class is defined with a number of different methods<sup>3</sup>. A planned improvement is to simply define a `calculateDistance()` function as part of the Processor class. The Processor class, in turn, can be subclassed into various types, each of which calculates distance in a slightly different way.

Similarly, preprocessing can be handled by separate instantiations and subclasses. Even data input and output can be modularized and separated. As written, the program only reads files from a local disk, but a relatively easy modification would allow files to be read from a local disk or from the network (for instance, Web pages from a site such as *Project Gutenberg* or *literature.org*). Users can therefore select functionality as needed on a module-by-module basis both in terms of feature as well as inference method; the current system incorporates four different approaches (Burrows; Juola 1997; Kukushkina et al.; Juola 2003).

From a broader perspective, this program provides a uniform framework under which competing theories of authorship attribution can both be compared and combined (to their hopefully mutual benefit). It also forms the basis of a simple user-friendly tool to allow users without special training to apply technologies for authorship attribution and to take advantage of new developments and methods as they become available. From a standpoint of practical epistemology, the existence of this tool should provide a starting point for improving the quality of authorship attribution as a forensic examination — by allowing the widespread use of the technology, and at the same time providing an easy method for



testing and evaluating different approaches to determine the necessary empirical validation and limitations.

On the other hand, this tool is also clearly a *research-quality* prototype, and additional work will be needed to implement a wide variety of methods, to determine and implement additional features, to establish a sufficiently user-friendly interface. Even questions such as the preferred method of output — dendrograms? MDS subspace projections? Fixed attribution assignments as in the present system? — are in theory open to discussion and revision. It is hoped that the input of research and user such as the present meeting will help guide this development.

- 
1. See, for example, the discussion of the cusum technique (Farrington) in (Holmes 1998).
  2. E.g. linear discriminant analysis of common function words (Burrows, Baayen et al; Juola & Baayen), orthographic cross-entropy (Juola, 1996), common byte N-grams (Keselj, 2004).
  3. For example, `crossEntDistance()` and `LZWDistance()`.

## Bibliography

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

Baayen, R. H., H. Van Halteren, A. Neijt, and F. Tweedie. "An experiment in authorship attribution." *Proceedings of JADT 2002*. St. Malo, 2002. 29-37.

Burrows, J. "'An Ocean where each Kind. . . ': Statistical analysis and some major determinants of literary style." *Computers and the Humanities* 23.4-5 (1989): 309-21.

Burrows, J. "Questions of authorship : Attribution and beyond." *Computers and the Humanities* 37.1 (2003): 5-32.

Farrington, J.M. *Analyzing for Authorship: A Guide to the Cusum Technique*. Cardiff: University of Wales Press, 1996.

Holmes, D. I. "Authorship attribution." *Computers and the Humanities* 28.2 (1994): 87-106.

Holmes, D. I. "The evolution of stylometry in humanities computing." *Literary and Linguistic Computing* 13.3 (1998): 111-7.

Juola, P. "What can we do with small corpora? Document categorization via cross-entropy." *Proceedings of an Interdisciplinary Workshop on Similarity and Categorization*. Edinburgh, UK: Department of Artificial Intelligence, University of Edinburgh, 1997. n. pag.

Juola, P. "The time course of language change." *Computers and the Humanities* 37.1 (2003): 77-96.

Juola, P. "Ad-hoc authorship attribution competition." *Proceedings of the 2004 Joint International Conference of the Association for Literary and Linguistic Computing and the Association for Computers and the Humanities (ALLC/ACH 2004)*. Goteborg, Sweden, 2004a. 175-176.

Juola, P. "On composership attribution." *Proceedings of the 2004 Joint International Conference of the Association for Literary and Linguistic Computing and the Association for Computers and the Humanities (ALLC/ACH 2004)*. Goteborg, Sweden, 2004b. 79-80.

Juola, P., and H. Baayen. "A controlled-corpus experiment in authorship attribution by cross-entropy." *Proceedings of ACH/ALLC-2003*. Athens, GA, 2003. n. pag.

Juola, P., and J. Sofko. "Proving and Improving Authorship Attribution Technologies." *Proceedings of CaSTA-2004*. Hamilton, ON, 2004. n. pag.

Keselj, V., and N. Cercone. "CNG Method with Weighted Voting." *Ad-hoc Authorship Attribution Contest Technical Report*.

Kucera, H., and W.N. Francis. *Computational Analysis of Present-day American English*. Providence: Brown University Press, 1967.

Kukushkina, O.V., A.A. Polikarpov, and D.V. Khmelev. "Using literal and grammatical statistics for authorship attribution." *Problemy Peredachi Informatii* 37.2 (2000): 172-184. Translated in *Problems of Information Transmission*.

Nerbonne, J. "The data deluge." *Literary and Linguistic Computing* (Forthcoming). [In Proceedings of the 2004 Joint International Conference of the Association for Literary and Linguistic Computing and the Association for Computers and the Humanities (ALLC/ACH 2004).]

## Social, Geographical, and Register Variation in Dutch: From Written MOGELIJK to Spoken MOK

---

**Karen Keune** (*karen.keune@mpi.nl*)

University of Nijmegen

**Mirjam Ernestus** (*mirjam.ernestus@mpi.nl*)

Max Planck Institute for Psycholinguistics

**Roeland van Hout** (*r.v.hout@let.ru.nl*)

University of Nijmegen

**Harald Baayen** (*baayen@mpi.nl*)

University of Nijmegen

---

In spontaneous speech words are often pronounced in reduced form. Some words are reduced to such an extent that an orthographic transcription would be very different from the orthographic norm. An example from Dutch is the word *MOGE-LIJK* ('possible'), which can be pronounced not only as *MO.GE.LEK* but also as *MO.GEK*, *MO.LEK*, or even as *MOK*.

Strongly reduced word forms are difficult to interpret without syntactic or semantic context. When speakers of Dutch are presented with the word *mok* in isolation, they tend not to be able to assign a meaning to this string of phonemes. It is only when the word is embedded in a sentence that its meaning becomes available. Interestingly, listeners who understood the meaning of *MOK* tend to think they heard the full, unreduced form *MOGELIJK*. A central question in the research on the comprehension of reduced words is what aspects of the linguistic context allow the listener to access the associated semantics (Kemps et al.).

An important predictor for the degree of reduction in speech production is lexical frequency, as demonstrated by Jurafsky et al. for function words. The more often a function word is used in speech, the more likely it is to undergo reduction, in line with Zipf's law of abbreviation. Furthermore, the degree of reduction is modulated by the extent to which a word is predictable from its context. In addition, frequency of occurrence has been shown to affect the realization of word final dental plosives in monomorphemic words (Bybee), and a negative correlation between frequency and acoustic length

has been observed for several kinds of derived words in Dutch, including words with the suffix *-LIJK* (Pluymaekers et al.).

It is an open question to what extent the use of reduced forms is co-determined by social, geographical, and stylistic factors. Various corpus-based studies have shed light on variation in the use of language. Biber identified different varieties of English (and also other languages) by means of factor analyses of the frequencies of a broad range of morphological and syntactic variables (Biber). In the domain of literary studies, Burrows demonstrated regional differences in English narrative, diachronic change in literary texts, and even sex-specific differences in the writing of English historians born before 1850 (see, e.g. Burrows). Studies in authorship attribution revealed, furthermore, that differences in speech habits can sometimes be traced down even to the level of individual language users (Holmes; Baayen et al.). Finally, it has been shown that derivational affixes are used to a different extent in spoken and written registers (Baayen; Plag et al.).

The aim of the present study is to investigate the extent to which the use of words in *-LIJK* varies systematically as a function of speech register, the speaker's sex, level of education, and of whether the speaker lives in Flanders or in the Netherlands. For spoken Dutch, we address the more specific question to what degree these factors (and contextual predictability) codetermine the extent to which words in *-LIJK* are reduced.

We first studied the social and geographic variation in the frequency of use of words in *-LIJK* in a corpus of Dutch newspapers. We selected all occurrences of 80 high-frequency words in *-LIJK* from seven newspapers using a 2 by 3 factorial design. We distinguished between Flemish and Dutch newspapers (Country) and contrasted quality newspapers, national newspapers, and regional newspapers (Register). In parallel, we conducted a study using the same design based on the 80 most frequent function words (pronouns, auxiliaries, connectives, determiners, numerals, etc.), following Burrows. In both analyses, we observed significant and remarkably similar regional and stylistic differentiation. This suggests that the syntactic habits of journalists (as revealed by their use of function words) are consistent with their habits with respect to the use of adverbs and adjectives in *-LIJK*.

Next, we explored the variation in frequency of use of words in *-LIJK* in spoken Dutch. We selected 32 high-frequency words in *-LIJK* from the subcorpora of spontaneous, unscripted speech in the Corpus of Spoken Dutch (CGN), using a 2x2x2 factorial design in which we contrasted speakers from Flanders with speakers from the Netherlands (Country), men with women (Sex), and highly educated with less educated speakers (Education). As before, we carried out a parallel study using the most frequent function words. This time, we observed a marked difference between the function words and the words in *-LIJK*. Speakers with a higher education level tended to use

words in *-LIJK* more often. For the Netherlands (but not for Flanders), this mirrors the finding that the quality newspaper made more intensive use of this suffix as well. The analysis of the function words, by contrast, revealed that men made less use of function words compared to women, suggesting a slightly higher information density (carried by content words) for men. In addition to these main effects, we observed marked (and significant) differences in how individual function words as well as individual words in *-LIJK* were used by men and women in the two countries as a function of their education level.

Finally, we investigated the social and regional variation in the degrees of reduction of words in *-LIJK* for 14 words that occurred sufficiently often in the different subcorpora of the CGN defined by our factorial design contrasting Country, Sex and Education, and that revealed substantial degrees of reduction. Two transcribers classified the degree of reduction for a total of 946 tokens. We considered two kinds of reduction, one primarily affecting the suffix, the other affecting the vowel in the word initial syllable. Both analyses show that in Flanders speakers reduce less than in the Netherlands. The reduction involving the suffix is more prominent for men compared to women. Moreover, highly educated Flemish speakers use fewer reduced forms than do less highly educated Flemish speakers. Finally, there were significant differences in the extent to which the individual words underwent reduction that we could trace back to the speaker's region.

In addition to these social and regional factors, the degree of reduction was significantly co-determined by two linguistic factors: the word's position in the sentence, and the extent to which the word is predictable from its context. We used the Mutual Information measure to gauge contextual predictivity. Words in *-LIJK* with a high mutual information, i.e., words that exhibited a high degree of predictability from the preceding word, revealed more reduction: As the information load of a word in *-LIJK* decreases, its formal distinctiveness in production decreases as well. In this respect, highly-reduced and semantically opaque forms in *-LIJK* such as *TUUK* (for *NATUURLIJK*, 'of course') and *EIK* (for *EIGENLIJK*, 'in fact') are becoming similar to function words. With respect to the word's position in the sentence, we found that words in *-LIJK* that occurred in sentence-final position revealed little reduction. This is as expected given that words in sentence final position are often lengthened.

For our analyses, we made extensive use of multilevel modeling of covariance, a statistical technique that offers two advantages compared to principal components analysis, factor analysis, and correspondence analysis. First of all, multilevel modeling allows the researcher to directly assess the significance of the predictors in the model, as well as how the individual words interact with these predictors. In other words, instead of using both a clustering technique such as principal components

analysis and a technique for group separation such as discriminant analysis, we were able to fit a single statistical model to the data that allows us both to trace what predictors are significant, and to visualize their effects. The second advantage of multilevel modeling is that it offers the researcher the possibility to include covariates such as mutual information in the model.

Although derived words are generally classified as open-class words, as opposed to the closed class function words, it is noteworthy that the suffix *-LIJK* is hardly productive. Furthermore, we have shown that if the information load of a word in *-LIJK* decreases, its formal distinctiveness in production decreases as well. Thus, high-frequency forms in *-LIJK* are becoming more similar to function words with respect to their lack of productivity and compositionality, with respect to their being social and stylistic markers, and with respect to their acoustic form.

## Bibliography

- Baayen, R.H., H. Van Halteren, and F. Tweedie. "Outside the cave of shadows: Using syntactic annotation to enhance authorship attribution." *Literary and Linguistic Computing* 11 (1996): 121-131.
- Baayen, R.H. "Derivational productivity and text typology." *Journal of Quantitative Linguistics* 1 (1994): 16-34.
- Biber, D. *Dimensions of register variation*. Cambridge: Cambridge University Press, 1995.
- Burrows, J.F. "Computers and the study of literature." *Computers and Written Texts*. Ed. C.S. Butler. Oxford: Blackwell, 1992. 167-204.
- Bybee, J.L. *Phonology and language use*. Cambridge: Cambridge University Press, 2001.
- Holmes, D.I. "Authorship attribution." *Computers and the Humanities* 28.2 (1994): 87-106.
- Jurafsky, D., A. Bell, M. Gregory, and W.D. Raymond. "Probabilistic relations between words: Evidence from reduction in lexical production." *Frequency and the emergence of linguistic structure*. Ed. J.L. Bybee and P. Hopper. Amsterdam: John Benjamins, 2001. 229-254.
- Kemps, R., M. Ernestus, R. Schreuder, and R.H. Baayen. "Processing reduced word forms: The suffix restoration effect." *Brain and Language* 90 (2004): 117-127.
- Plag, I., C. Dalton-Puffer, and R.H. Baayen. "Productivity and register." *Journal of English Language and Linguistics* 3 (1999): 209-288.

Pluymaekers, M., M. Ernestus, and R.H. Baayen. *Lexical frequency and acoustic reduction in spoken Dutch*. In preparation.

## **The *Edition Production Technology (EPT)* and the *ARCHway* and *Electronic Boethius* Projects**

---

**Kevin Kiernan** ([kiernan@uky.edu](mailto:kiernan@uky.edu))

*University of Kentucky, English*

**Dorothy Porter** ([dporter@uky.edu](mailto:dporter@uky.edu))

*University of Kentucky, Research in Computing for Humanities*

**Alex Dekhtyar** ([dekhtyar@cs.uky.edu](mailto:dekhtyar@cs.uky.edu))

*University of Kentucky, Computer Science*

**Ionut Emil Iacob** ([ionut@ms.uky.edu](mailto:ionut@ms.uky.edu))

*University of Kentucky, Computer Science*

**Jerzy W. Jaromczyk** ([jurek@cs.uky.edu](mailto:jurek@cs.uky.edu))

*University of Kentucky, Computer Science*

**Neil Moore** ([neil@s-z.org](mailto:neil@s-z.org))

*University of Kentucky, Computer Science*

---

### **Session Statement**

**T**his session is based on the collaborative research and interdisciplinary education that have gone into the development of a generic *Edition Production Technology (EPT)* for building image-based electronic editions of damaged Old English manuscripts and, by extension, any representation of digitized cultural materials for contemporary users.

The concept of an effective, modular, extensible, *Edition Production Toolkit (EPT)* arose from the difficulties encountered while producing an electronic edition of the Beowulf manuscript. To solve these problems for the *Electronic Boethius* project, we set out to create a modular Java and XML software framework, including an edition production management system, a native XML database, graphical user interfaces, and a suite of editorial tools, customized to the needs of textual scholars in the humanities. The goal was to allow for the efficient assembly of complex scholarly editions from high-resolution digital facsimiles and XML-encoded texts, apparatus and ancillary materials. While our general approach was sound and the *Electronic Boethius* project in a few months

developed a suite of Java editorial tools operating under an XML framework, and successfully used these tools to begin the edition, the extensible development of the *Electronic Boethius* toolkit was hampered by the lack of computer science expertise in software engineering.

We were accordingly fortunate to attract computer scientists to join in the *ARCHway Project*, which deeply involved them and their students in an interdisciplinary effort to create an overarching technology for image-based electronic editions. Guided by the *Eclipse* programming environment, *ARCHway* has established an infrastructure for collaborative research and teaching between computer science and the humanities. Our interdisciplinary teams, working together at each stage, have designed formal methodologies for collaborative teaching and research, based on practical goals. *Eclipse*, our chosen programming environment, maintains an open-standards architecture with modular, extensible, interoperable components to coordinate research and development of novel methods, tools, and associated technologies in a teaching and learning environment involving undergraduate and graduate students. *EPT* has guided the definition and coordination of well-encapsulated collaborative student projects from semester to semester in specified research projects related to documenting, editing, storing, accessing, and searching image-based electronic editions.

The complementary projects allowed our research teams to pursue these shared goals from both a specific and practical standpoint, with the *Electronic Boethius* and the *Electronic Beowulf* projects, to a more general and theoretical standpoint, with the *ARCHway Project*. In the following papers, we first present how specific problems of preparing an electronic edition from damaged Old English manuscripts help to define the range of tools required in the *EPT*; we then show how the computer scientists designed and implemented an *EPT* with specific components crucial to image-based, document-centric editing; and finally, we present the *EPT*'s architecture, centering on the utilities that constitute its underlying infrastructure.

## Using EPT to Build an Image-Based Electronic Edition of Alfred's Boethius

**Kevin Kiernan and Dorothy Carr Porter**

*EPT* is well suited to build an Image-Based Electronic Edition (IBEE) of Alfred the Great's Old English version of Boethius's *Consolation of Philosophy*. There are two surviving Old English manuscripts of this text, but they present the text in very different ways. The first, the tenth-century BL MS Cotton Otho A. vi, is the only prose and verse translation. The other complete manuscript is a later, twelfth-century, entirely prose version in Oxford, Bodleian Library MS, Bodley 180. There is also an indispensable, post-medieval source, however, in a seventeenth-century transcript and collation of the two

manuscripts by Francis Junius, an edition in the making now preserved in Oxford, Bodleian Library MS, Junius 12. In 1731 the earlier Cottonian MS was badly burned in the terrible Cotton Library fire, but ultraviolet discloses much of the seemingly lost text and Junius's transcripts and collations preserve most of the rest, while Bodley 180 provides critical variants. No modern editions have taken full advantage of these rich and diverse materials, and the two "standard" editions, supposedly based on Otho A. vi, respectively present a prose edition, stripped of the verse, and a verse edition, stripped of the prose. To provide a base text for XML encoding, the editor compiled a reconstructed version of Otho A. vi by reinserting the verse sections where they belong. At this point it is ready for *EPT*.

The purpose of an image-based edition is to reveal as openly and fully as possible the primary sources underlying the modern edition. Traditional print editions tend to conceal these sources by radically reformatting their structures, by providing modern punctuation, by underplaying their damaged states, by erasing their scribal peculiarities and semantic cruces through printed emendations and conjectural restorations, and by generally relegating the complex evidence the manuscripts hold to concise, uncomplicated, textual notes. An image-based edition makes these concessions, as well, to render an alien text accessible to today's readers, but it also provides ready access to the ultimate sources by linking, for example, all textual notes to the manuscript context.

In this presentation we will illustrate how the *EPT* provides the means for describing the manuscript using XML markup, and relating the folio and areas of the folio to the text that resides on that folio. The *EPT* enables us to associate images and sections of images with the relevant markup (folios with folio markup, damaged areas with damage markup, letters with markup describing the letter form, etc.), while at the same time associating text with whole or portions of single images, or multiple images of the same folio taken under different lighting conditions (daylight, ultraviolet, fiberoptic).

By providing support for pervasive, complex, image-based encoding, the *EPT* inevitably exacerbates the problem of overlapping markup. Jacob and Dekhtyar address the *EPT*'s approach to this problem in greater detail in this session; here it is enough to say that the *EPT* does support overlapping markup using multiple DTDs, and that if the DTDs are well-designed the humanities scholar need not worry about the semantics of XML markup, and can concentrate instead on the main tasks of editing, such as the semantics of an Old English text.

The *EPT* includes three fundamental tools for image-based encoding. The *ImagText* tool, working in cooperation with *xMarkup* and *xTagger* tools, supports general image-based encoding, allowing the editor to tag any element defined in the DTDs. *xMarkup* reads the DTDs and automatically creates

templates that the editor configures, assigning meaningful or otherwise convenient aliases for elements, attributes, and attribute values, and arranging the elements into logical editorial groups, such as Start Edition, Condition, Codicology, Paleography, Restoration, with all their subsets of tags, which will very likely differ from the organization of the DTDs. Through *ImagText*, the editor views images side-by-side with the corresponding text (viewed through *xTagger*), and describes them with reference to one another using the templates provided by xMarkup. Otho A. vi has suffered severe damage, both from fire and from later preservation attempts. Thus, much of the tagging relates to manuscript condition, highlighting where the manuscript is damaged and linking it with the transcript or edition. This information ensures that the final edition will show clearly what text in the edition comes directly from the manuscript, what text is slightly damaged, and what text is damaged to the point of illegibility and thus either copied from another manuscript or otherwise restored by the editor.

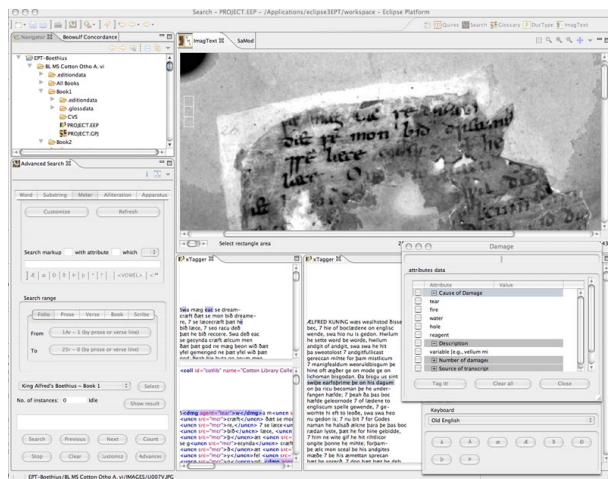


Figure 1. A snapshot of EPT illustrating image-based encoding through *ImagText*, *xMarkup*, and *xTagger* (including an XML view). The figure also shows the Keyboard panel and Search Tool.

The OverLay allows the editor to encode the text in reference to multiple images of the same folio. Images taken under special conditions, such as ultraviolet fluorescence, often provide clearer textual evidence than those digitized in natural light. Through OverLay an editor can examine minute differences between digital images, laying one image on top of another, selecting a section of the image, and using a slider to change the transparency of the top layer, moving between them. The editor can save the combined images created by OverLay and open them in *ImagText*, marking them up there as they would any other image.

The DucType is an example of a markup template that has been configured to deal with the specific encoding problem of describing letter forms. Paleography, the study of handwritten texts, is central to the scholarship of medieval manuscripts. Description of letterforms and the style of individual scribes

have traditionally been limited to general descriptions in manuscript catalog entries, or in introductions to manuscript facsimiles. Using XML, we are now able to incorporate paleographical description into the edition content character-by-character, providing individual letters with their own descriptions. This advanced markup will enable users of the finished edition to search for letters based on specific characteristics.

Specialized templates require specialized configuration. The editor configures the Letter template using the Letter Template tool, through which he assigns meaningful aliases to element and attribute names, adding attribute values as he discovers new letterforms in the manuscript. The Letter Template tool also enables the editor to clip and save sample letters, which the editor can reference later, comparing them to other letterforms in the manuscript.

Although the *EPT* enables the editor to practice image-based electronic editing, it is the DTDs that provide the underlying structure for describing the manuscript. We design our DTDs as extensions of TEI, defining some new elements and adding new attributes to existing TEI elements. We will illustrate how we are using TEI for image-based encoding, and how our extensions allow for a more complete manuscript description than TEI alone. We will discuss how we adapt TEI elements for image based encoding, and we will also describe our new attributes, which assist the *EPT* in supporting links between text and image. We will also introduce some of our new elements, including `<offset>`, an empty element which marks an area in the manuscript where text from the facing page has bled onto the folio, in cases obscuring the manuscript text, and `<offsettext>`, which marks the text on the facing page corresponding to the offset. The `<offset>` and `<offsettext>` regions can then be compared using OverLay. We will also discuss our markup for paleographical description and the restoration of text visible under special lighting, not visible under regular lighting.

The *Electronic Boethius* Project is funded by a Collaborative Research Award from the National Endowment for the Humanities and the Andrew W. Mellon Foundation, and is sponsored by The British Library and the Bodleian Library, Oxford, who are providing digital images of the relevant documents. We are working in collaboration with the complementary print-based *Alfredian Boethius* project at Oxford, directed by Malcolm Godden.

## Bibliography

### Primary Sources

British Library MS Cotton Otho A. vi.

Oxford Bodleian Library MS Bodley 180.

Oxford Bodleian Library MS Junius 12.

### Editions

Krapp, George Philip, ed. *The Paris Psalter and the Meters of Boethius*. The Anglo-Saxon Poetic Records 5. New York: Columbia University Press, 1932.

Robinson, Fred C., and E.G. Stanley, eds. *Old English Verse Texts from Many Sources: A Comprehensive Collection*. Early English Manuscripts in Facsimile 23. Copenhagen, Denmark: Rosenkilde and Bagger, 1991.

Sedgefield, Walter J., and E.G. Stanley, eds. *King Alfred's Anglo-Saxon Version of Boethius, de Consolatione Philosophiae*. Oxford: Clarendon Press, 1899.

### Secondary Sources

Bauman, Syd, and Terry Catapano. "TEI and the Encoding of the Physical Structure of Books." *Computers and the Humanities* 33 (1999): 113-127.

Clark, James, ed. *XSL Transformations (XSLT) 1.0*. W3C Recommendation, 16 November 1999. Accessed 2005-04-07. <<http://www.w3.org/TR/xslt>>

Johansson, Karl Gunnar. "Computing Medieval Primary Sources from the Vadstena Monastery: Arguments for the Primary Source Text." *Literary and Linguistic Computing* 19.1 (2004): 93-104.

Kiernan, Kevin. "Image-based Electronic Editing of Alfred the Great's Boethius." *Making Sense: Constructing Meaning in Early English*. Ed. Antonette diPaolo Healey and Kevin Kiernan. Forthcoming. (In progress, expected Richard Rawlinson Center Series, Medieval Institute Press, 2006.)

Kiernan, Kevin, Jerzy W. Jaromczyk, Alex Dekhtyar, Dorothy Carr Porter, Kenneth Hawley, Sandeep Bodapati, and Ionut Emil Iacob. "The ARCHway Project: Architecture for Research in Computing for Humanities through Research, Teaching, and Learning." *Literary and Linguistic Computing* (Forthcoming in 2005). (Special issue, papers from Joint International Conference of the Association for Computers and the Humanities and the Association for Literary and Linguistic Computing, 2003.)

Kiernan, Kevin. "The nathwylc Scribe and the Beowulf Palimpsest." *Poetry, Place and Gender: Studies in Medieval Culture in Honor of Helen Damico*. Ed. Catherine E. Karkov and Nancy van Deusen. Kalamazoo, MI: Medieval Institute Press, Forthcoming in 2005.

Kiernan, Kevin, W. Brent Seales, and James Griffoen. "The Reappearances of St. Basil the Great in British Library MS

Cotton Otho B. x." *Computers and the Humanities* 36 (2002): 7-26. (Image-based Humanities Computing. ed. Matthew Kirschenbaum.)

Kiernan, Kevin. "Digital Facsimiles in Editing: Some Guidelines for Editors of Image-based Scholarly Editions." *Electronic Textual Editing*. Ed. John Unsworth, Katherine O'Brien O'Keefe and Lou Burnard. Modern Language Association and the TEI Consortium, 2005.

Kiernan, Kevin. "Odd Couples in Ælfric's Julian and Basilissa in British Library Cotton MS Otho B. x." *Beatus vir: Studies in Anglo-Saxon and Old Norse Manuscripts in Memory of Phillip Pulsiano*. Ed. Kirsten Wolf and A.N. Doane. Tempe, AZ: Medieval and Renaissance Texts and Studies (MRTS), Forthcoming in 2005.

Lecolinet, Eric, Laurent Robert, and Francois Role. "Text-image Coupling for Editing Literary Sources." *Computers and the Humanities* 36 (2002): 49-73. (Image-based Humanities Computing. ed. Matthew Kirschenbaum.)

Prescott, Andrew. "'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*. Ed. C.J. Wright. British Library Publications, 1997. 391-454.

Seales, W. Brent, James Griffoen, Kevin Kiernan, C.J. Yuan, and Linda Cantara. "The Digital Atheneum: New Technologies for Restoring and Preserving Old Documents." *Computers in Libraries* 20.2 (February 2000): 26-30. Accessed 2005-04-07. <<http://www.infotoday.com/cilmag/feb00/seales.htm>>

Sperberg-McQueen, C.M., and Lou Burnard, eds. *Guidelines for Electronic Text Encoding and Interchange; XML-compatible edition*. Chicago and Oxford: TEI P4, 2001. XML conversion by Syd Bauman, Lou Burnard, Steven DeRose, and Sebastian Rahtz.

Unsworth, John. "Reconsidering and Revising the MLA Committee on Scholarly Editions' Guidelines for Scholarly Editions." *Panel on "New Directions for Digital Textuality."* 2001 Conference of the Society for Textual Scholarship. 19 April 2001. Accessed 2005-04-07. <<http://www.iath.virginia.edu/~jmu2m/sts2001.html>>

Yergeau, Francois, Tim Bray, Jean Paoli, C.M. Sperberg-McQueen, and Eve Maler, eds. *Extensible Markup Language (XML) 1.0 (Third Edition)*. W3C Recommendation, 4 February 2004. Accessed 2005-04-07. <<http://www.w3.org/TR/2004/REC-xml-20040204/>>

## Building Tools for Image-Based Electronic Editions

Alex Dekhtyar and Ionut E. Iacob

The *EPT* serves to organize the raw materials of digital scholarship – digital image and text files – and, using specialized encoding, builds these materials into a usable electronic edition. This edition will include a wide variety of editorial information, including organizational description of both physical (books, folios, lines) and semantic (sentences, words), glossarial and metrical description, and description of the condition of the physical object, notably how that condition interacts with the text on the page. For this reason, it is vital for a successful IBEE that the *EPT* enable the editor to create links between the images and text.

The eXtensible Markup Language (XML) is preferred by the humanities computing community as data support for electronic text encoding, most notably through Guidelines of the Text Encoding Initiative. Although XML does not well capture complex text structures (its strict hierarchical organization severely limits its usefulness in describing, for example, both physical and textual organization in a single file), its relative simplicity recommends it over more powerful but complex representations. Moreover, XML is well supported by software processing tools, from databases, parsers and editors (supporting syntax coloring and on-the-fly validation) to query engines and XML transformations. Many good XML editors are available at no, or very low, cost, which makes XML an even more attractive choice for humanities text encoding.

Building an electronic edition is a tedious enterprise. The editor using traditional XML software must encode editorial information while remaining mindful of XML syntax and the limits imposed by its use. A misplaced tag can keep an XML file from validating, and often an editor will have to choose between encoding different aspects of the manuscript text or risk overlapping markup (for example, the physical organization of a folio – the lines as they appear on the page – may conflict with the sentence structure of the text). Things become more complicated when images are involved. The editor has to keep track of images and record relationships between text and image, not just relating entire folios to the text on that folio, but identifying corresponding regions of text and image. The unfortunate result of this process is that as the complexity of the encoding increases, the editor must concentrate on the syntax of encoding rather than on the details of the text of the manuscript or edition. Our goal was to design tools that allow the editor to concentrate on the act of editing, rather than focus on issues of XML syntax and validity.

As James Clark points out, there are two main classes of XML editors: *text editors* and *structural editors*. The key difference between these two kinds of editors is the way markup is

introduced. Structural editors focus on data-centric encoding, and the editing process begins with markup. The human editor adds content to an encoding template, in a manner similar to entering items in a database. This is in contrast to text editors, which focus on document-centric editing and begin with the textual content (PCDATA). The editor inserts markup into (or deletes it from) the content one tag at a time. The text editor approach is much preferable for humanities editing in general and image-based encoding specifically, as it gives the human editor control over exactly what markup is entered where in the text. This control is important for image-based editing, as it facilitates the recording of image-text relationships by allowing the human editor to select specific sections of text and, with the right software support, relate that text to the corresponding sections of image. Another issue that arises in document-centric encoding is that the XML document may not be valid during the editing process: the order in which the editor introduces the markup in the text may depend not on the requirements of the DTD, but rather on the *modus operandi* of the human editor (which in turn depends on the semantics of the features to be encoded).

Thus, an image-based XML editor has to have the following features:

- Hide the XML syntax if requested. The focus of the human editor should be on text semantics and how images and text are connected. Instead of displaying the complete XML, show where markup exists by highlighting the relevant text in the display. The editor may at times wish to examine the XML encoding. In that case, the XML editor should provide a system for filtering out unwanted markup, showing only those elements that the editor wishes to see.
- Allow text markup by enabling the editor to select the range of content to be marked up and the tag (and attribute values) to be inserted. Among tag attributes, at least one is dedicated to link text and corresponding image or image region.
- Provide support for the editor to connect the markup with the corresponding manuscript image and a specific region in the image. While the editor selects the related areas, the information for mapping the image to the text should be saved automatically by the software — the editor should not have to concern himself with creating image maps or noting image coordinates.
- Assure document well-formedness and provide support for (partial) validation in such a way that it is transparent to the human editor. Imposing validity constraints for update operations might be too prohibitive in text encoding applications: not every update operation (or a set of consecutive update operations) yields a valid document. The software takes further update decisions based on the current status of a document. At the same time, it is important to be able to verify at each moment of time that



the current XML fragment is 'on track', i.e., that the human editor has not committed any structural error while introducing the markup (in which case markup deletion is required). We call this *potential validation* and we designed and implemented an algorithm for checking potential validity of document-centric XML documents.

- Provide support for searching for both text and structure, and for searching the encoding of image features described in the XML markup. There are three main types of searches that the editor can perform in an IBEE. First, the text search, through which the editor can search for a string of characters in the edition content. Second the structural search – this information describes how various text and image features are interrelated (words in certain lines or sentences, holes on the folio in the middle of sentences, etc.). And finally, image feature searches. Given a specified region on the image, the software will find all encoded features related to that region or, conversely, will find all image regions corresponding to a given text range or descriptor (for example, find all image regions with corresponding damage markup).

The architecture of our image-based XML editor is presented in Figure 2.

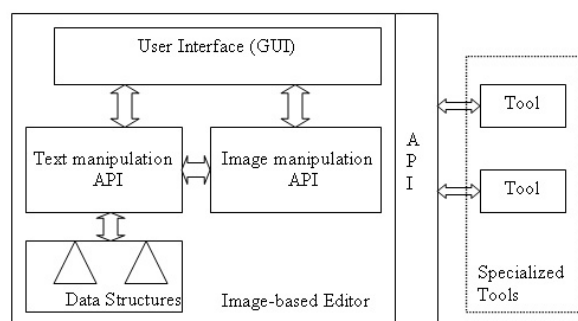


Figure 2. Image-based XML editor

In this paper, we will describe how we designed and built the *EPT* and its individual components to incorporate those elements that we found most important for image-based, document-centric editing.

## Bibliography

Brown, Michael S., and W. Brent Seales. "The Digital Atheneum: New Approaches for Preserving, Restoring, and Analyzing Damaged Manuscripts." *Proceedings of the First ACM/IEEE-CS Joint Conference on Digital Libraries*. New York: ACM Press, 2001. 437-443.

Brown, Michael S., W. Brent Seales, Kevin Kiernan, and James Griffioen. "3D Acquisition and Restoration of Medieval

Manuscript." *Communications of the ACM: Special Issue on Digital Libraries*. May 2001.

Clark, James. "Incremental XML Parsing and Validation in a Text Editor." Presentation at XML 2003, Philadelphia. December 2003.

Hayes, Deborah. "Glossing Damaged Manuscripts: an Example from Ælfric's Lives of Saints." Presentation at Digital Resources for the Humanities (DRH01). University of London, London, UK. 10 July 2001.

Kiernan, Kevin, Alex Dekhtyar, Jerzy W. Jaromczyk, Dorothy Carr Porter, and Ionut Emil Iacob. "Edition Production Technology (EPT) and the ARCHway Project." *DigiCULT.Info* 8 (August 2004): 36-38.

Seales, W. Brent, James Griffioen, Kevin Kiernan, C.J. Yuan, and Linda Cantara. "The Digital Atheneum: New Technologies for Restoring and Preserving Old Documents." *Computers in Libraries* 20.2 (February 2000): 26-30. Accessed 2005-04-07. <<http://www.infotoday.com/cilmag/feb00/seales.htm>>

Sperberg-McQueen, C.M., and Lou Burnard, eds. *Guidelines for Electronic Text Encoding and Interchange*. Chicago and Oxford: TEI P4, 2001.

Yergeau, Francois, Tim Bray, Jean Paoli, C.M. Sperberg-McQueen, and Eve Maler, eds. *Extensible Markup Language (XML) 1.0 (Third Edition)*. W3C Recommendation, 4 February 2004. Accessed 2005-04-07. <<http://www.w3.org/TR/2004/REC-xml-20040204/>>

Yuan, C.J., and W. Brent Seales. "Guided Linking: Efficiently Making Image-to-Transcript Correspondence." *Proceedings of the First ACM/IEEE-CS Joint Conference on Digital Libraries*. New York: ACM Press, 2001. 471. Accessed 2005-04-07. <<http://www.infotoday.com/cilmag/feb00/seales.htm>>

## The ARCHway Software Infrastructure: a platform and utilities for building electronic editions

Jerzy W. Jaromczyk and Neil Moore

In this paper we will discuss the implementation of *EPT*'s architecture, specifically focusing on those utilities that form its underlying skeleton or infrastructure. These utilities support the consistent management of diverse sources of data while providing an extensible framework for building and organizing the editorial tools discussed in the previous papers.

The *EPT*'s architecture is based on the plugin, an encapsulated and independent software unit that "plugs into" a larger whole, extending its functionality. An editorial workbench built from many individual tools gives the editor the freedom to pick and

choose tools for the tasks at hand. We selected *Eclipse* as the platform for both the development and deployment of the *EPT* because it seemed it would well support such a free and configurable approach to the design of the editorial workbench. In this presentation we will briefly describe the *Eclipse* platform and the functionality and implementation of a selection of *EPT* plugins that provide the infrastructure for accessing data, organizing and annotating resources and defining different models for electronic editions.

*Eclipse* is an open-source platform designed to serve as an extensible integrated development environment (*Eclipse Platform Technical Overview*). Originally developed by IBM, *Eclipse* is now maintained by the Eclipse Foundation and an extensive user's community. Although initially intended for software development, *Eclipse's* open-ended plugin-based architecture allows users to extend it to support an unlimited variety of other tasks including software deployment.

The organization of *Eclipse* is a collection of plugins: loosely-coupled software components, often developed independently of one another, which communicate with each other and with which users can interact using well-defined interfaces. Much like the more familiar web browser plugins (such as the Macromedia *Flash* plugin, Sun's *Java* plugin, and Adobe's *Acrobat Reader* plugin), *Eclipse* plugins 'hook into' so-called extension points defined elsewhere in the application, extending and enhancing the application's existing functionality as well as adding completely new features. However, *Eclipse* differs notably from most other extensible software. Most significantly, the platform itself is built from scores of plugins which themselves extend other plugins and provide additional extension points; almost everything in *Eclipse* is a plugin. This may be contrasted with, for example, web browsers, where plugins extend an underlying monolithic software system and rarely interact with one another.

The extensibility of the plugin architecture will be a tremendous benefit to the users of the *EPT*. Humanities researchers have various needs and editorial styles; with a plugin system, they can create a personal editing workbench containing only the tools they need, without losing the advantages of a coherent interface and uniform access to data. Furthermore, scholars with specific needs unforeseen by the *EPT* developers may collaborate with programmers to develop their own editing tools or modify existing ones. The *EPT's* plugin architecture allows users to develop new tools separately from and independent of the *EPT* and then plug them into the *EPT* extension points, providing a seamlessly integrated experience. In effect, such plugins become an integral part of a customized version of *EPT*, on equal footing with the many tools that make up the *EPT* proper.

The *EPT* organizes its plugins in a series of layers, with each layer building upon, using, and extending the layers below. We

will discuss three of these layers in our presentation. The bottommost layer is called the Data Layer. The plugins making up the Data Layer provide a consistent set of operations for managing, reading, and storing various types of edition data files such as images, configuration files, textual transcripts, marked-up edition documents, and XML document type declarations (DTDs). The Data Layer provides a single interface for accessing all data, regardless of where that data is stored — in the local file system, a database (see Dekhtyar et al.), or a remote site. Plugins called data source drivers extend the Data Layer by providing functionality to access resources through different means. Currently the *EPT* contains two data source drivers, one for accessing files located within the file system of the computer running the *EPT*, and one for accessing resources stored on a remote web server, using the HTTP protocol for the World Wide Web. The flexibility of the Data Layer framework allows the user to implement a wide variety of data source drivers; users could build drivers that transparently compress or encrypt data, drivers to maintain data in a relational database, and many others.

On top of the Data Layer sits the Project Explorer, which provides a higher-level view of the resources comprising an electronic edition project. Project Explorer provides a user interface for viewing and managing the logical structure — the model — of a project (as opposed to its physical structure, as an edition project may take its components from several different data repositories). Project Explorer provides a rich set of extension points so that other plugins may contribute resources to the Explorer view. Such contributions provide various actions (e.g., launch a tool, display an image, etc.), which operate on the model, and on the resources themselves.

The Resource Registry, built atop the Data Layer and Project Explorer, enables the user to organize, categorize, and manage collections of similar resources. For example, one collection might contain all the manuscript images comprising the electronic edition. Each collection has a schema, a list of attributes applicable to all resources in the collection. The editor defines collections and their schemas in the Resource Registry, and then adds resources to those collections, describing them by specifying the attribute values for each resource. For example, the schema for manuscript images might contain attributes describing the folio name (038v, for example); the image format (JPEG, GIF, TIFF, etc.); the type of lighting used when digitizing the image (e.g., overhead white light, ultraviolet light, or fiber-optic backlight); the provenance of the image files; and so forth. The Resource Registry contributes items to the Project Explorer, arranging the resources by a user-defined ordering of their attributes. Other plugins can issue queries to the Resource Registry asking for resources with certain attributes, for example all the ultraviolet manuscript images which are in the JPEG format.

The utilities described above form part of the infrastructure for the *EPT*. They provide a workbench within which a user can arrange various specialized tools, such as ones described in the other papers in this session, in convenient combinations capable of solving complex tasks in the production and presentation of image-based electronic editions.

## Bibliography

Dekhtyar, Alex, et al. "Database Support for Image-based Electronic Editions." *Proceedings, 10th International Workshop on Multimedia Information Systems (MIS 2004)*, August 25–27, 2004, College Park, MD. 2004. 147-156.

*Eclipse*. Accessed 2005-04-12. <<http://www.eclipse.org/>>

*Eclipse Platform, Technical Overview*. Accessed 2005-04-12. <<http://www.eclipse.org/whitepapers/eclipse-overview.pdf>>

Jaromczyk, Jerzy W., and Sandeep Bodapati. "An Architecture Promoting Collaborative Research, Teaching and Learning." *Proceedings, Joint International Conference of the Association for Computers and the Humanities and the Association for Literary and Linguistic Computing, May 29–June 2, 2003, Athens, GA*. 2003. 10.

Jaromczyk, Jerzy W., and Neil Moore. "Geometric data structures for multihierarchical XML tagging of manuscripts." *Proceedings, 20th European Workshop on Computational Geometry, Seville, Spain, March 2004*. 2004. Accessed 2005-04-14. <<http://www.us.es/ewcg04/Articles/jaromczyk.ps>>

Kiernan, Kevin, Jerzy W. Jaromczyk, Alex Dekhtyar, Dorothy Carr Porter, Kenneth Hawley, Sandeep Bodapati, and Ionut Emil Iacob. "The ARCHway Project: Architecture for Research in Computing for Humanities through Research, Teaching, and Learning." *Literary and Linguistic Computing* (Forthcoming in 2005). (Special issue, papers from Joint International Conference of the Association for Computers and the Humanities and the Association for Literary and Linguistic Computing, 2003.)

## Historical Lexicons in Medieval and Early Modern English and French

**Anne Lancashire** ([anne@chass.utoronto.ca](mailto:anne@chass.utoronto.ca))

University of Toronto

**Jennifer Roberts-Smith**

([roberts-smith@sympatico.ca](mailto:roberts-smith@sympatico.ca))

University of Toronto

**Brian Merrilees** ([brian.merrilees@utoronto.ca](mailto:brian.merrilees@utoronto.ca))

University of Toronto

The *TAPoR* (*Text Analysis Portal for Research*) network operates a Lexical Analysis Laboratory in the Roberts Library at the University of Toronto. The lexicographical projects associated with *TAPoR* Toronto include the *Old English Dictionary*, *Aalma* (a database in medieval French), the *Lexicons of Early Modern English*, and the *Mayors and Sheriffs of London* (to date, to 1558). These four projects share a scholarly objective, the accurate representation of ancient languages, both terms and names, so that researchers in all *TAPoR* institutions can access them, whether in SQL databases or in textbases created with *XTeXT* (Isagn Inc.), through Geoffrey Rockwell's *TAPoR* portal at McMaster University.

Anne Lancashire is converting her index of the *Mayors and Sheriffs of London* (*MASL*), recently published in a book on medieval London history, into an online database so that it can be enhanced and updated well into the future. The conversion process was not an easy one. Relationships among dates and proper names that appear straightforward on the page become problematic within a database designed especially for them. The book is not identical with the structured intelligence that goes into an SQL database. Lancashire describes how working with a web development group of digital librarians expands the historical research that went into the book.

Jennifer Roberts-Smith, a doctoral student in English at Toronto, discusses how *Lexicons of Early Modern English*, a text-database of over 120 glossaries and dictionaries published in England between 1480 and 1702, contributes to her reassessment of the prosody of Shakespeare and his fellow dramatists. Roberts-Smith, a director and a member of ACTRA, shows that the language of metrics in the Early Modern period is a musical one. From words like *tune* as used by lexicographers like Robert Cawdrey, it is possible to infer that

Elizabethan and Jacobean dramatic verse has "an inherent temporal rhythm that guides actors as to the relative pace of their delivery." Many historical scholars have observed that Renaissance English, lexically, offers many of the same problems to a modern reader as does a foreign tongue. This estrangement may now be extended to metrics.

Brian Merrilees is undertaking an online edition of fifteen manuscript glossaries of medieval French that vary widely but are, as a group, termed the *Aalma*. One challenge in this project is to find a way to give a full collation of variant readings from many manuscripts at the same time as to maintain the integrity of the various manuscript versions. For a long time Merrilees has worked very productively with *WordCruncher*, a 1980s interactive concordancer. Faced with the challenge of constructing an online hybrid, at once a collection of well-edited texts and a large index of variant readings, Merrilees is experimenting with a new generation of SQL database and *XTeXT* textbase technologies.

Lancashire, Roberts-Smith, and Merrilees differently show how, once digitized within a structured form, lexical and onomastical materials conspire to become a semantic web.

(Ian Lancashire)

## Mayors and Sheriffs of London (MASL)

### Anne Lancashire

*Mayors and Sheriffs of London (MASL)* is a searchable database covering at present all mayors and sheriffs of the city from 1190 to the accession of Elizabeth I in 1558. This list initially appeared in print as an Appendix (pp. 308-355) in London historian Caroline M. Barron's *London in the Later Middle Ages*, published by Oxford University Press in early 2004.

Such reference lists of London mayors and sheriffs have existed from early times: in medieval and early modern chronicle histories (in manuscript and in print), and in urban historian John Stow's *A Survey of London* (first published in 1598). Most of these early lists are, however, incomplete, and contain many inaccuracies, so that British institutions such as the Public Record Office and the Corporation of London Records Office have had to compile their own handlists, for reference, for staff working with London records. While more accurate than the original lists, these handlists are also not complete; the names, for example, of mayors and sheriffs replacing those who died in office are sometimes omitted, and the specific replacement dates are almost never included. Before the more modern handlists were compiled, C.L. Kingsford, a scholar working in the early 20th century on London chronicles, had already corrected Stow's *Survey of London* list in a two-volume edition of the *Survey* published in 1908; but Kingsford did not fill out to any significant extent the material Stow had provided. Above all, none of these early lists or the more recent records offices'

lists normally include the livery company memberships of the two sheriffs (and occasionally, early on, of the mayor as well) holding office each year.

*MASL* includes for each mayor and sheriff not only specific term-of-office dates (including, for mid-term replacements, the dates of both election and swearing-in where these dates differ and can be found) but also, above all, the occupation or company membership of each individual, where this could be ascertained from early manuscript and/or print records. Company membership of mayors and sheriffs is an important factor for London historians to consider, given the importance of the companies from earliest times, as political organizations, in the government of the city. *MASL* allows its users to search for mayors and sheriffs by name, by year, or by company; it can also provide chronological listings of mayors and sheriffs for a defined period of time; and it notes the sources of all information provided, through a searchable list of references. It is currently in the late stages of construction, with links being provisionally added to the web sites — where these exist (as they increasingly do) — of the companies themselves. Once the 1190-1558 list has been tested and debugged, gradually more years will be added, in chronological order, so that eventually the list will reach the present day. Gradually also references such as biographical sources on the various mayors and sheriffs will be introduced. As a database, *MASL* can be revised, corrected (as more historical work is done), and extended in whatever directions its users find helpful.

A number of problems have been encountered in moving *MASL* from print form into an electronic database. For example, the spelling of names in the early period was not stable; and any one mayor or sheriff might have a name spelled in several different ways. When the user consults a printed list, s/he visually and rapidly scans the printed pages looking for an approximation or version of the name in which s/he is interested. The user of a searchable database, however, must type in a name, choosing a spelling. If this is not the spelling the database has used, the name will not be found. We have therefore had to wrestle with what to do about alternative spellings. Another problem is that a printed list will begin with an introduction explaining how it is to be used: for example, what its date ranges mean; but the user of an electronic database does not normally stop to read an introduction but simply types in the name or date s/he is looking for. How does one convey to such a user the information s/he needs, with a list as complex and problem-based as *MASL*, so that s/he will not misinterpret the information s/he is finding?

Yet another problem involves livery companies. At any given time in London's early history, about 100 of these companies existed — composed at first of members of specific occupations or trades, such as goldsmiths, grocers, shoemakers, ironmongers, and weavers — at first, for the purposes of

regulation of trade, community religious worship, and the provision of social assistance to members in need, but increasingly, from the mid-15th century on, for purposes such as business connections of all kinds, political clout, and social status. Many of these companies still exist today, functioning as essentially private clubs, and as still an important part of political, social, and business connections in the city. Many of the companies from the earlier periods, however, no longer exist (some companies died; others grew up to replace them and in turn expired); and in order to provide appropriate information on company memberships to *MASL* users, who will not — unlike the users of the printed list — have a book on London history in their hands, a good deal of extra work will be involved in eventually linking information on these now extinct companies to their names in the database, along with providing links to the web sites of still-flourishing companies which have developed a web presence. Further expansions then also become possible: for example, references to reliable company histories.

*MASL*, in short, can be developed into an extremely useful reference tool for various aspects of early London history, and can be continuously updated and expanded, but presents interesting challenges in the conversion from print to database. One major advantage of the database, however, as correctable and expandable, lies in the fact that users will be able to make corrections and to add material from their own research; they will be invited to send to the *MASL* editor their corrections and additional information, which will then be checked and, credited to them, added to the site. Reference works such as *MASL*, which can be easily updated in this way, and also linked to other historical sites and records collections, would seem to be especially suited to electronic database format.

## Bibliography

Barron, Caroline M. *London in the Later Middle Ages*. Oxford: Oxford University Press, 2004.

Beaven, Alfred B. *The Aldermen of the City of London*. 2 vols. London: Corporation of London, 1908-13.

Brooke, Christopher N.L., and Gillian Keir. *London 800-1216: The Shaping of a City*. London: Secker and Warburg, 1975.

Hughes, A., and Gillian Keir. *Public Record Office: Lists and Indexes*, #9. London: Public Record Office, 1898; rpt. 1963.

Lancashire, Anne. *London Civic Theatre: City Drama and Pageantry from Roman Times to 1558*. Cambridge: Cambridge University Press, 2002.

Reynolds, Susan. "The Rulers of London in the Twelfth Century." *History* 57 (February 1972): 337-357.

Stow, John. Ed. Charles Lethbridge Kingsford. *A Survey of London*. 2 vols. Oxford: Clarendon Press, 1908.

## In tune with the times?: English Renaissance metrics and the Lexicons of Early Modern English

Jennifer Roberts-Smith

The *Lexicons of Early Modern English* is a database of more than 120 keyed, thoroughly proofed electronic transcriptions, minimally tagged for lexical content, of lexical works dating from 1480 to 1702. They are accompanied by a bibliography and a search engine. *LEME*'s coherence, simplicity and accuracy are its key attractions.

These conditions are essential to a project like mine, which I suspect is typical of many humanities research projects in that it requires efficient access to a large corpus of related works free from the interpretive intervention of modern habits of thinking. Trusting an instinct I felt as an actor, I set out to discover whether Shakespeare systematically used syllable-duration in the metre of his plays. There are, of course, no scanned transcriptions of complete verse works for the stage dating from Shakespeare's lifetime extant; nor were Shakespeare and most of his contemporary playwrights inclined to write prosodies. However, the analogy between metre and music that is implied by my hypothesis — that if syllables have durations, they are like musical notes — was commonplace in the works of literary prosodists writing during Shakespeare's lifetime. This analogy has been attributed by scholars for at least 100 years as belonging to the sixteenth-century movement which aimed to dignify English poetry by showing that it was similar to Classical poetry. But modern linguists have shown that the analogy was linguistically sound: Philip Sidney, for one, accurately identified the phonological durations of the English syllables he employed in his quantitative verse (Kristin Hanson 2001). Sidney, however, was arguably an elitist, whose work was not published during his lifetime and whose esoteric quantitative metrical experiments even he lost faith in: he abandoned them. The question is, were the musical-metrical prosodists also arguably esoteric elitists? They use musical vocabulary to describe English, as well as Classical, metres; but can their accounts of English metres be trusted to be accurate, empirical, and accessible to the ordinary auditor, rather than eccentric, abstract, and nationalistically ambitious? *LEME* makes this question answerable.

*LEME*'s evidence is provided in five ways. First, the corpus is comprised entirely of lexical works. This limits the potential for the errors of interpretation that can easily be made when a modern reader is trying to deduce meaning from a word's literary context: lexical works name the things in the world to which words point and place words in the context of their synonyms. Second, *LEME* returns search results in the context of entire word entries, identifies the sections of works in which results appear, and provides electronic links to the entire works cited in the search results. This minimises the alternative hazard

of interpreting a citation too far out of context. Third, the *LEME* works represent a broad range of linguistic contexts: they are bilingual dictionaries, monolingual dictionaries, glossaries of hard or foreign words, rhetoric handbooks, grammars, glossaries of technical terms or 'words of art'. Patterns of word usage can hence be analysed according to topical context and register. Fourth, since *LEME* searches entire lexical works, not merely representative quotations as, for example, the *Oxford English Dictionary* does, patterns of word usage can be analysed statistically. It is possible, for example, to say confidently that the English metrical term *tune* is used in four of the thirteen works in the corpus in which it occurs, exclusively in the sense employed by George Puttenham (attrib.) in his *Arte of English Poesie*. (Roberts-Smith 2003) Fifth, the minimal tagging structure of the *LEME* database provides efficient access to patterns of equivalence that may not be apparent in un-tagged searchable texts. The *LEME* headword tag, for example, is used to identify words being explained, a status indicated by a variety of conventions in the original texts which may not be apparent in the keyed text: the graphical arrangement of words on a page, perhaps, or the kind of type in which the word is set. A search for the word *tune* will return a result showing that in none of 251 occurrences of the word in the *LEME* corpus is it used as a headword (Roberts-Smith 2003). These five avenues of access have allowed me to assess the demographic and chronological currency of the usages of a series of metrical terms common in literary prosodies.

For prosodists such as Puttenham, Webbe, Gascoigne, and Campion, these terms fall into two categories: English words, like *tune* or *ditty*, and words borrowed from Classical languages, like *accent* and *iambic*. When they use the unfamiliar Classical terms, they are at pains to explain them using the plain English words; but their explanations do not always match those given in contemporary bilingual and monolingual English dictionaries, as they are represented in *LEME* database. *LEME* shows that English Renaissance prosodists' lexical eccentricities reveal the limitations of their comparison between English metre and Classical metre. In some cases, they find it necessary to broaden the signification of Classical metrical terms to include elements they believe to be unique to English metre (stress, for example); in other cases, they limit Classical terms to exclude elements of English metre which have fallen out of fashion (rhyme, for example). These observations will not be surprising to scholars of English metrics: they show the neo-Classical reform project at work. What may be surprising is that a comparison of the lexical ranges of Classical and English metrical terms in the Lexicons database reveals that the English terms were inherently musical: they indicated syllable duration and vocal pitch in addition to stress and number and they could almost always be used to refer to spoken or sung compositions with equal accuracy and frequency. So the comparison between metre and music was widespread, commonplace, and English; it was not

merely a function of the metrical reform project which coloured the prosodies.

In combination with lexical analyses of the same words in Shakespeare's works, this evidence argues that Shakespeare thought of his own verse in musical terms. A survey of Shakespeare's works shows that his usage of metrical terms matches those found in the dictionaries rather than those found in the prosodies. In other words, this aural and vernacular poet still saw the gulf between Classical metrics and traditional, native English metrics. He thought of himself as writing English metre, tuneful iambics, perhaps.

The *LEME* database, then, has provided me with an empirical basis for further investigation of my topic. If Shakespeare equates *tune* with *time* in *As You Like It* (5.3.32) does he think of the metre of his own dramatic verse, which William Webbe the prosodist and Robert Cawdrey the lexicographer would both call its *tune*, as literally incorporating *time* in the form of relative syllable duration? Is Shakespeare writing musical verse with an inherent temporal rhythm that guides actors as to the relative pace of their delivery? If this is the case, do we need to revisit our understanding of the English Renaissance dramatic iambic, to locate it in the context of the native English metrical tradition revealed by the lexical works in the *LEME* database?

## Bibliography

- Campion, Thomas. "Observations in the art of English poesie." *Elizabethan Critical Essays*. Ed. G.G. Smith. London: Oxford University Press, 1602; 1904. 327-355.
- Cawdry, Robert. "A table alphabeticall." *Lexicons of Early Modern English*. Ed. I. Lancashire. Toronto: University of Toronto Libraries and University of Toronto Press, Alpha vers. Dec. 10, 2004. <<http://link.library.utoronto.ca/leme/public/index.cfm>>
- Gascoigne, George. "Certayne Notes of Instruction." *Elizabethan Critical Essays*. Ed. G.G. Smith. London: Oxford University Press, 1574; 1904. 46-57.
- Hanson, Kristen. "Quantitative meter in English: the lesson of Sir Philip Sidney." *English Language & Linguistics* 5.1 (2001): 41-91.
- Lancashire, I., ed. *Lexicons of Early Modern English*. Toronto: University of Toronto Libraries and University of Toronto Press, Alpha vers. Dec. 10, 2004. <<http://link.library.utoronto.ca/leme/public/index.cfm>>
- Puttenham, George, attrib. "The Arte of Poesie." *Representative Poetry Online (Vers. 3.0)*. Ed. I. Lancashire. Toronto: Web Development Group, InformationTechnology Services, U. Toronto Library, 1589; 2002. Accessed 2003-03-10. <<http://eir.library.utoronto.ca/>

[rpo/display/displayprose.cfm?prosenum=17&sub\\_file=puttenham\\_artofp\\_all.html](http://www.eir.library.utoronto.ca/rpo/display/displayprose.cfm?prosenum=17&sub_file=puttenham_artofp_all.html)

Puttenham, Richard, attrib. "The Arte of English Poesie." *Early English Books Online*. ProQuest Information and Learning, 1589. Accessed 2003-03-10. <[http://eir.library.utoronto.ca/rpo/display/displayprose.cfm?prosenum=17&sub\\_file=puttenham\\_artofp\\_all.html](http://eir.library.utoronto.ca/rpo/display/displayprose.cfm?prosenum=17&sub_file=puttenham_artofp_all.html)> STC 20519.

Roberts-Smith, Jennifer. *Musical-metrical vocabulary in Early Modern English: a preliminary lexicon*. Toronto: Renaissance Society of America, 2003.

Roberts-Smith, Jennifer. "Puttenham rehabilitated: the significance of 'tune' in The Arte of English Poesie." *Computing in the Humanities Working Papers/Text Technology* 12.1 (September 2003). Accessed 2005-04-13. <[http://www.chass.utoronto.ca/epc/chwp/CHC2003/Roberts\\_Smith2.htm](http://www.chass.utoronto.ca/epc/chwp/CHC2003/Roberts_Smith2.htm)>

Shakespeare, William. Ed. Agnes Latham. *As You Like It*. London: Routledge, 1991. 5.3.37-41.

Sidney, Philip. Ed. J. Robertson. *The Countess of Pembroke's Arcadia*. Oxford: Clarendon Press, 1590; 1973.

Webbe, William. "A Discourse of English Poetrie." *Elizabethan Critical Essays*. Ed. G.G. Smith. London: Oxford University Press, 1602; 1904. 226-302.

## The Aalma Project: research in early French lexicography

### Brian Merrilees

'Aalma' was the name given by the great French romaniste Mario Roques to a group of Latin-French glossaries compiled in the fourteenth and fifteenth centuries and which served as language learning tools in Medieval France. These glossaries are of particular interest in the history of French lexicography both on account of their number, fifteen known to date, and on account of their importance in the development of French vocabulary. Roques edited one text, the version found in Bibliothèque nationale de France, ms. lat 13032, in 1938; another from Lille had been edited in the 19th century, but no edition exists that takes into account all versions of the *Aalma*. The aim of my research group at the University of Toronto is to produce, first of all, searchable on-line editions of what we believe to be the key versions of the *Aalma* and in addition add to a database we have been assembling for the past several years of medieval French lexicographical material. Later a printed edition will be envisaged.

To date we have transcribed five versions of the *Aalma*, Paris, BnF. lat. 13032, Metz Bibl. mun. 510 et 1182, Lille, Bibl. mun. 147, Salins, Bibl. mun. 44. We are currently transcribing Exeter

Cathedral, ms. 3517 and St Omer, Bibl. mun. 644, and we have done a sample letter (B) from Paris, BnF, lat. 14748, Paris, BnF, lat. 17881, Paris, BnF, lat. 7679 and Epinal, Bibl. mun. 224. In addition we have a full transcription of one of the first dictionaries printed in France, the *Catholicum parvum* done in Paris around 1484 by Antoine Caillaut, who used Metz 510 as his printer's copy.

All material transcribed is added to a local database of lexicographical materials using an old DOS program, *WordCruncher*, developed in Utah in the 1980's. This admirably simple program has been a boon to non-technical scholars like myself. Text preparation is minimal, no high level of mark-up is required and the indexing process is rapid. Given that everything in our database has been copied, laboriously, from manuscript and microform, this is a great advantage and we have been able to use this program above all others in our analytical work.

Our only online venture so far has been a simple database search of the *Aalma* in BnF lat. 13032. Here we checked and corrected the Roques text and set up a search using Active Server Page. Again this was simple and effective. It is still on our website:

<[http://www.chass.utoronto.ca/~merrilee/2003/searchset\\_temp.htm](http://www.chass.utoronto.ca/~merrilee/2003/searchset_temp.htm)>

There are also excerpts from other dictionaries can also be found there under the 'Research' rubric on my website: <<http://www.brianmerrilees.com>>.

This, however, is not enough. We aim to expand the Paris 13032 base to include several other versions of the *Aalma*, probably one version at a time. We are looking for suitably uncomplicated software that will translate the *Aalma* texts into a coherent base. Our next experiment will be with *XTeX*, an Ontario-based text search-engine, and from there we shall seek other proposals.

What we seek essentially is a minimally marked text entry from which maximal benefit can be drawn. If this can be done in a DOS context, as we have found with *WordCruncher*, then it behooves us to find something equally simple for on-line presentation.

On a different tack, we have applied a database program to the comparison of 10 versions of the *Aalma*, using *Microsoft Access*. Here we entered the letter B from those versions and obtained a preliminary indication of the relationships between the various versions. Here is a small sample (not all versions contain each lemma):

Bacca ('bay')

Paris 13032	braie, fruit d'olive ou de lorier et aucune foiz est mis pour tout fruit, principalement d'arbres sauvages
Metz 510	fruit de lorier ou aucune foiz pour tout fruit sauvaige

Metz 1182 baie, fais (sic) d'olive ou de lorier  
Lille 147 fruit d'olivier  
Salins 44 fruit d'olive ou d'olivier fructus olive vel lauri  
quandque pro quoque fructu ponitur ...  
Exeter 3517 fruit de olive ou de lorier .i. fructus lauri vel olive  
St Omer 644 braie, fruit d'olive ou de lorier et aucune fois  
pour tous fruis et especialment de arbres sauvages

Biceps ('two-headed')

Paris 13032 cil qui a ii chiex  
Metz 510 celui qui a deux testes  
Metz 1182 celui o celle qui a deux chiefs  
Lille 147 qui ha ii testes  
Salins 44 ce qui ha deux chief  
Exeter 3517 qui a ii. chiefz ille qui habet duo capita  
St Omer 644 ce qui a ii quiefs

We have recently had microfilms of two of the versions digitized and can therefore put the manuscript image alongside the text as we transcribe it on a large screen. A second large screen sits beside the main screen for calling up references, such as our own *WordCruncher* database, the *Trésor de la Langue française*, Douglas Walker's *Lexique de l'ancien français*, the emerging *Dictionnaire du moyen français* from ATILF in Nancy and the *Patriologia latina* site produced by Chadwyck Healey.

Transcribing from microform or manuscript is a slow process. Our task is nonetheless made easier with the technologies and electronic resources at our disposal.

## Bibliography

*ATILF (Analyse et Traitement de la Langue Française)*. Centre Nationale de la Recherche Scientifique / Université Nancy 2. Accessed 2005-04-14. <<http://atilf.atilf.fr>>

Edwards, William, and Brian Merrilees, eds. *Dictionarius familiaris et compendiosus : le dictionnaire latin-français de Guillaume Le Talleur*. Corpus Christianorum : Continuatio mediaevalis: Lexica Latina Medii Aevii : Nouveau Recueil des lexiques latins-français du moyen âge 2. Turnhout: Brepols Publishers, 2002.

Merrilees, Brian, and William Edwards, eds. *Dictionarius Firminis Verris : Dictionnaire latin-français de Firmin Le Ver*. Corpus Christianorum : Continuatio mediaevalis: Lexica Latina Medii Aevii : Nouveau Recueil des lexiques latins-français du moyen âge 1. Turnhout: Brepols Publishers, 1994.

Merrilees, Brian, William Edwards, and Michelle Troberg. "Vers une nouvelle édition du glossaire latin-français l'Aalma." « *Pour acquérir honneur et pris* » : *Mélanges de moyen français offerts à Giuseppe Di Stefano*. Ed. Maria Colombo Timelli and Claudio Galderisi. Montréal: CERES, 2004. 287-292.

Naïs, Hélène. *La lexicographie au Moyen Âge*. Ed. Claude Buridant. Villeneuve d'Ascq: Presses Universitaires de Lille, 1986. 185-196. Lexique 4

*Patriologia latina*. ProQuest Information and Learning Company. Accessed 2005-04-14. <<http://pld.chadwyck.com>>

Roques, Mario. *Recueil général de lexiques français du moyen âge II*. Paris: Champion, 1938.

*Le Trésor de la Langue Française*. Centre Nationale de la Recherche Scientifique / Université Nancy 2. Accessed 2005-04-14. <<http://atilf.atilf.fr>>

Walker, Douglas, ed. *Lexique d'ancien français*. University of Calgary. Accessed 2005-04-14. <<http://www.acs.ucalgary.ca/~dcwalker/Dictionary/dict.html>>



# Cybertextuality and Text Analysis

*Ian Lancashire* (ian.lancashire@utoronto.ca)  
University of Toronto

A cybertext is any oral, written, mental, or machine-generated language act viewed from within cybernetics, the study of communication and control in living organisms and machines, a theory invented by the American mathematician Norbert Wiener (1894-1964). We author cybertexts by steering or governing their making according to the persistent feedback we receive from all those who observe them. Insofar as computing humanists use software to analyze both literary works and machine-made texts like concordances, they may be said to owe something to cybernetics. Our analytic programs simulate part or all of the messaging and feedback process, some acting as creators, others as reader-listeners or noisy channels. The basis for our software is ultimately how language cognition works. Given that we create most of our own oral and written utterances unselfconsciously, we first encounter them as strangers and observers, not as authors; and our observation always begins with modelling the sense data we have received from ourselves. These mental models act as feedback and help shape the next sentences we make. Cybertextual cycles, each an unselfconscious utterance and a partially conscious modelling and response to it, steer our composition even if no one but ourselves is present to reply to what we utter.

I propose that cybertextual cycles, enacted in cognition, partly shape the idiolect or personal style exhibited by the texts we make. We use text-analysis tools today to detect the idiosyncratic patterns of flat, atemporal documents, but all texts, being cybertextual, unfold in time. An author's silent feedback to his own utterings pulses wave-like in the emerging text, but how can these characteristic waves, that is, the cybertextual style, be recovered from flat documents? Usability software offers some tools for this purpose, as do keyloggers, protocol analysis, and word-processing programs. One way to advance text-analysis methodology in a post-concordancer age is to investigate cybertextual style by recording and analyzing the behaviour of living authors as they write. Usability software like *Morae*, because it externalizes working memory, can capture the tic-tocs of cognitive style.

## Bibliography

- Aarseth, Espen J. *Cybertext: Perspectives on Ergodic Literature*. Baltimore and London: Johns Hopkins University Press, 1997.
- Baddeley, Alan. "Working Memory and Language: An Overview." *Journal of Communication Disorders* 36.3 (2003): 228-66.
- Galison, Peter. "The Ontology of the Enemy: Norbert Wiener and the Cybernetic Vision." *Critical Inquiry* 21 (1994): 677-99.
- Haraway, Donna. "A Manifesto for Cyborgs: Science, Technology, and Socialist Feminism in the 1980s." *Socialist Review* 80 (1985): 65-108.
- Hayles, N. Katherine. *How We Became Posthuman: Virtual Bodies in Cybernetics, Literature, and Informatics*. Chicago: University of Chicago Press, 1999.
- Hayles, N. Katherine. *Writing Machines*. Cambridge, Mass.: MIT Press, 2002.
- Lancashire, Ian. "Cybertextuality." *TEXT Technology* 13.1 (May 2005).
- Lancashire, Ian. "Cognitive Stylistics and the Literary Imagination." *A Blackwell Companion to Digital Humanities*. Oxford: Blackwell's, 2004.
- Lieberman, Philip. *Human Language and Our Reptilian Brain: The Subcortical Bases of Speech, Syntax, and Thought*. Cambridge, Mass.: Harvard University Press, 2000.
- Masani, R.P. *Norbert Wiener 1894-1964*. Basel: Birkhäuser, 1990.
- Miller, G.A. "The Magical Number Seven, plus or minus Two: Some Limits on our Capacity for Processing Information." *Psychological Review* 63 (1956): 89-97.
- Morae*. Okemos, MI: TechSmith, 2004. Accessed 2004-05-27. <<http://www.techsmith.com/default.asp>>
- Olson, David R. *The World on Paper: The Conceptual and Cognitive Implications of Writing and Reading*. Cambridge: Cambridge University Press, 1994.
- Pierce, John R. *An Introduction to Information Theory: Symbols, Signals & Noise*. 2nd edition. New York: Dover, 1980.
- Poeppel, David, and Gregory Hickok. "Towards a New Functional Anatomy of Language." *Cognition* 11 (2004): 1-12.
- Smith, John B., Dana Kay Smith, and Eileen Kupstas. "Automated Protocol Analysis." *Human-computer Interaction* 8 (1993): 101-45.

Wiener, Norbert. *Cybernetics or Control and Communication in the Animal and the Machine*. 2nd edition. Cambridge, Mass.: MIT Press, 1961.

Wiener, Norbert. *The Human Use of Human Beings: Cybernetics and Society*. 2nd edition. New York: Hearst, 1967.

## Using Markup for Multivariate Analyses in the Prosopographical Study "Formation for the Public Sphere"

---

*Monica Langerth Zetterman*  
([monica.langerth@ped.uu.se](mailto:monica.langerth@ped.uu.se))

*Digital Literature, Uppsala University*

---

### Introduction

This paper aims to illustrate how markup might be applied for multiple purposes in research.<sup>1</sup> Here, the TEI/XML encoding scheme<sup>2</sup> was used as a research tool when producing a collective biography in a sociological prosopographical study on prominent Swedish female pioneers around the turn of the century 1900.<sup>3</sup> In this collective biography the markup is used for the exploration of biographical information. Although the markup provided was done for a special reason, namely to extract specific data in order to apply multivariate analyse methods, such as correspondence analysis<sup>4</sup>, it also provides means for presenting, filtering and indexing the material.

### Background

The main purpose of the project *Formation for the public sphere. A Collective Biography of Stockholm Women 1880--1920* is to investigate the social strategies of the first generations of women entering the public sphere in Sweden. This period was of crucial significance for women engaged in philanthropy, reform pedagogy, modern health care, literature and music. These women's strategies, investments and careers differed from their male contemporaries and their contributions are not easily recognisable. In order to discern and interpret their contributions to the establishment of the modern welfare state institutions, a modern educational system and the modern cultural fields, methods from the French sociological tradition founded by Pierre Bourdieu have been used.

A central endeavour is to collect information on the women's social origin, social intercourse, their networks, educational trajectories and matrimonial status. Such information is here called 'assets' or 'capital'. In Bourdieu's sense certain types of capital are acknowledged within certain social groups but not by everyone (Bourdieu, 1992). Each field that is sufficiently

autonomous has its own rules for inclusion, exclusion and rewards, and specific species of capital.<sup>5</sup> By analysing the distribution of certain types of capital among the pioneer women we try to map the structure, the hierarchies and the polarities of domains like female culture, education and philanthropy.

Since we favour the collection of data which is sociologically interpretable data it is important to collect information on names, dates and places, e.g. where and when and with whom she lived, where she worked, in order to trace the "meeting places" and networks. Hence a mandatory core set of data was, whenever possible, harvested to depict some of the most crucial assets:

- Social origin: father's and mother's occupation, education, positions. Number of brothers and sisters. Woman's and parents' place of birth and place of upbringing.
- Educational capital: kind of basic and further education. Sojourns abroad.
- Social capital: influential relatives, matrimonial status, number of children, housing, member of state commissions, foundations.
- Economic capital: wealth, earthly goods and relations to patrons.
- Political and religious capital: positions in political/religious organisations, standpoints in such matters.
- Specific symbolic capital: assets being valued either within certain fields or domains or within women's networks.

Of course many of the biographical texts cover much more, but the main aim is trying to make the collection of these core data as comprehensive as possible for each woman.

## Modelling the Data

There are two kinds of datasets of biographical accounts called capital descriptions. One of the sets consists of one hundred rather extensive texts written in running prose text by the researchers, aimed to be published in for example historical journals or biographical handbooks. The scholars have explored archival material such as letters, diaries or estate reports, as well as printed newspapers, journals or books — and of course existing biographies and autobiographies. The other set is more than 1200 condensed texts based on excerpted information.<sup>6</sup> The excerpts have been transcribed from two volumes, one from 1914 and one from 1921 with biographic articles on prominent Swedish women. Both kinds of texts have been provided with markup according to the TEI guidelines with the additional TEI tag set "Names and Dates" to encode proper names, date periods and precise dates.<sup>7</sup>

Similar to the much more extensive and ambitious *Orlando* project<sup>8</sup> we produce texts and we apply descriptive

model-driven and interpretative markup. Unlike the *Orlando* project, though, we have not developed a DTD for this project.

In our content model each woman corresponds to a main division that contains subdivisions and further subdivisions. In principle each subdivision or sub-subdivision corresponds to one type of capital such as `<div type="soc.orig">`. The basic features marked in a subdivision are:

- a date(range),
- an event or an occurrence,
- name(s).

For each of these elements there is a reciprocal relationship to a biography and an arbitrary category respectively. This means that a division is only instantiated when there is a relationship or an occurrence of a certain kind. Obviously, some instances are mandatory: we need for example to record a name and a birth date in order to instantiate a record for a woman. We use the hierarchies to extract exactly the data we need for the analyses. Thus, we can extract the content "Paris" out of the element `placeName` when and only when it occurs within `div type="travel"`. We might thus extract information on that the woman in question did travel to Paris, and we do not have to bother with all other Parises that might appear: books published in Paris, dresses made in Paris, father born in Paris. Obviously a consistent use of the content model providing for the accuracy of the markup become very important since so much depends on it.

Our approach to apply the content model and do the encoding in running prose is not unproblematic. We have faced, and are still facing, many rather difficult decisions on the ways to encode and how to denote the aspects we want to capture. Should chronology take precedence over events or the other way around? The consistent use of divisions corresponding to the content model is important, since it guarantees that the internal hierarchy of elements is no hindrance for finding the specific data.

The corresponding types of data (i.e. the core set) from the two different datasets have been merged into one dataset for further processing. We extracted the data needed for analyses into tables using XSL in order to import the tables into the statistical software SPSS. SPSS is used for converting string values to numerical values and for organising, combining, classifying and aggregating variables. After preparations in SPSS the datasets were imported into the SPAD software where the correspondence analyses are done.

Since the material is heterogeneous, it calls for some measures to guarantee the consistency of data. If we had chosen a database solution, such as the relational database used by the project on the prosopography of the Byzantine empire, it would be a matter of course to opt for a controlled vocabulary, as do

Bradley and Short. Thus the categorisation is undertaken prior to data input into order to ensure the congruency of the recorded aspects:

Of course, this kind of interpretation of a source — by assigning some aspect to fit into categories — is in fact very similar to an important element of most scholarly work: classification and categorisation are standard part of scholarly practice.

(ibid. 9)

This is very true also in our project. The major decisions on the categorisation have been taken prior to the collection of information and prior to the writing of the biographies.

## Closing

What might separate this project undertaking from other similar prosopographic projects is the aim to maximize portability. The material is available through an ordinary web browser, downloadable and reusable with intact markup. Users should be able to import the material to their own systems or add additional markup and not be forced to stay on our website in order to explore the material or perform their own analyses. Another difference is that the encoded data is used as input to multivariate quantitative analyses, as will be illustrated at the presentation. Thanks to the encoding we can provide enhanced navigation, presentation of different views of the material and filtering possibilities, which will be demonstrated in the paper presentation. Altogether we have found that the TEI encoding scheme has been serving as quite a valuable tool in this kind of collaborative research practice.

- 
1. Parts of the content in this paper is based on work in progress, a forthcoming article co-authored with Prof. Donald Broady, titled "TEI markup as research tool in the prosopographic study Formation for the public sphere". In our collaborative work Prof. Broady answer for the sociological and historical content and the author of this paper for the markup, application and statistical analyses.
  2. See <http://www.tei-c.org/> and Sperberg-McQueen & Burnard
  3. See <http://www.skeptron.ilu.uu.se/broady/sc/ffo.htm> . The project is directed by Donald Broady and funded by the Bank of Sweden Tercentenary Foundation.
  4. *l'Analyse des Données*, introduced by Jean-Paul Benzecri, a geometer-statistician, in the 1960. The method is done by modelling data sets as clouds of points in multidimensional Euclidian spaces and then interpreting the data in the cloud of points (Lebart et al.). Cf. Bourdieu (1984) for applications and some explanations.
  5. See Broady for a proposed definition on Bourdieuan prosopography. See also the study on the French academic field

*Homo Academicus* Bourdieu 1984) for an example of Bourdieu's prosopography.

6. Provided that the copyright issues may be solved, there should in due time be a freely available digital version. Meanwhile the access is restricted to the researchers and for teaching purposes.
7. cf. Sperberg-McQueen and Burnard, 2002, pp. 499-516 <<http://www.tei-c.org/P4X/ND.html>>
8. For information on the *Orlando* project, documenting "the scholarly history of women's writing in the British Isles." see <<http://www.ualberta.ca/ORLANDO/>> . See also e.g. Grundy et al.

## Bibliography

- Bourdieu, P. *Homo academicus*. Paris: Minuit, 1984. English translation: *Homo Academicus*. Polity Press, Cambridge, 1988.
- Bourdieu, P. *Les règles de l'art. Genèse et structure du champ littéraire*. Paris: Seuil, 1992. English translation: *The Rules of Art. Genesis and Structure of the Literary Field*. Polity Press, Cambridge, 1996.
- Bradley, J., and H. Short. "Using Formal Structures to Create Complex Relationships: The Prosopography of the Byzantine Empire--A Case Study." Ed. K.S.B. Keats-Rohan. Oxford: Unit for Prosopographical Research, Linacre Collage, 2002. Preprint available at <<http://pigeon.cch.kcl.ac.uk/docs/papers/pbe-leeds>> .
- Broady, D. "French prosopography. Definition and suggested readings." *Poetics* 30 (2002): 381-385.
- Grundy, I., P. Clements, S. Brown, T. Butler, R. Cameron, G. Coulombe, S. Fisher, and J. Wood. "Date ChronStructs: Dynamic Chronology in the Orlando Project." *Literary and Linguistic Computing* 15:3 (2000): 265-89.
- Lebart, L., A. Salem, and L. Berry. *Exploring Textual Data*. Dordrecht: Kluwer Academic, 1998.
- Sperberg-McQueen, C.M., and L. Burnard, eds. *Guidelines for Electronic Text Encoding and Interchange (TEI P4)*. Oxford, Providence, Charlottenville, Bergen: TEI Consortium, 2002.

---

# Markup of Educational Content

---

*Monica Langerth Zetterman*

*(monica.langerth@ped.uu.se)*

*Digital Literature, Uppsala University*

---

The purpose of this presentation is to illustrate the implementation of methods and tools for separating markup, from the actual XML document. This is being done in the project Educational Content Markup<sup>1</sup> where the object is to explore and develop tools and methods for the creation and management of encoded, modularised and portable digital content to be used in educational settings.<sup>2</sup>

The underlying idea in the project is to enable the production, use and reuse of document content in different fields of application and to support various groups of users and end-users. Concerning the fields of application our focus is on teaching and learning in higher education and the production of academic research publications. Hence, one group of users are teachers creating e.g. collections of documents to be used in classes. We picture both students and teachers as a group of end-users. Another group of users are scholars and other content providers concerned with document creation and the production process.

In this project we have chosen to differentiate two kinds of markup in order to distinguish fields of application and groups of users. Thus, the differentiation described below guides our exploration of methods and the development of tools.

- Original/internal markup: Markup placed within the document - as opposed to stand-off markup- in this context often capturing the structure of a document or some specific features such as personal names. Either the markup was already added to the document by another part or the markup is created during the document creation process. Markup is automatically added by controlled input using templates in the DiVA system<sup>3</sup> or semi-automatically by the content producer.
- Additional markup: Markup added after the creation process which can be done with optional annotation techniques such as *Annotea* or the *TEI* guidelines<sup>4</sup>. The additional markup might be added and stored internally in the document, or, annotations might be added and stored externally in separate XML-files, not altering the any markup internally added to the document. For example *Annotea* allows annotations to be attached to resources without modifying the original/internal markup.

It is important to point out that our two views of markup can be, and often are different, but this does not mean that the methods are non-exclusive. However, a markup scheme designed to catch in-depth descriptions of a chosen aspect, rather than describing the internal structure of a document, serves a special purpose and therefore anticipates a different data model than a structured oriented scheme. That is one reason why we, in this project, think it is useful to separate the additional markup from the original/internal markup.

Other reasons to separate different kinds of markup have to do with different kinds of uses and user-groups. By various means of organising, filtering and presentation, one and the same content might be utilized in a range of subject areas and for different educational levels, from primary to higher education or in-service and in-company training. In trying to accomplish this, several different XML techniques and tools are being used for the implementation of a prototype of an educational content management system, currently consisting of three related parts.

1. In one part of the project, a prototype provides means to take the descriptive markup added to the internal XML markup and treat it as external annotations. The browser is used as the markup client, using *Annotea* protocol both on the client side and both on the server side. At the backend a RDF<sup>5</sup> database is used as the content store. This architecture allows usage of arbitrary application profiles and an ability to integrate the information with other tools.<sup>6</sup>
2. In the second part we explore tools and schemes for adding external markup - stand-off markup - to documents already annotated with TEI/XML markup of personal names, dates and places. Here we will test schemas providing uses of authority files such as the first draft of *MADS* which is designed for description of agents (e.g. people, organisations) and terms (e.g. topics, geographical places).
3. In the third part we are exploring applicable methods and tools for adding markup when authoring. The idea is to integrate the process into the author's natural workflow such as the software they use for text editing (e.g. *OpenOffice* or *MS Office*). These kinds of tools should support and facilitate work performed by the author during the production and revision of texts.

This poster will provide demonstration of a prototype tool implementation and uses, examples of markup, and a description of current status in the project progress. We are looking forward to get a chance to exchange ideas and well-informed feedback on the underlying ideas, choice of methods and implementation of tools in this project.

---

1. The project Educational Content Markup is a joint project and the partners are: The research program Digital Literature and the

Electronic Publishing Centre, both Uppsala University. The Swedish Royal Library, the National Agency for Education in Sweden, Ekelunds Publishing Company and Center for IT in Northern Sweden. The project is funded by the Swedish Agency for Innovation Systems. The author of this poster presentation is co-ordinating the project. For participants and further information, see <http://www.skeptron.ilu.uu.se/broadly/dl/mu.htm> .

2. For further information on the research program Digital Literature publications and research areas, see: <http://www.skeptron.ilu.uu.se/broadly/dl/> .
3. *DiVA* is the Swedish acronym for *Academic Archive Online*. For further information on the *DiVA* system and related publications, see <http://publications.uu.se/epcentre/projects.xsql> , for the technical report, see: [http://publications.uu.se/epcentre/diverse/hardware\\_software.pdf](http://publications.uu.se/epcentre/diverse/hardware_software.pdf) .
4. See <http://www.tei-c.org/Activities/SO/sow06.html> .
5. Resource Description Framework, see <http://www.w3.org/RDF/> .
6. For a technical report on the prototype, see Engman.

## Bibliography

*Annotea*. Accessed 2005/03/04. <http://www.w3.org/2001/Annotea/>

Engman, J. *Handling fine-meshed filtering from archive to resource level*. Master thesis, Umeå University, 2004.

*Metadata Authority Description Schema*. Accessed 2005-03-14. <http://www.loc.gov/standards/mads/>

---

# Creating an Archives Management System at the University of Maryland Libraries

---

**Jennie A. Levine** ([levjen@umd.edu](mailto:levjen@umd.edu))

*University of Maryland*

**Amit Kumar** ([kumaramit01@gmail.com](mailto:kumaramit01@gmail.com))

*University of Illinois at Urbana-Champaign*

**Susan Schreibman** ([sschreib@umd.edu](mailto:sschreib@umd.edu))

*University of Maryland*

**Jennifer Evans** ([jennifer.evans@nara.gov](mailto:jennifer.evans@nara.gov))

*National Archives and Records Administration*

---

**E**ncoded Archival Description (EAD) is an XML-based standard used to encode archival finding aids that reflects the hierarchical nature of archival collections and that provides a structure for describing the whole of a collection, as well as its components (Pearce-Moses 2004).

Archives approach EAD in different ways. The tools available through the official EAD website at the Library of Congress (<http://www.loc.gov/ead/>), such as the EAD Cookbook, focus on making it possible for archivists to tag their existing paper/word-processed finding aid content. This approach works well for smaller institutions, but with over 400 finding aids, manual tagging was not a viable solution at the University of Maryland.

In order to help define the ideal system for our institution, the staff of the Archives and Manuscripts Department identified several main themes. Streamlining workflow was the largest focus. Finding aids are the natural end result of a series of archival procedures that begin with appraisal, selection, arrangement, and finally, description. Since these processes are related, the creation of a system to help the department streamline workflow while also simplifying and demystifying the creation of an EAD document seemed the most practical course of action. Standardization was another concern. We wanted to create a system that organized our finding aid data in such a way that a future generation could easily port it to a new system. Search capability was a third requirement. Searching across collections would greatly aid us both with our reference services and with our processing. Fourthly, we wanted to create a distinct, usable, and attractive public interface.

The department had already created a database in Microsoft Access to keep track of basic information, such as collection title, collection size, and location. As a first step, we modified this database to hold several more collection-related fields and also added a report that allowed staff to create electronic accession sheets (the first step in the record-keeping process for any collection acquired by the department).

Since much of the accession sheet information becomes part of the later EAD document, the new database tables and structure were based on the structure of EAD. Many of the fields were named after their corresponding EAD tags.

The database structure is relatively simple. A main table, named *archdescdid* after one of the main components of an EAD document, contains the bulk of the information. A handful of smaller tables tie together the *<eadheader>* information and the deeper-level descriptive information located in the *<desc>* sections of the EAD document. The biggest challenge was figuring out a way to design a table that would accurately represent the "Box Inventory" section of the paper finding aids so that data entry was simple for staff, but that would also easily convert into the EAD tag structure.

The decision to use *Microsoft Access* as the primary database was based on a number of rationale, although the primary ones were staff familiarity and widespread availability of the product within the institution. Several other institutions use web forms for entering EAD information into a database, and while this method is very flexible and allows the system to be easily shared across institutions, it would not allow the department to carry out some of the other collection management tasks.<sup>1</sup>

In the absence of a skilled programmer, *Microsoft Access* adequately served the purposes of the early phases of the project. Little to no programming expertise was necessary to create functional database forms and reports. There were, however, some weaknesses in the Microsoft Access software that put the project on hold at a crucial point: the plan to create the EAD document using the Microsoft Access report features would not work; the reports could not handle the text in large memo fields. A model for the conversion from Microsoft Access to EAD came from the Australian Heritage Document Management System, which was created by the Australian Science and Technology Heritage Center. It also used an advanced Microsoft Access database and ASTHC staff was helpful in discussing the system. After examining their approach, the University of Maryland realized that the assistance of a programmer would be needed to properly extract the data from Microsoft Access.

The Archives and Manuscripts Department thus approached Maryland Institute for Technology in the Humanities (MITH) to assist with the programming support needed, as well as the project management skills to convert the Microsoft Access

database into a series of outputs (primarily finding aids and subject guides), as well as create an online publishing system with a robust search and browse interfaces, and an administrative management system.

The software itself is comprised of two independent systems: a converter program written in Java that communicates with the Microsoft Access database using Java Database Connectivity (JDBC), and a web application with an XML Content Management System.<sup>2</sup> The web application is based on Java Servlet API with Model View Controller architecture.

The converter application creates a list of finding aids in the database and a user can click and generate the EAD-compliant XML document.

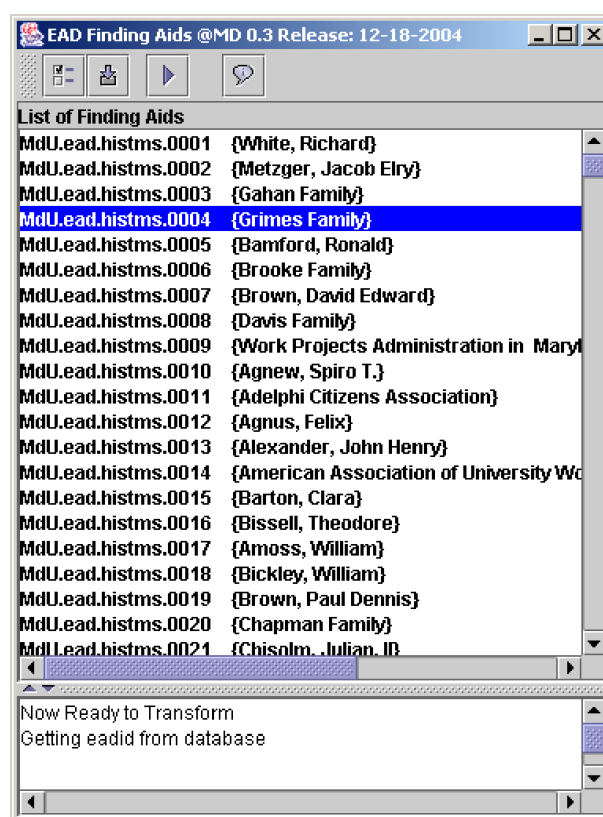


Figure 1: The converter application which transforms the finding aids from the Microsoft Access database into EAD-compliant XML.

These documents are then uploaded and indexed by the web application. The web application also generates the subject guides and finding aids using XSL style sheets.

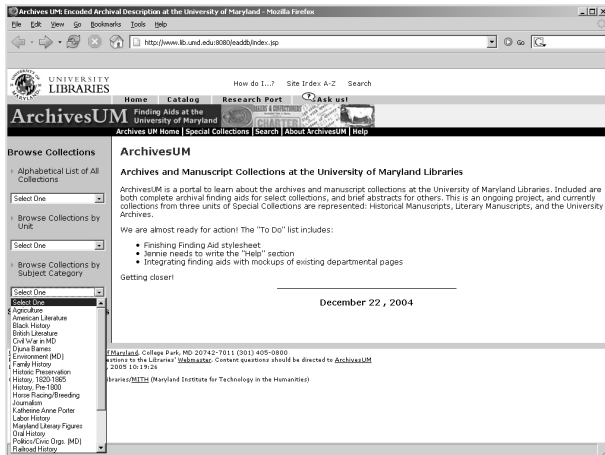


Figure 2. The home page of ArchivesUM, with a pull-down menu listing the subject guides which are generated through a combination of static HTML and dynamically-generated content from the database.

Via the administrative interface, the repository editor can upload, delete, and convert finding aids to HTML. This pre-processing of the XML document was built into the system so that the finding aids did not have to be converted to HTML at the time of the request. Figure 3 shows a result page ranked in order of relevance. It was decided that in the first instance, all collections would be represented in the database through an abstract. As finding aids are converted, they will be made available through the archive management system. As Figure 3 shows, the interface makes it clear when the finding aid is available:

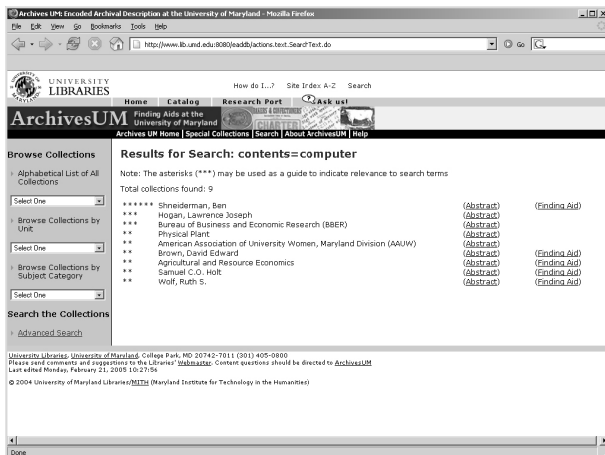


Figure 3: A ranked result page indicating which items have finding aids available.

Generating subject guides proved a greater challenge. Although it would have been easy to generate the subject guides on the fly, it was felt that these needed to be converted into static HTML pages and mounted on the Internet. Subject guides indexed by Google and other search engines has proved to be the most popular way for potential users to find the University of Maryland's archival resources. Thus, a feature was built into

the administrative interface to create the subject guides through a combination of static text and abstracts generated from the EAD document, where <abstract> tags with different "type" attributes are located.

Various nodes of the <eadheader> and <archdesc> are indexed with Lucene and a query interface is provided to search and browse the finding aid.<sup>3</sup> The use of Lucene as a search index enables compound searches for phrases in the box inventory, collection title, author, scope, and subject fields of the EAD document.

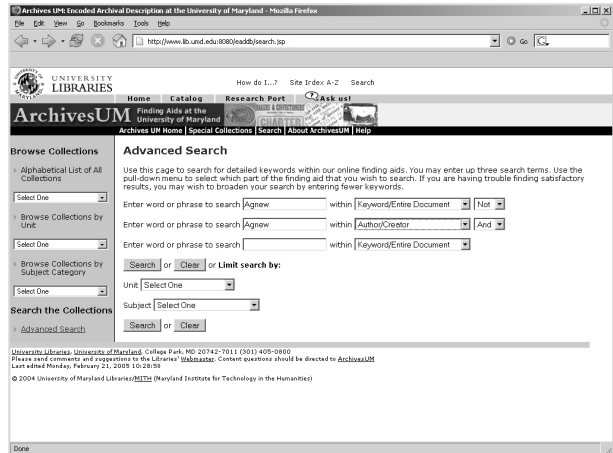


Figure 4: A search page enabling users to perform complex searches based on information in different parts of the EAD finding aid.

The Archives and Manuscripts Department staff and MITH worked together in the development of several XSLT style sheets for various parts of the website. In many ways, this proved to be the most difficult task. The hierarchical nature of the display of a finding aid made design of the final, and most important, style sheet extremely complicated. Other repositories provided examples to build from, but since the EAD of the <desc> section of a document varies widely from institution to institution, advanced customization was necessary.

The administrative interface provides an interface to upload an XSL style sheet, so that the website administrator can change the design of the finding aids and subject guides. Much of the software code for this project has been borrowed from *teiPublisher*. Moreover, although staff developed the system for use with finding aids in three of the units within the University of Maryland's Archives and Manuscripts Department, staff constructed it with the possibility that other archival units on campus could use it, as well as staff in repositories across the University of Maryland system. While each repository will have its own Microsoft Access database so it may generate reports unique to its holdings, there will be one EAD repository, which will give users unprecedented access to search across archival units and institutions in a way not possible currently.



This paper will thus address the theoretical, practical, and programming decisions that contributed to the design of this archival management system.

1. *Virginia Heritage Guides to Manuscript and Archival Collections in Virginia; Online Archive of California.*
2. JDBC <<http://java.sun.com/products/jdbc/>> .
3. Lucene <<http://jakarta.apache.org/lucene/docs/index.html>> .

## Bibliography

Dooley, Jackie M., ed. *Encoded Archival Description: Context, Theory, and Case Studies*. Chicago: Society of American Archivists, 1998.

*EAD Help Pages - Software Products*. EAD Roundtable of the Society of American Archivists, 2003. Accessed 2003-08-13. <<http://jefferson.village.virginia.edu/ead/products.html>>

*Encoded Archival Description (EAD): Official EAD Version 2002 Website*. Library of Congress, 2002. Accessed 2005-03-21. <<http://www.loc.gov/ead/>>

Feeney, Kathleen. "Retrieval of Archival Finding Aids Using World-Wide-Web Search Engines." *American Archivist* 62.2 (Fall 1999): 206-228.

*Heritage Document Management System*. Australian Science and Technology Heritage Center, 2003. Accessed 2003-04-15. <<http://www.austehc.unimelb.edu.au/HDMS/findingaids.html>>

Miller, Fredric, ed. *Arranging and Describing Archives and Manuscripts*. Chicago: Society of American Archivists, 1990.

*Online Archive of California*. University of California, 2004. Accessed 2005-03-21. <<http://www.cdlib.org/inquire/projects/oac/toolkit/>>

Pearce-Moses, Richard. "Encoded Archival Description." *A Glossary of Archival and Records Terminology*. Website: Society of American Archivists, 2004. <<http://www.archivists.org/glossary/>>

*teiPublisher*. Accessed 2005-05-19. <<http://teipublisher.sourceforge.net/docs/index.php>>

*Virginia Heritage Guides to Manuscript and Archival Collections in Virginia*. University of Virginia, 2004. Accessed 2004-11-04. <<http://www.lib.virginia.edu/vhp/index.html>>

## Story Generators: Models and Approaches for the Generation of Literary Artefacts

**Birte Lönneker**

(birte.loenneker@uni-hamburg.de)

Universität Hamburg

**Jan Christoph Meister**

(jan-c-meister@uni-hamburg.de)

Universität Hamburg

**Pablo Gervás** (pgervas@sip.ucm.es)

Universidad Complutense de Madrid

**Federico Peinado** (fpeinado@fdi.ucm.es)

Uni Madrid

**Michael Mateas** (michaelm@cc.gatech.edu)

Georgia Institute of Technology

This session is concerned with the automated creation of fiction or "literary artefacts" that might take the form of prose, poetry or drama. Special focus is placed upon those approaches that include the generation of narrative structures and therefore use some kind of story model. First attempts in automated story generation date back to the 1970s, with the implementation of Meehan's *TALE-SPIN* (1977) based on the achievement of character plans and Klein's automatic novel writer (1973/1979) that simulates the effects of generated events in the narrative universe. Currently, story generators enjoy a phase of revival, both as stand-alone systems or embedded components. Most of them make reference to an explicit model of narrative, but the approaches used are diverse: they range from story grammars in the generative vein to the conceptually inspired engagement-reflection cycle. Real-life applications include the generation of a set of plot plans for screen writers in a commercial entertainment environment, who could use the automatically created story pool as a source of inspiration, and the generation of new kinds of interactive dramas (video games).

The creation process of literary artefacts is of particular relevance to Literary and Humanities Computing. Not only does it provide methods of simulating and modelling narrative processes, but it identifies basic and combinatorial elements of story and discourse. The definition of these elements can in

turn help scholars involved in the analysis of narrative to produce annotation (mark-up) that might be re-used in the story generation models.

Research into Story Generators is by nature an interdisciplinary project and as such constitutes an exemplary case for Humanities Computing efforts that are of a more speculative kind, as opposed to application oriented approaches. While the humanities – and more particularly, narratology – are called upon to clarify and systematize basic concepts and theoretical models of narration, computer scientists and AI researchers try to translate these models into workable system architectures and processes. Accordingly, our session brings together one theoretical and two applied approaches to the generation of literary artefacts. The theoretical paper presents an attempt towards formulating an 'ideal' Story Generator Architecture based on a narratological model of story generation. The applied papers discuss the role of the story model in a stand-alone fairy tale generator (*ProtoPropp*), and the problem of story management in games, using *Façade* and other story (drama) management architectures as examples.

## "Dream on": Designing the ideal Story Generator Algorithm

**Birte Lönneker and Jan Christoph Meister**

*Deep Blue* may have beaten Kasparov at chess —but whether computers will ever be able to generate well-formed and aesthetically pleasing narratives is still subject to dispute (Bringsjord/Ferrucci; Pérez y Pérez/Sharples). Most AI researchers have come to the conclusion that the generation of natural language narratives that are both domain independent and Turing test compliant is a 'killer application': it defines the outer limit of computational creativity.

The current paper reports on our research into Story Generator Algorithms (SGAs), that is, computational systems designed to generate natural language narratives. Though attempts at designing and implementing SGAs date back to the early 1970s they have not received a lot of attention in Humanities Computing (HC) circles. The relevance of these experimental and speculative approaches seemed rather limited in the light of practical HC desiderata such as the definition of mark-up conventions, document type definitions, and standards for digital resource building, to mention but a few. However, contrary to this pragmatist line of reasoning we would like to argue that SGAs in their abstract models make explicit some of the cardinal assumptions underlying our intuitive human models of narrative, which in turn have filtered down into the practice of humanities computing in whose object domain narratives play a dominant role. Our own research methodology is therefore in part empirical - we aim to survey and classify the types of SGAs developed thus far - and in part theoretical. The current paper focuses on one of the theoretically oriented

tasks: the design of the architecture for a hypothetical 'ideal' SGA that would be able to emulate advanced, aesthetically validated human storytelling capability.

The system architecture of this ideal SGA is derived from advanced models of storytelling developed in narratology, i.e. the humanities methodology dedicated to the scientific study of narratives. Figure 1 shows the architecture with its four domains:

1. the goal domain, in which several kinds of story-telling goals are offered, for a random or user selection;
2. the knowledge domain, in which static knowledge is represented in concepts and their interrelations in an ontology, to which language-specific lexica as well as a case-base of previous and system-generated stories are related;
3. the *histoire* domain, containing three modules concerned with the question: "what happens?", or with the production of the content of the story;
4. the *discours* domain, combining two modules that aim to answer the question: how is the content presented?

The two system complexes labelled *histoire domain* and *discours domain* mirror the two main 'levels' of narratological description introduced by structuralist scholars, *histoire* and *discours* (Todorov). However, the 'level'-metaphor used by narratologists —a residue of the structuralist 'deep layer' vs. 'surface layer' dichotomy – misleadingly implies a generative hierarchy which is at the same time uni-linear and strictly bottom-up, starting at the *histoire* level. We prefer instead to use the non-hierarchical 'domain' metaphor because it is better suited to accommodate backtracking procedures. Those procedures are necessary during the generation of the final product that should eventually reflect the intertwined results of constrained operations on knowledge pertaining to both *histoire* and *discours*; therefore, the backtracking is possibly recursive and iterative within and between both domains. With respect to implemented models for the generation of natural language artefacts, this view is more in line with the engagement-reflection model of the story generator *MEXICA* (Pérez y Pérez/Sharples) than with those of the story generator *BRUTUS* (Bringsjord/Ferrucci) or of generators for technical texts (cf. Reiter), all of which basically use a unidirectional pipeline model. In the two remaining sections, we will briefly and exemplarily present key elements of our model.

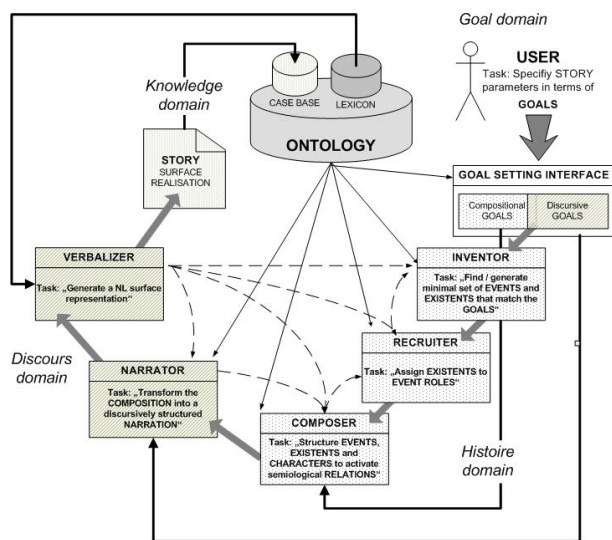


Figure 1: Architecture of an ideal SGA

### The *histoire* domain

According to a minimal consensus definition, the *histoire* (story) is a chronological sequence of narrated events, together with their participants. Some narratologists have identified actions as the constitutive elements of stories, where an action is a special type of event, intentionally caused by at least one of the participants. Often it is also claimed that in order to form a story, the events need to be causally related (cf. Rimmon-Kenan 16–17). However, narratology and related approaches have identified further types of relations between elements in the story domain that might as well contribute to storiness. For example, our previous formulation of a computational model of 'episode' (Meister) has shown that the episode is an interpretive construct of several events (four in that model) based on the activation of various semiological (semantic) relations, including contradictories.

In the system architecture shown above, events are represented as classes or frames and have at least the following slots (properties): A *follows* slot, the filler (value) of which points to the previous event; the slots *parallels* and *causes*, pointing to the respective events; the slots *involvesExistent* and *causedByExistent*, pointing to the respective participants (modelled as existents); the slot *changesStateInto*, which as a filler has an attribute-value-pair of one of the involved existents (the effect of the event is the resulting state of the affected existent); and, finally, an *isIntentional* slot to indicate the intentionality of the event. Some of the fillers are mandatory, others not; some slots can host a list of fillers. Existents like characters and objects (cf. Chatman) are also modelled as concepts with slots (attributes) and fillers (values).

The model defines further narratological notions in the *histoire* domain in terms of events and their relatedness; in particular,

*story schema*, *closed event sequence* and *story* (cf. Forster 93-94) can be defined as follows:

1. A story schema applies to a group of stories whose *histoire* is similar, for example fairy tales, legends, or detective stories. It constitutes a predefined closed set of existent classes and event classes together with their relations. The *ProtoPropp* generator presented in another paper of this session constitutes an example for a story schema based approach.
2. A closed event sequence is a temporal succession of events with a constant (sub)set of participants and a start event and end event. The temporal succession can be obtained from the fillers in the *follows* slots of the events.
3. Instead of explicitly defining what a story is, the model allows a comparison of different event sequences with respect to their storiness which is measured in terms of a) the numerical ratio between all events and action events in the considered event sequence; b) the numerical ratio between the events and the causal relations in the event sequence.

### The *discours* domain

The *discours* domain should take into account several *discours* parameters, or aspects of text description defined in structuralist narratology. Such *discours* parameters include for example the **Order** of the presented events (which might differ from the order in which they actually occur in the story) and the **Frequency** with which the same or similar event(s) are presented (Genette), or the **Mediation-of-relatedness**, to name but a few. Currently, we work with a set of twelve *discours* parameters.

Every parameter has a list of subconcepts, which represent the actual phenomena they subsume. For example, the Order parameter has the subclasses *anachrony* and *synchrony*, with *anachrony* comprising the phenomena *flashback*, *flashforward*, and *achrony*. One of the aims of the project is to identify a standard phenomenon (default) in every parameter class. For example, in unmarked narrative texts, events are most likely to be presented in their chronological order (Order-subclass: *synchrony*) and motivation relations such as 'causality' are most likely to be left implicit (Mediation-of-relatedness subclass: *implicit*).

*Discours* parameters operate on, or apply to, specific types of elements belonging to the *histoire* domain. For example, the Order parameter operates on events and event sequences. Furthermore, each *discours* parameter supplies a modification rule that states in which way it affects the story representation during the preparation of the *discours*. Thus the *flashforward* parameter states that the affected story element (event) be

moved one element (event) back in the sequence of presented events, i.e. towards the start of the story.

### Future work

Future work will include the representation of the aesthetical or communicative effect of each of the *discours* parameters and the study of application restrictions of these parameters and combinations of them. It is important to note that those effects and restrictions are not absolute, but depend on the types of events and existents and their relations used in the *histoire* domain. Therefore, the 'communication' between operations performed in both domains, described as backtracking above, will be necessary in the generation process.

While the implementation of the hypothetical 'ideal' SGA outlined in the above may seem practically impossible, it proves to be a dream that raises fundamental questions — if nothing else by challenging our HC methodologies which, by and large, have hitherto concentrated on managing static humanist data, yet shied away from tackling the conceptual threshold of dynamism and recursivity inherent in most semantic artefacts.

### Bibliography

Bringsjord, Selmer, and David A. Ferrucci. *Artificial Intelligence and Literary Creativity. Inside the Mind of BRUTUS, a Storytelling Machine*. Mahwah: Lawrence Erlbaum, 1999.

Chatman, Seymour Benjamin. *Story and discourse*. Ithaca: Cornell University Press, 1978.

Forster, Edward Morgan. *Aspects of the Novel*. London: Arnold, 1974.

Genette, Gérard. *Narrative Discourse*. Ithaca: Cornell University Press, 1980.

Meister, Jan Christoph. *Computing Action*. New York/Berlin: de Gruyter, 2003.

Reiter, Ehud. "Has a Consensus NL Generation Architecture Appeared, and is it Psycholinguistically Plausible?" *Proceedings of the Seventh International Workshop on Natural Language Generation (INLGW-1994)*. Kennebunkport, Maine, USA, 1994. 163–170.

Rimmon-Kenan, Shlomith. *Narrative Fiction*. Contemporary Poetics. London/New York: Routledge, 1983.

Todorov, Tzvetan. "Les catégories du récit littéraire." *Communications* 8 (1966): 125–151.

## A generative and case-based implementation of Proppian morphology

Federico Peinado and Pablo Gervás

Automatic construction of story plots has always been a longed-for utopian dream in the entertainment industry, specially in the more commercial genres that are fueled by a large number of story plots with only a medium threshold on plot quality, such as TV series or video games. Although few professionals would contemplate full automation of the creative processes involved in plot writing, many would certainly welcome a fast prototyping tool that could produce a large number of acceptable plots involving a given set of initial circumstances or restrictions on the kind of characters that should be involved. Such a collection of plots might provide inspiration, initiate new ideas, or possibly even include a few plot sketches worthy of revision. Subsequent selection and revision of these plot sketches by professional screen writers could produce revised, fully human-authored valid plots. By making such a collection of tentative plots available to company screen writers, a smaller number of writers might be able to provide the material needed to keep the technical teams in work.

*ProtoPropp* is a software application that can generate new stories. The domain of the application is the Russian Folk Tale, and we have used Vladimir Propp's formalization of the structure of such tales (Propp) as a guideline for the generation process.

The particular method employed for plot generation is case-based reasoning (Riesbeck & Schank). Case-based reasoning (CBR) is an AI method that relies on reusing solutions to problems solved in the past to solve problems in the present. The concept of case identifies a pair formed by a problem and its solution. The main idea is to store past problems and their associated solutions as cases, in what is called a *case base*. Whenever a new problem has to be solved, the case base is searched for similar problems faced in the past (retrieval step), and the solution of the best matching past case is adapted to the new problem (adaptation step), to account for differences between the new problem and the old one. The full CBR cycle involves two additional steps (revision and update) which provide the means for enriching the case base with the results of solving new problems using this method, but those steps are not relevant to the current endeavour. Traditional CBR is very useful in domains where the information to be handled is too complex to model explicitly, and in which there is an easily accessible store of previously solved problems and their solutions. The ProtoPropp application considers a plot generation problem in terms of Proppian functions, and a solution to that problem in terms of the assignment of a conceptual representation of the plot of a story. This involves transcribing existing folk tales into conceptual representations

of their contents and associating them with elements of Proppian morphology. This is done by resorting to a formalized knowledge base of concepts, organised into a taxonomy, which explicitly includes the relations between them. Such a knowledge base, following current terminology in AI, is referred to as an *ontology* (Gruber 1994). The use of explicit conceptual knowledge to guide the CBR process characterises *Knowledge Intensive CBR* (Díaz-Agudo & González-Calero 2003).

In this project we propose a Knowledge Intensive Case-Based Reasoning (KI-CBR) approach to the problem of generating story plots from a case base of existing stories analyzed in terms of Proppian functions. A case-based reasoning process is defined to generate plots from a user query, with two important phases: retrieval of old stories, and adaptation to build a new one. The user query specifies an initial setting for the story, and the ontology is used to measure during the generation process the semantic distance between concepts specified by the user and those that appear in the texts.

**ProppianOnto: the formalized knowledge**

ProppianOnto is the name of the ontology developed to implement the formalization of Propp's theory (see Figure 1). It is built using OWL description logic according to current standards (Bechhofer et al. 2004). It includes concepts like PlotCase, Character or ProppianFunction that our system needs to reason about stories.

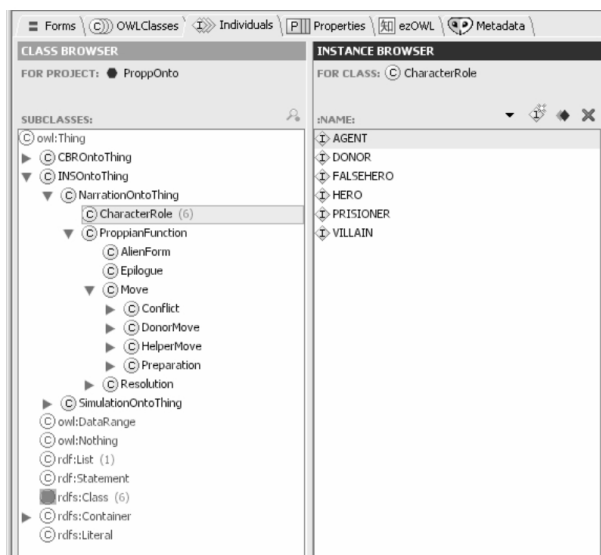


Figure 1: Implementation of ProppianOnto

The cases are plots taken from the Alexander Afanasiev's collection of Russian Fairy Tales. According to our interpretation of Propp, a plot case can have structures of two types.

PlotCase	PlotCase
1..n Character	1..n Character
1..n MoveCase	1 AlienForm
1 Resolution	0..1 Epilogue
0..1 Epilogue	

The elements in these structures can themselves have structure. The structure of a move case and a Proppian function are shown below.

MoveCase	ProppianFunction
1..n Character	1..n Character
1..n ProppianFunction	1..n Event
	1..n Place
	1..n Item
	before: 0..1 ProppianFunction
	after: 0..1 ProppianFunction

The structure of resolutions, epilogues and alien forms are similar to the move case.

**ProtoPropp: the generation process**

The generation of a new tale starts with a query presented by the user which represents the constraints that the desired tale should fulfil. The query is composed by filling in a form created which has six boxes for the main character roles (hero, villain, prisoner, donor, helper and false hero) and five boxes for the main Proppian functions (preparation, conflict, donor move, helper move and resolution), as it can be seen in Figure 2.

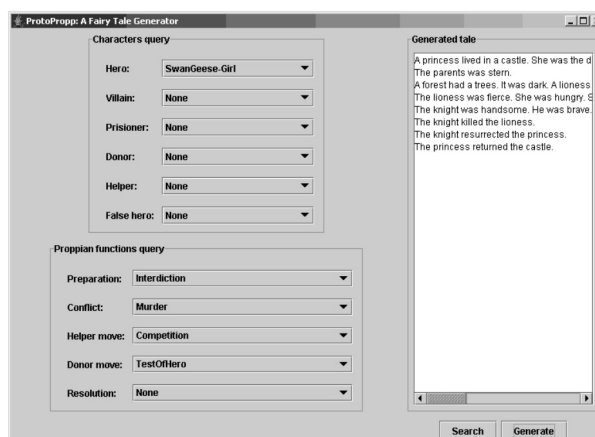


Figure 2: Author interface of ProtoPropp

Based on this query, the system attempts to find the most similar tales in the case base.

Because of the hierarchical structure of the conceptual taxonomy of the ontology, the given constraints can be fulfilled either in terms of strict similarity — the requested concept occurs in the tale — or in terms of conceptual similarity - either a specialization or an abstraction of the requested concept occurs in the tale.

1. *There is (at least) one retrieved tale with all the constraints in the query.* The system gives the option of adding more constraints.
2. *There are no retrieved tales fulfilling all the given constraints but at least one retrieved tale fulfils some of them.* The system selects the tale fulfilling the highest number of stated constraints. A second retrieval process is initiated, searching for a *different tale* satisfying the failed constraints. In this second process, the system can relax the similarity to ensure that some tale is found. That tale is mixed with the first retrieved tale to create a new one, adding or changing restrictions to the first tale, depending on the inter-relations between Proppian functions and other logical considerations, and generating from scratch the elements that the system was not able to find in the case base.
3. *There are no retrieved tales satisfying any of the query restrictions.* The system offers the possibility of adding more restrictions or changing the current restrictions to create a new story.

Whenever more restrictions are needed, the system always has the default option of filling them in randomly.

The resulting tale is obtained in an 'abstract form', as a conceptual representation of the ingredients that make it up. The last step is to use a natural language generator - cFROGS (García-Ibáñez) - to convert this conceptual representation to a simple text that can be easily read by the user.

### Conclusions and future work

This system shows how narrative structure theories can be implemented in a computational generator of stories. It is a long road, but every step we take will help us to learn more about the way stories are built. The current version of the prototype is restricted to single move plot cases. Complex stories made of more than one move case need more control over the dependencies between structural elements, and their construction should be directed by appropriate inference over the concepts represented explicitly within the system. For this strategy to succeed, the existing knowledge base must be extended to include any narratological concepts that may play a role in the adequate resolution of any conflicts arising from the existence of complex dependencies between the structural elements of a plot.

## Bibliography

- Bechhofer, S., et al. *OWL Web Ontology Language Reference*. W3C, 2004. Accessed 2005-04-06. <<http://www.w3.org/TR/2004/REC-owl-ref-20040210/>>
- Díaz-Agudo, B., and P.A. González-Calero. "Knowledge Intensive CBR through Ontologies." *Expert Update* 6.1 (2003): 44-54.
- García Ibáñez, C., et al. "Una arquitectura software para el desarrollo de aplicaciones de generación de lenguaje natural." *Procesamiento de Lenguaje Natural* 33 (2004): 111-118.
- Gruber, T.R. "Towards Principles for the Design of Ontologies Used for Knowledge Sharing." *Formal Ontology in Conceptual Analysis and Knowledge Representation*. Ed. N. Guarino and R. Poli. n.p.: Kluwer, 1994. n. pag..
- Propp, V. *Morphology of the Folktale*. Austin and London: University of Texas Press, 1968.
- Riesbeck, C.K., and R.C. Schank. *Inside Case-Based Reasoning*. Evanston, Illinois: Lawrence Erlbaum, 1989.

## Beyond Story Graphs: Story Management in Game Worlds

### Michael Mateas

Bringing truly interactive story structures to computer games is a hotly debated topic within the worlds of computer game design and academic game studies. For some designers and theorists, interactive story worlds are a holy grail of game design (e.g. Murray, Crawford), while for others narrative is antithetical to interactive experiences, destroying the high-agency, procedural potential of games (e.g. Eskelinen, Frasca). The heart of the tension between games and narrative lies in player agency. A player is said to have agency when she can form intentions with respect to the experience, take action with respect to those intentions, and interpret responses in terms of the action and intentions. Those who argue against narrative games point to the predetermined or predestined nature of narrative; strong narrative structures have complex sequences of cause and effect, complex character relationships and sequences of character interactions. Since player interaction can at any moment disrupt this narrative structure, the only way to maintain the structure is to remove or severely limit the player's ability to effect the structure. This eliminates so-called 'global' agency, forcing the player down a predetermined path. Thus ludologists argue that if narrative must inevitably mean a diminishment in player agency, it should not be used in game design.

Contemporary games do seem to support the ludologist position. In all contemporary story-based games the story structure is

completely fixed, or has an extremely simple branching structure. The player has local agency, that is, can move around the environment and interact with objects and non-player characters, but the narrative structure is a linear sequence of cut scenes unlocked during the gameplay. In order to provide global narrative agency, computational and design methods must be devised that can incorporate player interaction into large scale story structures. This problem can be best understood by contrasting it with story graphs.

The standard 'best practice' in interactive narrative is the story graph, where each node represents a story event and each arc represents player actions. In a story graph the author has manually unwound all possible paths through the narrative space. However, the manual authoring overhead of story graphs means that, in practice, they tend to have a very small branching factor (quasi-linear) and a small number of nodes (limited story-level variation), leading precisely to the lack of global (story) agency evident in the contemporary game scene.

### Story management

The alternative to the story graph is story management. The first story manager was proposed by Brenda Laurel in her thesis on interactive drama (Laurel), and further developed by AI research groups exploring interactive narrative. A story manager replaces the graph structure with a policy for story event selection. The author still creates the nodes of the story graph, where nodes represent story events such as scenes or individual character actions (depending on the granularity of global agency). However, rather than manually linking the nodes, that author instead creates a selection policy for story events; story events are activated as a function of the history of the story so far and the actions performed by the player. The story policy implicitly defines a story graph; theoretically, one can imagine unrolling the policy into a graph by recording the story function's response to all possible inputs (story histories + player action). The whole point of the story management approach, however, is to keep the graph implicit. By implicitly specifying graphs via a story policy, authors can create interactive stories that would be impractical to explicitly specify as graphs, and can thus create experiences with rich global agency.

In order to define a story policy, the author must specify:

- A representation of the desired story. In order for the story policy to select the next story event, it will need some model of the desired story (what a good story looks like within the domain) so as to decide what direction the story should move in given the story history plus player actions.
- A collection of story events. The story events may correspond to discrete units of storyness, such as scenes or dramatic beats, or may be more abstract story moves that

manipulate the world in such a way as to make a desirable story happen in the future.

- A function that, given a model of the desired story, the story history, and the player actions, selects a story event.
- When story event selection happens. In general, a game world presents the player with a continuous, real-time experience while story guidance only happens at discrete points. This presents the design problem of deciding when guidance should happen.

### Example story managers

In this paper I survey three approaches to story management, the beat-based drama manager of the interactive drama *Façade* (Mateas & Stern 2003), Magerko and Laird's IDA (2003), and the search based drama manager (SBDM) first defined by Bates and Weyhrauch (1992; 1997), describing the different design decisions made by each approach with respect to the four design questions above.

In the *Façade* drama manager the story events are inspired by dramatic beats (McKee), the smallest units of dramatic value change. The desired story is modeled by one or more story value arcs (in *Façade*, the tension story value), and by declarative knowledge represented on each beat. This declarative knowledge includes:

- one or more preconditions, tests over facts pertaining to the episodic memory of the story-so-far that must be true for the beat to be potentially selectable;
- one or more priority tests that, given a satisfied precondition, boost the importance of a beat being selected;
- one or more weight tests that, given a satisfied precondition and highest-priority, boost the probability of a beat being selected;
- one or more effects that describe how the beat, assuming it is successfully executed in the world, will change the story values.

This knowledge, plus the desired story value arc(s), is used to compute a probability distribution over possible next beats; beats are selected by drawing from this changing distribution. When a beat is selected it activates a collection of behaviors that support the autonomous characters in carrying out the beat. These character-specific behaviors, which model the intentional structure of the characters, are written so as incorporate the player's moment-by-moment activity into the performance of the beat. If the player's activity deviates too far from the context assumed by the beat, the beat is aborted and a new one selected. Beat selection occurs on beat success (the beat successfully accomplishes the drama value change) or failure (the player's activity violates the beat context).

In SBDM, a player's concrete experience in the world is captured by a sequence of Player Moves, abstract plot points that a player's activity can cause to happen. A single Player Move may encapsulate 5 or 10 minutes of concrete player activity in the world - moving around, picking up objects, interacting with characters and so forth. When the concrete activity accomplishes a plot point, then a Player Move is recognized. A SBDM has a set of System Moves available that can materially alter the world (e.g. move objects around, change goals in characters' heads, etc.) in such a way as to encourage or obviate a Player Move. System Moves give the SBDM a way to warp the world around the player so as to make certain Player Moves more or less likely. Besides the System Moves, the author also provides the SBDM with a story-specific evaluation function that, given a complete sequence of Player and System Moves, returns a number indicating the "goodness" of the story. Whenever the drama manager recognizes a Player Move occurring in the world, it projects all possible future histories of Player and System moves, evaluates the resulting total histories with the evaluation function, and backs these evaluations up the search tree (in a manner similar to game-tree search) to decide which system move to make next that is most likely to cause a good total story to happen.

### **The future of story management**

Story management virtualizes the links of a story graph; while the nodes of the graph must be authored, the possible paths through the graph remain implicit. The future of story management is to virtualize the nodes of the story graph as well; the nodes (story events) will be dynamically constructed as needed. The challenge will be to adapt algorithmic story generators to incorporate interaction. All artificial intelligence-based models of story generation, including story grammars, character modeling, and author modeling, assume that all elements of the story are under the complete control of the generator. In interactive narrative, however, the player can perform actions at any time that may compromise the current causal structures established by the generator. In the context of story management, generation must be able to dynamically adapt to player action.

### **Bibliography**

Bates, J. "Virtual Reality, Art, and Entertainment." *Presence: The Journal of Teleoperators and Virtual Environments* 1.1 (1992): 133-138.

Cavazza, M., F. Charles, and S.J. Mead. "Sex, Lies and Videogames: an Interactive Storytelling Prototype." *Proceedings of the AAAI 2002 Symposium on Artificial Intelligence and Interactive Entertainment*. Ed. K. Forbus and M.S. El-Nasr. 2002. 13-17.

Crawford, C. *Chris Crawford on Interactive Storytelling*. n.p.: New Riders, 2004.

Eskelinen, M. "Towards Computer Game Studies." *Proceedings of SIGGRAPH 2001, Art Gallery, Art and Culture Papers*. Ed. K. Forbus and M.S. El-Nasr. 2001. 83-87.

Frasca, G. "Ludologists love stories too: notes from a debate that never took place." *Proceedings of Level Up 2003, Utrecht, The Netherlands*. 2003. Accessed 2005-04-06. <[http://ludology.org/articles/Frasca\\_LevelUp2003.pdf](http://ludology.org/articles/Frasca_LevelUp2003.pdf)>

Lamstein, A., and M. Mateas. "Search Based Drama Management." *Working Notes of the AAAI Workshop Challenges in Game AI, AAAI*. 2004.

Laurel, B. *Towards the Design of a Computer-Based Interactive Fantasy System*. Ph.D. Dissertation., The Ohio State University, 1986.

Magerko, B., and J. Laifo. "Building an Interactive Drama Architecture with a High Degree of Interactivity." Paper delivered at the 1st International Conference on Technologies for Interactive Digital Storytelling and Entertainment (= TIDSE '03), Darmstadt, Germany, March 2003. 2003.

Mateas, M., and A. Stern. "A Behavior Language for Story-Based Believable Agents." *Working notes of the Artificial Intelligence and Interactive Entertainment Symposium, AAAI Spring Symposium Series*. AAAI Press, 2002.

Mateas, M., and A. Stern. "Integrating plot, character and natural language processing in the interactive drama Façade." Paper delivered at the 1st International Conference on Technologies for Interactive Digital Storytelling and Entertainment (= TIDSE '03), Darmstadt, Germany, March 2003. 2003.

Mateas, M. "An Oz-Centric Review of Interactive Drama and Believable Agents." *AI Today: Recent Trends and Developments*. (= *LNAI 1600*). Ed. M. Wooldridge and M. Veloso. Berlin, New York: Springer, 1999. n. pag..

McKee, R. *Story: Substance, Structure, Style, and the Principles of Screenwriting*. New York, NY: HarperCollins, 1997.

Murray, J. *Hamlet on the Holodeck*. Cambridge, MA: MIT Press, 1998.

Pérez y Pérez, Rafael, and Mike Sharples. "Three computer-based models of storytelling: BRUTUS, MINSTREL and MEXICA." *Knowledge-Based Systems* 17 (2004): 15-29.

Weyhrauch, P. *Guiding Interactive Drama*. Ph.D. Dissertation, Tech report CMU-CS-97-109, Carnegie Mellon University, 1997.



## Mysteries in Time and Space: Historical Computing

---

**John Lutz** (*jlutz@uvic.ca*)

*History University of Victoria*

**Patrick Dunae** (*dunae@cliomedia.ca*)

*Malaspina University College*

**John Bonnett** (*John.Bonnett@nrc-cnrc.gc.ca*)

*National Research Council of Canada*

---

This session highlights three of the most innovative projects in Canadian historical computing, all aimed at deploying the new technologies new pedagogical ideas to allow us to teach humanities and history in ways we have never been able to teach before. The session is organized by the Canadian Committee on History and Computing, a committee of the Canadian Historical Association and the leading body in historical computing in Canada.

The three papers in the session highlight different aspects of the history-technology interface: one involving 3-D reconstructions, the second the linking of historical maps and census information to contemporary GIS based maps, and the third the creation of virtual archives around historic crimes aimed at a new pedagogy of teaching history.

### History viewed from the side: Future Directions for Historical Representations Using 3D Environments

**John Bonnett, Ph.D.**

Historians, as a rule, make bad futurists. Their object of concern is the past. And more to the point, their experience suggests that would-be prophets more often get things wrong than right. That being said, in the context of history and computing, there is a case to be made that scholars should consider how current changes in communication technologies will transform future practice in the historical discipline. The purpose of this paper is to argue that if historians mean to appropriate the computer, and specifically 3D objects and 3D environments as media for representation, they will have to pay a specific price. They will need to revisit the aesthetics of their discipline. They will need to devise conventions to govern narration, representation and documentation for 3D-immersive environments, specifically virtual reality and augmented reality. This paper will explore

two aesthetic innovations that potentially might be applied in 3D environments.

The first innovation is the *sideshadow*, a literary device that can be used to communicate that systems such as cities have the potential to evolve in multiple directions, while only selecting one. The sideshadow is a representation that complements the representation of history as it was. Its purpose is to suggest the possibility of alternate histories, paths of growth and change that could have been visited by a city, but were not. The second innovation is the *siderepresentation*. Its purpose is to communicate that objects contained in a virtual environment are the product of interpretation. Different individuals will have different opinions on how a given object — such as a building — should be represented in a virtual environment. Conventions will need to be devised to communicate when a given object is a point of contention, and enable exploration of the differing perspectives available to them in a given virtual space.

John Bonnett is a Research Officer with the National Research Council of Canada. A historian by training, he recently completed a thesis devoted to the writings of Harold Innis, the Canadian communication theorist. He is also the principle developer of the *3D Virtual Buildings Project*, an initiative designed to enable students to generate models of historic environments using 3D software, and to develop critical thinking skills via model construction. He is the chair of the Canadian Historical Association's committee for history and computing.

### Mystifying History: the Great Mysteries Project

**John Lutz**

The paper describes the marriage of the new pedagogy to the potential of the new technology to create virtual archives around a single event or subject. It focuses on the work of the *Great Unsolved Mysteries in Canadian History Project* which has created three virtual archives of historic documents and images, each focused on a compelling murder, and proposes to create 10 more.

The project has taken the innovative approach that history can and should be fun and internet technology can help make it so. It is not just that each of the websites uses the near universal interest in the mysterious and the macabre to draw students in. What is novel about this project is a teaching method which focuses on giving students the skills to be historians, while providing them with the context needed to make historical judgments. The standard method for teaching history has tended to make students passive consumers of a history which others have put together for them. Not only is this usually uninteresting but students do not acquire the historical skills that would allow

them to tell good from bad history: assessing credibility, critical reading, reading for bias, or creating a narrative.

This involves giving students access to archival and other historical records compiled from the main archival repositories, transcribed and translated, so they avoid the most frustrating activity of reading archaic, faded writing and are able to engage in their own language. Previously, only professional historians could justify the time and resources to travel to several archives to compile the necessary documentary record.

The project draws on the new historical pedagogy, sometimes called *active learning* or *Document Based Inquiry*, aimed at making students historian-detectives. It is a staged approach to teaching the skills of an historian as early as middle school and making the tasks more complex through high school and to university. The session title "Mystifying History" is ironic since the main point of using historical mysteries is that through the detective work needed to solve the crimes, the project de-mystifies the complex skills that go into creating a historical narration and argument.

Each virtual archives houses approximately 250 unique documents, 100,000-125,000 words when transcribed (the equivalent of a scholarly book), 100 photographs, 10 maps and the next phase will include 3-dimensional reconstructions of a component of the historical landscape. The project is already in use in over 300 high schools in Canada, the United States, Australia, Germany, Great Britain, and elsewhere. Teachers' guides are available and the websites are available in French and English. They are provided to schools totally free of charge.

John Lutz is an associate professor of history at the University of Victoria and is co-director of the *Great Unsolved Mysteries in Canadian History* project. Phase One of the project has won two prestigious North American Awards, the NAWEB and MERLOT awards.

## **Virtual Victoria. Visualizing a Victorian city with digital maps, views and GIS**

**Patrick A. Dunae**

### **Abstract**

What did a late 19th century Canadian city look like? What did it feel like? How was residential and commercial space utilized? We can address these questions in a computer-mediated application that allows us to visualize a Victorian city. In the prototype described in this paper, we focused on Victoria, British Columbia, circa 1891.

The application comprises several components. It utilizes attribute data derived from census records, tax rolls and street directories. It features digital image maps based on bird's eye

views and panoramic photographs of the city in 1891. The image maps connect streets, buildings and activities shown in the prints and photographs to our attribute data. The data and the images provide the foundation for an historical GIS of the city. Using these resources along with video, photo imaging and 3-D modeling software, we have created scenarios where users can 'fly over' the city and to zoom in and out of 1891 streetscapes. The scenarios and our GIS layer will be available at web site called *Virtual Victoria*. At the *Virtual Victoria* web site, students, researchers and the general public can visualize and interact with a late Victorian city.

### **Introduction**

In the early 1960s, G. M. Young, the distinguished historian of Victorian England, famously exhorted students to "read until you can hear people talking." He meant, of course, that students could best connect intellectually and emotionally to the past by immersing themselves in the literature of the period. Historians of Victorian Canada encourage students in similar ways. We enjoin students to read contemporary novels, newspapers and memoirs as a way of connecting with the past. Like Young, we want students to be so acutely attuned that they can hear people from the past talking.

But we have an advantage over history practitioners forty years ago. We have the advantage of digital convergence. Empowered by multimedia and other computer technologies, we can see the Victorians more clearly than ever before. Using archival records and new technologies, we can create a milieu where students can visualize a Victorian city and an historical environment where perceptive students can "hear people talking." This is the promise and allure of humanities computing for the historian.

We are exceptionally well positioned in this city, because we have broad foundation to build on. In terms of digital historical data, Victoria may be one of the best-documented cities in Canada. Thanks to an initiative started about ten years ago at the University of Victoria and Malaspina University-College, we have an extraordinarily rich dataset to work with. We have a 100% sample of the nominal census records, plus digital versions of city directories and tax assessment rolls from 1881 to 1901. These records are readily accessible on our Vancouver Island web site at vi — an abbreviation for "Vancouver Island" — history. The *vihistory* web site is located at <http://vihistory.ca/>.

Historical records available on the *vihistory.ca* web site comprise the attribute data for this application. But while our study is grounded with those records, I will start our discussion from a different perspective — from the air.

## I. Visualizing the City with Lithographic Views

Panoramic lithographic views — also known as panoramic maps and 'bird's eye views' — are one of the best sources available to historians who want to visualize, understand and represent nineteenth century cities in Canada. Of course, the views were idealized and somewhat fanciful. In the main, however, panoramic views provide a very good representation of urban space. The 1889 panoramic view of Victoria is remarkably accurate in its depiction of the city and so is a valuable component in this project. We have created an effect that enables viewers to connect the lithograph with a sequence of contemporary photographs taken from the top Victoria's highest building in 1890. We have also linked the images to our dataset of census records and street directories, thus creating an interactive faux GIS.

## II. Visualizing & Populating the Victorian City with GIS

Urban historians and historical geographers face another challenge. We not only want to visualize the Victorian city, we want to populate the Victorian city. We want to populate it with the residents who actually lived and worked in the urban spaces we are endeavouring to re-create. In the second half of this presentation, I will discuss the power and potential of a full — rather than a faux — GIS application. I will describe briefly some exciting projects already developed by urban historians and historical geographers in the United States and Great Britain. I will demonstrate how we are building an historical GIS of Victoria using historical census records and digital cadastral layer of the modern city of Victoria. With technical assistance from GIS professionals in the Capital Regional District of Victoria and consultants from ESRI Canada, we have started on a prototype that has immense potential for students and researchers. In this part of my presentation deal with datasets and simulations generated by *ArcGIS* and a 3-D modeling program, *CommunityViz*.

The presentation will conclude by connecting 21st century 3-D models with the isometric models seen in Victorian lithographs. In my closing remarks, I will emphasize the value of utilizing archival records with new technologies, and the promise of research methods and digital applications derived from other disciplines.

---

# Reaching Out: What do Scholars Want from Electronic Resources?

---

*Shawn Martin* ([shawnmar@umich.edu](mailto:shawnmar@umich.edu))

*University of Michigan*

---

**T**he potentials for teaching and learning using technology are tremendous. Now, more than ever before, computers have the ability to spread scholarship around the globe, teach students with new methodologies, and engage with primary resources in ways previously unimaginable. The interest among humanities computing scholars has also grown. In fact at ACH/ALLC last year, Claire Warwick and Ray Siemens et al. gave some excellent papers on the humanities scholar and humanities computing in the 21st century. Additionally, in the most recent version of *College and Research Libraries* (September 2004), a survey was conducted specifically among historians to determine what electronic resources they use. The interest in this is obviously growing, and the University of Michigan as both a producer of large digital projects as well as a user of such resources is an interesting testing ground for this kind of survey data. Theoretically, Michigan should be a potential model for high usage and innovative research and teaching. In many cases it is; nevertheless, when one looks at the use of electronic resources in the humanities across campus and their use in both the classroom and innovative research, it is not what it could be. The same is true at other universities. At many universities across the U.S. and Canada, including those with similar large scale digitization efforts, use remains relatively low and new potentials of electronic resources remain untapped. Why?

During 2004 and 2005, the *Text Creation Partnership* (TCP) project, one of the largest digital projects at the University of Michigan, has undertaken several studies to answer that exact question, specifically for its own resources but also including other similar projects. Rather than asking the question of what the humanities scholar is and wants, TCP has sought to answer what is inhibiting use of current resources and what can the community do to enhance their experience with already existing resources. With the help of the School of Information and Departments of English and History at the University of Michigan as well as the cooperation of librarians and scholars throughout the US, UK, and Canada, TCP has set out to determine how its databases and other similar resources are being used, what potentials scholars could see for use, what are

the barriers inhibiting use, and what the community can do to reach out to those scholars who may not have used electronic technology as much as they perhaps could. The results have been divergent and quite interesting.

So far in this continuing study, the TCP has tried to cover as many bases as possible. Project staff and students at relevant departments have interviewed many relevant scholars in the field (both those who have used technology and those who have not), sent out surveys to faculty and students, held a focus group of University of Michigan faculty and two with faculty from outside the University of Michigan, completed interviews with faculty throughout the U.S. and Canada, created sample syllabi and educational materials for scholarly review, and surveyed the use of electronic resources and citations within scholarly literature. Responses have ranged tremendously. Some scholars see interface as the primary concern; such resources are not designed to do the kind of search they want. Others see selection as a problem; the materials that databases choose to select are too narrow to be of use to scholars outside of that field or are too broad and produce too many results. Still others question the legitimacy of the source itself. How can an electronic copy be as good as seeing the original in a library? Other, more electronically oriented scholars, see the great value of accessibility of these resources, but are unaware of the added potential for research and teaching. The most common concern, however, is that scholars believe they would use these resources if they knew they existed. Many are unaware that their library subscribes to resources or that universities are sponsoring this kind of research. Others feel that there is no incentive within the university system for scholars to use these kinds of new resources. In all, the humanities computing community has a great deal to do to facilitate further use.

What kinds of things should the community do? Some answers are as simple as new interface tools and methods of interacting with the database itself. Yet, even more fundamentally, many feel that the community needs to raise awareness of these resources and create incentives to use them. Librarians can certainly be a part of this in helping to raise awareness among their local faculty. Faculty also need to be involved by raising awareness beyond their own community to those colleagues who may not be as aware of the potential for these kinds of resources. Some faculty suggested that the community work together to create even more grants, contests, or prizes to encourage innovative electronic publication and research. Others have suggested that, given time, researchers will realize the potential and use it.

These responses raise several questions. How should designers of electronic resources structure their databases to maximize use? How should librarians, scholars, and users of electronic resources reach out to the community to increase use? What obstacles are there in increasing use of technology in the

classroom or in scholarly research? What role will such sources play in the future? How will this change the study of the humanities? What benefits or detriments does it bring to the profession? What role does or should the humanities computing community play in all of this? By analyzing the data to these questions and by gaining insights from others, it is hoped that we can supplement and continue an ongoing dialog about what the community is, what it needs, and what should be doing either to change current practices or to enhance already existing ones. In all, there would seem to be much work to do to increase the potential of electronic resources in the humanities, and the questions and answers brought up in these surveys may help to focus thinking on many aspects of the profession.<sup>1</sup>

- 
1. I would also like to thank Julia Gardner and Kathy Schroeder for conducting interviews among scholars and working on the survey materials for this project.

## Bibliography

- Blouin, Francis X. "History and Memory: The Problem of Archive." *Publications of the Modern Language Association of America* 119.2 (March 2004): 296-298.
- Dalton, Margaret Stieg, and Laurie Charnigo. "Historians and Their Information Sources." *College and Research Libraries* 65.5 (September 2004): 400-425.
- Gould, Constance C. *Information Needs in the Humanities: An Assessment*. Stanford, CA: Research Libraries Group, 1988.
- 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." Paper presented at ALLC/ACH 2004, Gothenburg, 2004.
- Warwick, Claire. "No Such Thing as Humanities Computing? An analytical history of digital resource creation and computing in the humanities." Paper presented at ALLC/ACH 2004, Gothenburg, 2004.

# Human Computing: Modelling with Meaning

---

**Willard McCarty**

*King's College London*

**Meurig Beynon** ([wmb@dcs.warwick.ac.uk](mailto:wmb@dcs.warwick.ac.uk))

*University of Warwick*

**Steve Russ** ([sbr@dcs.warwick.ac.uk](mailto:sbr@dcs.warwick.ac.uk))

---

In bringing the humanities and computing together, the question of how computer *science* relates to the humanities has to be addressed. Most striking is the starkly different treatment of meaning in the humanities and in computer science. To ignore this issue is to risk investing our limited notion of computer science with unwarranted authority. The commonplace view of computer science suggests a monolithic image of computing, in which all activity reduces to the execution of formal algorithms. Computing in the wild, in contrast, is both incorrigibly plural, and rich in possibilities for marrying a science of computing with a computing of the humanities. This session is designed to explore one such possibility.

The standard way of construing computer science focuses on combinatorics, syntax and algorithms. Its guiding question is "*what can be automated?*" (Denning). The benefits of asking this question are undeniable — more efficient pattern-matching, more advanced data mining, better data representation and the like. But these benefits, and the question that elicits them, do not address the humanities intellectually. They pertain to a relationship analogous to that between an accountant and his or her calculator — hardly a promising one for computing practitioners and humanities scholars alike. If we wish to have a computing of the humanities, we need to be asking a rather different question: "*how can we best integrate automated processing with human thinking and acting?*"

*Empirical Modelling* (EM-website), the approach around which this session has been organized, reflects a radical shift from the logical and linguistic philosophical stance of theoretical computer science to one based on the pragmatic empiricism of William James. It has been developed by two of the authors, Beynon and Russ, at Warwick University. The third author, McCarty, has independently developed a convergent idea of modelling based on the tradition of experimental science, recent historical and philosophical analyses of experiment and the phenomenology of Martin Heidegger, Michael Polanyi and

others. The convergence indicates, all will argue, a highly promising basis for interchange between computing science and humanities computing. This basis takes us considerably further than previous attempts. (See e.g. Gardin; Koch; *Computing the Future* 1992; *Computing and the Humanities: Summary of a Roundtable Meeting* 1998; Orlandi; *Beyond Productivity* 2003. See also *Transforming Disciplines: Computer Science and the Humanities*, <http://www.carnegie.rice.edu/>.)

The first paper discusses the prospects for partnership between the humanities and computing from the alternative perspective afforded by Empirical Modelling. It identifies perceived dualities that separate the two cultures of science and art as the primary impediment to this partnership, and outlines how these can be dissolved in a vision for 'human computing'.

The second paper illustrates the key characteristics and potential for EM for the humanities with reference to a projected modelling exercise addressing the Erlkoenig theme (as represented in the work of Goethe, Schubert and Liszt). It also highlights how each of six varieties of modelling identified by McCarty in (2004) can be represented within an EM model.

The final paper discusses the implications of EM with reference to McCarty's account of the key role for modelling in the humanities (2005), and considers these in relation to James's "philosophic attitude of radical empiricism" and ideas from phenomenological sources.

## Computing in the Humanities - Servant or Partner?

**Meurig Beynon and Steve Russ**

The term *humanities computing* evokes two images of relationship: one in which computing is the servant, the other in which it is a partner. To traditional humanists computing-as-servant is unproblematic — who does not wish to be served? But the more challenging notion of computing-as-partner promises the greater intellectual rewards. This paper proposes 'Empirical Modelling' as the basis for a new vision of *human* computing through which a strong and fruitful partnership can be built.

## Humanities and computing in partnership?

When we trouble to take a close look, rather than simply to relegate computing below stairs, its relationship to the humanities seems deeply troubling: on the one hand, flawless manipulation of data; on the other, contingent interpretation. We are reminded of the familiar *two cultures* caricature of the relationship between arts and science (cf. Collini in Snow). Unfortunately, the majority view of computer science (CS) sits comfortably alongside this popular caricature. At the theoretical

end, where the designation *science* best fits, CS describes formal, objective meaning as a computational recipe. But at the practical end, where application programming is done, CS faces the fourth decade of a messy *software crisis*. Uncertain human situations, including scholarly ones, have not meshed well with the *science*. Hence the quite separate concerns of theoreticians and practitioners within the field. New trends in computing subvert their separation, however. The manner in which data is represented and presented to the scholar is now open to negotiation, and it has become clear that different modes and technologies for presentation have significant cognitive implications. Neither the programmer nor the scholar is well-adapted to cope with this state of affairs.

Modern developments in practical computing present a serious challenge to computer science as it is currently understood. The sharply differentiated treatment of formal and informal meanings of programs is oriented towards applications in which mathematics plays a central role. This traditional view of computation made good sense in its historical context, when the archetypal role for the computer was "automating routine processes". As Brian Cantwell Smith has argued in (Smith 1987, Smith 2002), a foundation for computation in logic may suit programs with a preconceived abstract functionality but is not well-adapted for dealing with the relationships between form and content encountered in modern computing practice. Through its capacity to generate rich experiences, the computer can liberate the imagination, and in principle suggest fertile new modes of interaction that defy preconception.

In acknowledging and exploiting the semantic impact of holistic experience, computing practice has made a transition that our science of computing has not. Trying to give a mathematical account of computing is like trying to account for musical experience solely by music theory. This motivates us to reappraise computing from a totally different perspective in which *experience* rather than logic has a privileged role.

The objections to this reorientation centre on perceived fundamental distinctions between kinds of experience. In commonsense thinking about computing and the humanities, for instance, we distinguish experience of physical reality, experience of the virtual world, experience that can be communicated — formally or informally — through language, experience that can be authenticated by scholarship or experiment, and affective experience such as is associated with the appreciation of works of art. Attributing an absolute status to these distinctions endorses the familiar fractured caricature of the relationship between sciences and arts, at the ends of a spectrum of experience leading from the material world to the miraculous. Both computing and the humanities have made significant intellectual and practical contributions to challenging the status of these distinctions. Consider, for instance, the ontological issues addressed by Gooding in his discussion of

the status of virtual experiments in science, and the analysis of poetic treatments of the metaphysical and the material in Heaney. The alternative vision for computing endorsed by 'Empirical Modelling' is rooted in a philosophical position proposed by William James where the distinctions between different varieties of experience are taken to be no more or less than matters of classification (James). This is potentially significant both in respect of aligning the science of computing with its practice, and in negotiating — and perhaps in due course, consummating — the marriage of humanities and computing.

### **Human Computing and Empirical Modelling**

This section takes up the idea of reappraising computing from a perspective in which experience rather than logic plays a privileged role. This involves turning from the relationship between computing and the humanities as disciplines to consider the more concrete relationship between humans and computers.

Through their enormous flexibility and power, and the ethereal medium of electronics, computers have greatly extended the machine metaphor. The activity of programming allows us to make new 'machines' of extraordinary range and variety. A widespread view compatible with this metaphor sees the computer characteristically as an 'information processor'. Underlying such a 'machine computing' outlook the role of logic is central from the specification to the verification of both programs and hardware.

There is, however, a perspective on computers and their use that is independent of the machine metaphor and more fundamental. It has always been present in computing but has been so over-shadowed by the viewpoint, and usefulness, of machine computing that it has often been overlooked.

Before making any use of the computer I need to be able to relate what I see and do on the computer with my situation in my own world outside the computer. For this I must be able to present a part of my world, or some phenomenon, on the computer in a recognisable fashion. When this is a matter of using the computer in a machine mode (e.g. for e-mail or word-processing) this act of representation is very familiar. But it is now possible to make computer models with which we can deliberately dwell upon our personal understanding of something of interest for its own sake, and without any functional use yet in mind.

This role for the computer of building artefacts with which to think and explore has been facilitated by the improving technological management of the electronic medium. This has become, like paint, or music or language, a medium for self-expression. The fluidity and flexibility of the medium make it a potential match for close integration with the 'stuff' of human thought and perception.

The contrast then, with the machine mode of the computer, is the capacity of computer artefacts to offer us direct, 'felt' experience of parts of our own worlds. It is a 'likeness' established through the correspondence between the experiences, on the one hand, of interacting with our world, and on the other hand, of interacting with the artefact. It is this emphasis on the way computer artefacts may be experienced as if for the first time, then explored and developed before definite meanings have emerged, that is the essence of what we mean by 'human computing'. Computer artefacts themselves now become a significant source of experience, and — especially in terms of the quality of experiential interaction — they may even be offering us a new kind of experience.

Some of the early pioneers of electronic computing had a vision not unlike that of human computing. For example, many of the sentiments of the enthusiasts for electronic analogue computing (Small) resonate strongly with our ideas, and Licklider looked forward to a time when "men and computers would work in intimate association." But in the 1960's the technology made any such use of computers very difficult. Since then spreadsheets have been the most successful software to embody the idea of human computing. It has, however, been the explicit aim of the Empirical Modelling (EM) project at Warwick to develop principles and tools that give priority to experience rather than logic, and that promote the integration between human and computer processes that is at the heart of our vision for human computing.

The Empirical Modelling Project has been pioneered and led by Meurig Beynon at Warwick for over fifteen years. The work has been taken forward in large measure by many cohorts of third-year project students and many research students. The overall guiding principle has been the development of computer artefacts that offer similar experiences, through interaction, to those in some part of the modeller's own world. Fundamental practical concepts that have shaped the principles and the model-building tools are those of *observable*, *dependency* and *agency*. The characteristic activity of EM is the experimental identification of relevant observables associated with some phenomenon and of reliable patterns of dependency and agency among these observables. It is a modelling process that is more primitive than, and so prior to, the commitments inherent in programming. The approach is a broad one having relevance across the whole spectrum of computing. We shall introduce the ideas of definitive scripts and agent-oriented modelling by means of a small example and demonstration, and will give an overview of the on-line material available on EM (EM-website).

## Not in the notes: Erlkoenig as a case study in Human Computing

### Meurig Beynon

A companion paper (Paper 1) argues the need for a radically different perspective on computing that is particularly relevant to its role in the humanities. A key notion is dispelling the idea of an absolute duality in experience, and reinterpreting computing with respect to distinctions that rest on how experience is characterised. We can understand how this might work by recognising that semantic relations similar to those that arise in computer programming exist in the humanities. The pianist plays Chopin, but the score resembles a program. But where the computer scientist views the program as essentially defined by its precise abstract operational semantics, the musician — whether composer, pianist or analyst — takes a much more liberated view of the meaning of the musical score. The pianist is deemed to play a Chopin sonata, even though there are some wrong notes. Playing Chopin and playing the piano are both human skills that clearly admit no exact ultimate level of attainment, and the counterpoint between the two is a commonplace theme in music analysis and criticism. Particularly pertinent in this context is Mahler's remark that "what is best in music is not to be found in the notes" (Shapiro), and the well-attested fact that Chopin's use of rubato defied precise notation in a score (Schonberg).

A better understanding of the distinction between a musical score and a conventional computer program helps us motivate an alternative approach to computer-based modelling that can do fuller justice to the concept of humanities computing — that of Empirical Modelling (EM). The archetypal computer program is intended for machine interpretation, and is optimised for a specific function and context of use. Though the results of executing the program can be experienced by the human interpreter in the appropriate user role, any human interpretation of the program in execution is in general a most specialised exercise in interpreting machine operation that is of its essence unintelligible within the context of use. What is more, the degree of specialisation and optimisation of the program to function is typically such that the user-oriented interpretation disintegrates on changing the merest detail — all that remains to the programmer is to 'debug' the behaviour of the machine. Contrast the musical score. Though the aspiration of the pianist may be to trace the execution from beginning to end with the strictest adherence to the score, the process of interpretation resembles reading a computer program no more than it resembles reading a piano roll for a player-piano. (Indeed less, since in this analogy a computer program is typically more like a prescription for punching holes in a piano roll.) The pianist may enter the score at any point in time, extract melodic fragments, or adapt the written prescription in order to savour the experience of a particular chord, to shape the inflection of

a melody, or identify the essence of a technical difficulty. In this activity, in accordance with Mahler's dictum, the pianist will give ultimate priority not to being in every respect accurate to the score, but to evoking and communicating 'the felt experience'. In the spirit of Turner, the separation between the technical accomplishment and the musical effect is not a sharp duality: the two experiences of playing the piano and playing Chopin are blended in the experience of the human interpreter. The priority that is given to those aspects of the interpretation of the score that are least precisely documented is reflected in the way that we say: "She played Chopin's Revolutionary Study" rather than "She used the piano to execute Chopin's Revolutionary Study." This distinction between stances towards interpretation speaks to a yet deeper tension between the values of the humanities and the method-tool-use paradigm of the business IT culture (EM-website 055).

The principles of EM, and the respects in which they represent a radical departure from conventional thinking about computing with implications for the humanities, can be illustrated with reference to a study in modelling music. For this purpose, our choice of theme is Erlkoenig, as first dramatised in verse by Goethe, then set to music by Schubert, and later transcribed by Liszt for piano solo. The objective for this case study is to show how the application of EM principles and tools is suited in principle to the development of an auxiliary model that can serve a whole variety of different functions for the human interpreter. At present, the construction of such a model is in its earliest stages, but its broad conception can be outlined by drawing upon well-established experience of EM for a wide range of applications (EM-website, EM-archive). An important and characteristic theme of EM that echoes sentiments expressed about modelling by McCarty (2004) is that the potential scope and function of the model cannot be preconceived, nor will the model ever represent more than "a temporary state in the process of coming to know." In this respect, it is crucially different from a conventional program in having no preconceived formal specification, and being intended and open for indefinite extension and elaboration.

The case study has been chosen to highlight a number of key issues: the fundamental significance of the shift in perspective towards the radical empiricist outlook of William James (1996) rooted in the idea that 'one experience knows another'; pertinent aspects of EM from a technical modelling perspective, such as the role for observation, dependency and agency, the scope for invoking concurrent agents in the interpretation, and the merits of EM in respect of combining models; how each of the six varieties of modelling identified by McCarty (2004) can be represented within a single EM model.

The importance of a radical empiricist stance stems from the need to account for a treatment of meanings in the humanities that is far beyond the scope — though not perhaps the

aspirations — of the formal semanticists in computer science and AI (Smith 1987). Consider the audacity of the following extract from Maurice Brown's commentary on the Erlkoenig:

Even more remarkable, as was first pointed out by Sir Donald Tovey in a superb programme note on the song, is the treatment of the pianoforte when the child speaks. During the rest of the song we are observers: we watch the ride, we hear the child's voice and the father's reassuring answers. But only the child hears the Erlking, and the rocking, almost lulling, movement of the pianoforte accompaniment is the child's experience of the motion of the galloping horse, the warm protection of his father's arms, while he trembles at the sinister invitation. When he cries out, we revert to observers and the clamour of the hoofs, the rush of the wind, break again on our ears.

(Brown)

In Jamesian terms, both Brown and Tovey are testifying to the experience of a conjunction of two experiences (the texture of the musical accompaniment and the child's perspective on events) for which no formal explanation need be given. It is quite characteristic of such a conjunction that its recognition is to some extent enabled by a purely technical consideration — that these changes in texture come as an enormous relief to the accompanist, so taxing is the pianistic device that evokes the horse's unrelenting ride.

From a technical modelling perspective, Erlkoenig is a rich source of instances of agency, dependency and observation. EM makes use of techniques for distributed modelling (cf. EM-archive: claytontunnelSun1999) and animation (EM-archive: railwayYung1995) that can underpin concurrent engineering (EM-website: 034). The model-building can be framed with reference not only to the various perspectives of external agents (in this context, the poet, the composer, the singer, the accompanist, the translator etc) but also those internal to the drama itself (the father, the child, the Erlking, the horse). A vital aspect of EM is that model construction is not compromised by optimisation to performing some specific function, as in conventional programming, so that blending of models is pervasive, and there is openness to extension possibly even in the light of subsequent developments in tools and technology (cf. the new pianistic possibilities explored by Liszt in his transcription of Erlkoenig).

The status of EM as a radical generalisation of modelling with spreadsheets makes it possible to envisage a role for modelling extending that illustrated by McCarty in his *Analytical Onomasticon* to Ovid's *Metamorphoses* (2005, Chapter 1). Musical counterparts for the analogy, representation, map, diagram, simulation and experiment can be found in modelling Erlkoenig and identified in EM. Of particular interest is the combination in the context of a music of formal and informal semantic frameworks. One might for instance seek an authentic virtual reconstruction of an early performance of Erlkoenig as Schubert himself might have heard it (cf. Beacham), or wish



to elaborate on the semi-formal analysis of musical language of Erlkoenig that Cooke initiates in (Cooke). A precedent in EM for combining formal and informal semantic ingredients within a single model can be found in (EM-website: 051).

## Towards a philosophy of modelling for humanities computing

Meurig Beynon and Willard McCarty

In developing a persuasive philosophical stance on humanities computing, the first task is to relate its aspirations to the current vision of computer science. In (Paper 1), Beynon and Russ propose that an alternative science of computing is needed to bring computing and the humanities into a more fulfilling relationship. McCarty (2004) identifies a better understanding of "what modelling is" as key to making sense of humanities computing. This paper — to be read in conjunction with (McCarty 2004) — revisits McCarty's arguments in the context of the critique of traditional thinking about computing motivated by the study of Empirical Modelling (EM) (Paper 1).

Informally, McCarty's *Onomasticon* (McCarty 2005) may serve as an archetypal example of EM. Though it has been built using commercial spreadsheet and database software, rather than the special-purpose tools that have been developed for EM (EM-website), its development exploits the essential principles and concepts of EM. It is characteristic of this development that (to paraphrase McCarty 2005) the *Onomasticon*, however finely perfected, is better understood with reference to temporary states in the process of coming to know rather than a fixed structure of knowledge. In thinking of the *Onomasticon* in EM terms, the term *model on the computer* is preferred to McCarty's *computational model*. The principal reason for this is that the way in which EM views the semantics of the *Onomasticon* is quite different from what is understood by the computational semantics of the underlying computer program (cf. Cantwell Smith's discussion of semantic relations in Smith 1987). Specifically, the manner in which the EM model (the *Onomasticon*) represents the referent (Ovid's *Metamorphoses*) is that there is a repertoire of 'atomic interactions' that the modeller can make both with the model and with its referent and that these are perceived by the modeller (McCarty) to connect the experience of the model with that of its referent.

As McCarty's careful analysis of terminology (McCarty 2004) indicates, the dynamic and provisional quality of the model argues against describing the model as 'a representation' of its referent. For reasons discussed at length in (EM-website: 078), the terminology that William James introduced in considering relations between experience is preferred: "experience of the model *knows* experience of the referent". It is to be understood that the modeller will never be obliged to 'explain' why one experience knows another experience, nor to make any claims for the objectivity of this perceived relationship. This is the

essence of James's *Radical Empiricism* (James), that relations between experiences are themselves given in experience.

Though it is accepted usage to refer to the spreadsheet as a model of a financial situation, this is not the sense in which *model* is most commonly used in computer science. Expressions such as *model-checking*, *model-based reasoning*, mathematical *model* allude to far more abstract semantic relations that are by no means directly apprehendable in experience. When we conceive a model as a set of logical equations or constraints, the manner in which the model is experienced is outside the semantic scope. Invoking the alternative semantic framework of EM entails being more discriminating about kinds of computing activity, and motivates a reappraisal of what McCarty (2004) identifies as the "decisive criteria" for modelling by computer: *complete explicitness and absolute consistency and manipulability*.

Where consistency is concerned, it must be recognised that the experience a computer generates is not explicitly specified in every respect — at any rate not in the same sense that an abstract computation is explicitly specified. EM focuses on the experiential aspects of computer-based models, for which — as is appropriate for humanities computing in general — no presumption of complete explicitness and absolute consistency in informal semantics is required. Indeed, in (EM-website: 072), Beynon makes the case that the semantic framework of EM is aptly suited to dealing with situation, ignorance and nonsense ("the principle of SIN"). For this purpose, it is not the linguistic and logical frameworks supplied by Chomsky and Tarsky or the syntactic treatment of metaphor in logicist AI that are appropriate (EM-website: 050), but semantics closer in spirit to the thinking of Lakoff and Turner.

Where manipulability is concerned, it may seem that we can manipulate representations effectively using a computer because we can modify programs. The notorious difficulty of adapting conventional programs to meet new requirements is evidence that this contention cannot be taken at face value. And where "one experience knows another" is concerned, there are serious conceptual and practical objections to deeming the common debugging cycle (as in "stop execution of program P, fix line 235, recompile, run program P 'to the same point as it was before' — whoops ... I've introduced another bug — etc etc ...") to be an atomic transition in experience. In practice, manipulability is bound up with contextual and pragmatic issues that are entirely alien to the formal semantics of computation. This is consistent with McCarty's observation that "manipulation ... requires something that can be handled [in] a time-frame sufficiently brief that the emphasis falls on the process rather than its product" (McCarty 2004). For this purpose, the notion that "the experience of adjusting the computer model should know the experience of adjusting the interpretation of the referent" is precisely what is required.

The decisive emphasis of EM is on what is known in immediate experience, and what in William James's terms is associated with "the most intimate conjunctive relation .... that experienced between terms that form states of mind" (James 44-45). Within this apparently limited frame of "what experience knows another in-the-now" all kinds of conception of model are possible through assuming different kinds of context, observation and agency. This is the very subject of James's *Radical Empiricism*. James develops the story of knowledge to deal with expectations of what has not been experienced — knowledge that transcends direct experience. The rich quality of engagement with past and future experience that this demands is well-represented in EM, both in the characteristic inflection of the "what if?" interaction, and the capacity to replay the entire process of construction as one might in exposing the sequence by which the cells of a spreadsheet came to be defined. This facility, frequently exploited in presenting EM models, captures the aspiration for modelling identified by Dening — "[that we may] return to the past the past's own present, a present with all the possibilities still in it, with all the consequences of actions still unknown" .

In appreciating the shift of perspective in EM fully, it is vital to distinguish the semantics given in experience in a state-of-mind from semantics based on behaviours (as in program semantics (Smith 1987)) — even when these are guided by experience (as in Turner's treatment of narrative (Turner), and CantwellSmith's discussion of "the process semantics" (Smith 1987)). This is evidenced by the diversity of contexts behind the wide range of applications for EM (EM-website), and the associated diversity of models. As is illustrated in (Paper 2), EM can be used to generate just such rich varieties of model — analogy, experiment, simulation, map, diagram, representation — as are catalogued in McCarty (2004). This diversity is enabled precisely because an EM model is identified by a state and a body of latent anticipated interactions that can be more or less familiar and significant to the modeller, or any other human interpreter, and in this way serves as an interactive environment whose meaning is constrained only by the imagination. This delivers more than is envisaged by Minsky or Naur in respect of 'constructed models': beyond the confirmation of a theory, a place for "blind variation" in the sense of Vincenti — interaction "without complete or adequate guidance" potentially leading to discovery.

Several intriguing philosophical connections identified by McCarty (2005) are ripe for further scholarship and exploration. The suggestive links between EM thinking and the phenomenology of Polanyi and Heidegger echo the phenomenological interpretations of software development offered by Winograd and Flores, but also argue against invoking such interpretations in relation to traditional software practice. Of crucial importance in ensuring the universality of the concept

of modelling, and embracing activities that involve creation and discovery, is the ontological status of the model, the referent and the relation between them. The idea of an EM model as a construal invokes Vaihinger's "as if" : neither true nor false, but as construed for the purpose in hand. Such a stance even underwrites propositions such as "our constructions continue to work, no matter how violent the changes in scientific opinion may be" (cf. McCarty 2005) that might be seen as authorising absolute claims for EM models as physical artefacts. This outlook accords with James's contention that "subjectivity and objectivity are affairs not of what an experience is aboriginally made of, but of its classification" (James 141), and his perspective on the difficulties of understanding the direct products of experience: "But how the experiences ever get themselves made, or why their characters and relations are just such as appear, we can not begin to understand " (James 132-133). The pragmatic importance of this ontological stance for humanities computing is that it helps to dispel the mystique that surrounds high art and hard science: a mystique that is the pretext for divisive absolute partitions in experience.

## Bibliography

Beacham, Richard, and H. Denard. "Roman Theatre, Frescos, and Digital Visualisation: Intermedial Research." *Proceedings of the 4th Int Symposium on Virtual Reality, Archaeology and Cultural Heritage*. 2003. n. pag..

Brown, Maurice J.E. "Schubert Songs." *BBC Music Guides*. n.p.: BBC, 1967. 16.

*Computing and the Humanities: Summary of a Roundtable Meeting*. New York: American Council of Learned Societies, 1998. National Research Council; Coalition for Networked Information; National Initiative for a Networked Cultural Heritage; Two Ravens Institute; Computer Science and Telecommunications Board.

Cooke, Deryck. *The Language of Music*. Oxford: Oxford University Press, 1959. 98.

Dening, Greg. *Readings/Writings*. Chicago: University of Chicago Press, 1998.

Denning, Peter J. "What Is Computer Science?" *American Scientist* 73.1 (1985): 16-19.

*EM-archive*. Accessed 2005-04-08. <<http://empublic.dcs.warwick.ac.uk/projects/>>

*EM-website*. Accessed 2005-04-08. <<http://www.dcs.warwick.ac.uk/modelling/>>

Gardin, Jean Claude. "On the Way We Think and Write in the Humanities: A Computational Perspective." *Research in*

- Humanities Computing 1. Papers From the 1989 ACH-ALLC Conference*. Ed. Ian Lancashire. Oxford: Clarendon Press, 1991. 337-345.
- Gooding, David. *Experiment and the Making of Meaning: Human Agency in Scientific Observation*. n.p.: Kluwer, 1980.
- Hartmanis, Juris, and Herbert Lin, eds. *Computing the Future: A Broader Agenda for Computer Science and Engineering*. Washington DC: National Academies Press, 1992.
- Heaney, Seamus. "Joy or Night: Last Things in the Poetry of W.B. Yeats and Philip Larkin." *The Redress of Poetry (Oxford Lectures)*. London: Faber and Faber, 1995. 146-163.
- Heidegger, Martin. *Being and Time*. San Francisco: Harper, 1962. Revised edition.
- James, William. *Essays in Radical Empiricism*. n.p.: Bison Books, 1996. (First published 1912.)
- Koch, Christian. "On the Benefits of Interrelating Computer Science and the Humanities: The Case of Metaphor." *Computers and the Humanities* 25.5 (1991): 289-295.
- Lakoff, George, and Mark Johnson. *Metaphors We Live By*. Chicago: University of Chicago Press, 1980.
- Licklider, J.C.R. "Man-computer Symbiosis." *IRE Transaction in Human Factors in Electronics* March (1960): 4-11.
- McCarty, Willard. *Humanities Computing*. Houndmills, Basingstoke: Palgrave, 2005 (forthcoming).
- McCarty, Willard. "Modelling: A Study in Words and Meanings." *A Blackwell Companion to Digital Humanities*. Oxford: Blackwell's, 2004.
- Minsky, Marvin L. "Matter, Mind and Models." *Semantic Information Processing*. Ed. Marvin L. Minsky. Cambridge, MA: MIT Press, 1968. n. pag..
- Mitchell, William J., Alan S. Inouye, and Marjory S. Blumenthal, eds. *Beyond Productivity: Information Technology, Innovation, and Creativity*. Washington DC: National Academies Press, 2003.
- Naur, Peter. *Knowing and the Mystique of Logic and Rules*. Boston: Kluwer, 1995.
- Orlandi, Tito. "Is Humanities Computing a Discipline?" *Jahrbuch für Computerphilologie* 4 (2002): 51-5.
- Polanyi, Michael. *The Tacit Dimension*. New York: Doubleday and Co, 1966.
- Schonberg, Harold C. *The Great Pianists*. London: Victor Gollancz, 1965. 143-145.
- Shapiro, Nat, ed. *An Encyclopedia of Quotations About Music*. New York: Da Capo Press, 1977.
- Small, James. *The Analogue Alternative*. London: Routledge, 2001.
- Smith, Brian Cantwell. "Two Lessons of Logic." *Comput. Intell.* 3 (1987): 214-218.
- Smith, Brian Cantwell. "The Foundations of Computing." *Computationalism: New Directions*. Ed. Matthias Scheutz. Cambridge, Mass.: MIT Press, 2002. 23-58.
- Snow, C.P. *The Two Cultures*. Edited with introduction by Stefan Collini. Cambridge: Cambridge University Press, 1998.
- Tovey, Donald Francis. *Essays in Musical Analysis Vol. 5*. London: Oxford University Press, 1948. 194.
- Turner, Mark, ed. *The Literary Mind: The Origins of Thought and Language*. Oxford: Oxford University Press, 1996.
- Vaihinger, H. *Philosophy of As If: A System of the Theoretical, Practical and Religious Fictions of Mankind*. 2nd ed. n.p.: Routledge, 1984.
- Vincenti, Walter G. *What Engineers Know and How They Know It: Analytical Studies from Aeronautical History*. Baltimore: Johns Hopkins University Press, 1993.
- Winograd, Terry, and Fernando Flores. *Understanding Computers and Cognition: A New Foundation for Design*. Boston: Addison-Wesley, 1986.

# Heraldic Applications of Computational Linguistics, Computational Geometry and Image Processing

*Michael McKeag (R.McKeag@qub.ac.uk)*  
*The Queen's University of Belfast*

## Introduction

The official definition of a coat of arms is not the depiction of the arms but the blazon, which is the text describing the arms. It requires a skilled heraldic artist to render the arms from the blazon. Anyone well versed in heraldry can perform the inverse operation, to render the blazon from the arms.

The paper describes projects to automate these processes, by capturing the herald painter's skills and drawing upon computational linguistics (constructing the syntax and parsing the blazon) and computational geometry (positioning and sizing the components) and by using image processing techniques (identifying the components, layer by layer, that make up the shield). More importantly, the paper emphasizes that these techniques have wider application in the arts and humanities.

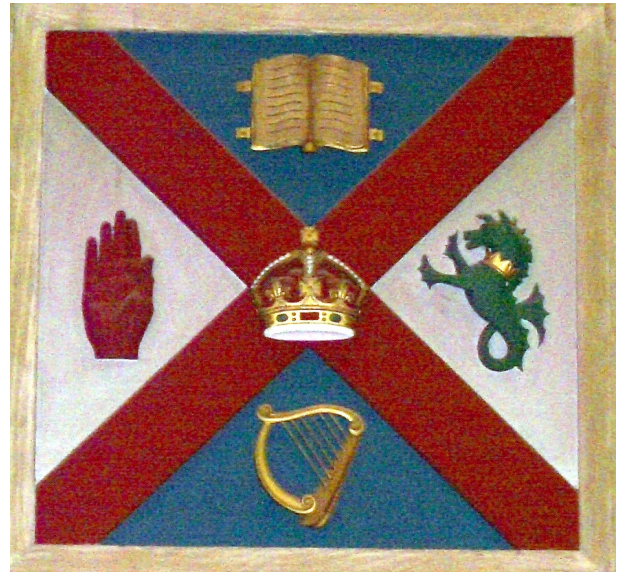
## Context

An important rationale for many computer applications is to capture skills so that tasks can be tackled by unskilled operators. In the visual arts, computers have long been used to help people draw or paint two dimensional pictures or to model and render three dimensional scenes.

One area that provides a considerable challenge is heraldry. It requires artistic skills to render the arms from the blazon, thereby limiting the rôle of those unskilled in the use of pen and paint. It also requires heraldic knowledge to depict or recognize arms.

The blazon defining the Queen's University's arms (Figure 1) is: *Per saltire azure and argent, on a saltire gules, between in chief an open book and in base a harp both proper, in dexter a hand coupé of the third, and in sinister a sea-horse vert gorged with a mural crown of the fourth, an Imperial crown of the last.* From this definition any competent heraldic artist can draw the shield whereas the quality of computer programs

that attempt to automate this process is at present limited and inferior. Much the same was said of the printer's art when desktop publishing was first attempted.



*Figure 1: The Queen's University of Belfast*

Several exploratory projects to automate this task have been undertaken at Queen's by final year undergraduates and by taught course postgraduates.<sup>1</sup> The results have been promising but, because of the time available to those students, limited. Now is the time to build upon those studies to produce a much more acceptable application that not only consolidates the work done to date but also addresses the more challenging questions that have yet to be tackled. This paper discusses the progress so far and outlines what problems remain and how they might be solved.

The artist's skill shows in the disposition and scale of the components. No program can rival the inventiveness of an artist but there remains the question of whether an acceptable rendition can be generated. Such an application draws upon computational linguistics and computational geometry; it also entails constructing a dictionary of heraldic terms and a corresponding library of heraldic images. It may be that a semi-automatic solution is necessary, allowing the user to make fine adjustments. In fact, some human intervention will be required because the set of components that may appear in a coat of arms is unbounded and new ones will inevitably be encountered and will have to be entered into the dictionary and library.

The inverse operation, of deconstructing a coat of arms to produce the blazon, has also been attempted in a student project (Flanagan) and is now to be tackled more seriously, using experience gained from another student project (Taylor) that uses image processing techniques to extract the contours and

other information from a map. Again, the problems and how they might be resolved are discussed in the paper.

## Benefits

The language of blazon has been in continual use since the Middle Ages. Based on the French of the time, and subsequently adapted, it has a reasonably precisely defined grammar and a substantial vocabulary that is extended as necessary. To most people that are neither heralds nor heraldists it appears to be arcane. This project should help to make heraldry more widely accessible; it should make heraldic and artistic skills available to the unskilled and it should also be a useful tool for heralds, at least for intermediate sketches if not for the final grant of arms.

It should also make the production of illustrated rolls of arms feasible as, at present, there are not enough heraldic artists to produce an illustrated version of, say, Burke's *General Armory*, which describes thousands of shields listed by the names of their owners. The inverse operation of identifying the owners of shields is generally accomplished using Papworth's *Ordinary*, which is not illustrated and can be difficult to use; automation can play a useful part here. Heraldic offices are now beginning to use computers in a more imaginative way and they will want to develop indexed and illustrated databases of arms.

More generally, it may encourage those working in the visual arts to develop more precise written descriptions of their artifacts. Heraldry is somewhat stylised and lends itself to precise formal descriptions; that is not the case for most art but it may be that the correspondence between blazon and image might encourage the development of more precise descriptions of other works of art, as happens with music or dance notation.

Computational techniques for the construction of grammars, the disambiguation of syntax and the parsing, processing and indexing of texts have wide application; we have experience with subsets of Latin, English, French, Spanish and Esperanto as well as blazon.

Geometrical and topological techniques for describing and manipulating two and three dimensional shapes are also widely applicable; we have experience of applying such techniques to mechanical and aeronautical engineering artifacts. It is now time to apply these methods to other areas of scholarship.

## Examples & Problems

By way of example, the following coats of arms (Figures 2a-2g) are taken from Foster's *Feudal Coats of Arms* and have been generated automatically by computer program from blazons, some of which have been rephrased.



Figure 2a: Or, on a cross gules five escallops argent.

**Bigod, Rauff**, of Settrington, Yorks  
Glover Roll &c.  
Ascribed also to **John** in the St. George Roll, and to **Sir Raffe**, of Norfolk, in the Parly. Roll.



Figure 2b: Azure, six eaglelets, 3, 2, 1, or.

**Biblesworth, Sir John**  
Ashmole Roll  
Also ascribed to **Walter**.



Figure 2c: Per pale or and vert, a lyon rampant gules.

**Bigod, Roger**, Earl of Norfolk,  
Earl Marshal of England  
at the battle of Falkirk 1298 and sealed the Barons' letter to the Pope 1301, pp xv, xxiv.



Figure 2d: Quarterly or and gules, a bend sable.

**Beauchamp, John, Walter, and William**, of Bedford

(H. III Roll) Parliamentary, Norfolk and St. George Rolls [*bend gules*; in other MSS. *bend sable* occurs].



Figure 2e: Azure, three cinquefoyles or.

**Bardolf, Sir Hugh**, a baron 1299  
Sealed the Barons' letter to the Pope 1301, pp xvii, xxiv;  
bore at the battle of Falkirk 1298 and at the siege of  
Carlawerock 1300.



Figure 2f: Or, a chevron azure.

**Bastard** (---), of Kitley, Devon  
Shirley.

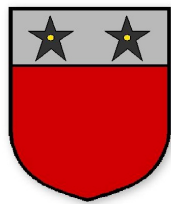


Figure 2g: Gules, on a chief argent two mullets sable pierced or.

**Bacon, Sir Edmond**  
at the first Dunstable tournament 1308.

Compound coats, like the arms of Queen's College Belfast (Figure 3) or those of the Royal University of Ireland (Figure 4), have still to be incorporated but this should not be too hard. Extending the dictionary of terms and the corresponding library of images will also be reasonably straightforward, utilising Elvin's *Dictionary of Heraldry* and images available online (Phillips). A much more difficult task will be to extend the formal grammar to cope with a larger range of blazons. In any language processing application the construction of a formal grammar is usually the most time consuming part as textbooks tend to define syntax informally or by means of examples.

Another difficult problem is the sensible handling of potential ambiguities.



Figure 3: Queen's College Belfast



Figure 4: The Royal University of Ireland

When we turn to the inverse operation, the decomposition of a picture into its components in order to derive the blazon, we meet a different set of problems. As with all image processing

applications, we have to remove noise before we can begin to identify components. Having removed a component we must then fill in the gap in the component that underlay it. We must finally construct the blazon, taking care to introduce the components in the right order and to conform to the standard conventions of blazoning.

A partly successful example is shown below (Figure 5). Instead of recognizing the background as *per fess or and azure*, it treats each half as a *fess*, i.e. a horizontal band across the middle of the shield. However it recognizes the diagonal *bend sable*. It is, though, well on the way to establishing that the blazon is *Per fess or and azure a bend sable*.

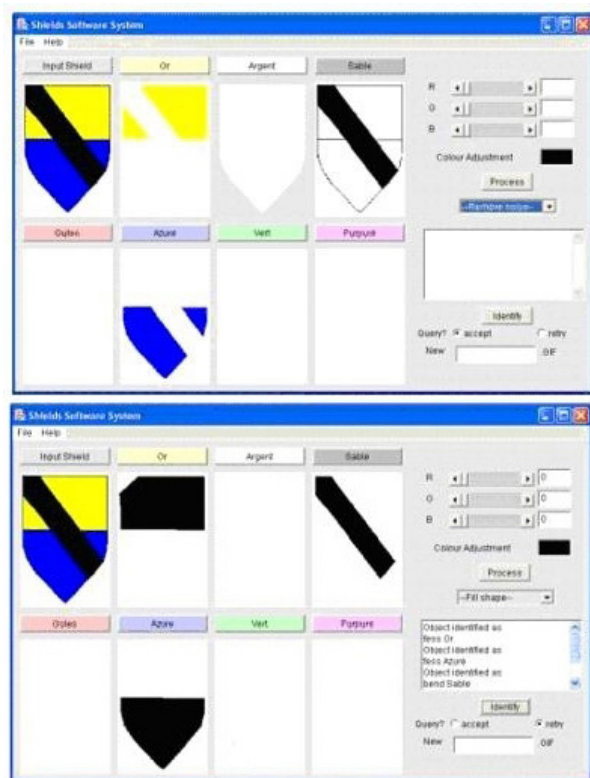


Figure 5: Stages in decomposing a coat of arms: colour separation, noise reduction, shape filling and component identification

1. See Carswell; Stewart; McVicker; Tinman; Black; Carson; Nicholl; McGuckin; O'Shea.

## Bibliography

Black, Sarah-Jane. *Heraldic Register and Ordinary*. 2001. (Student project.)

Burke, J.B. *General Armory*. Ramsbury, Wiltshire: Heraldry Today, 1961, 4th impression 1989. Facsimile of The General

Armory of England, Scotland, Ireland, and Wales; comprising a Registry of Armorial Bearings from the Earliest to the Present Time by J. B. Burke (1842, last edition 1884).

Carson, Gayle. *Coat of arms*. 2002. (Student project.)

Carswell, T. Peter. *Displaying coats of arms*. 1988. (Student project.)

Elvin, C.N. *Dictionary of Heraldry*. London: Heraldry Today, 1969, 2nd printing 1977. Facsimile of the original edition by C. N. Elvin (1889).

Flanagan, Lorna. *Shields software system*. 1998. (Student project.)

Foster, J. *The Dictionary of Heraldry – Feudal Coats of Arms and Pedigrees*. With an introduction by J.P.B. Brooke-Little. London: Bracken Books, 1989. Previously published as *Some Feudal Coats of Arms* by Joseph Foster, published by James Parker & Co. (1902).

McGuckin, Niall. *A coats of arms editing package*. 2003. (Student project.)

McVicker, Stephen. *The herald's editor*. 1997. (Student project.)

Nicholl, Paul R. *Coats of arms*. 2002. (Student project.)

O'Shea, Laura. *Automated heraldic artistry*. 2004. (Student project.)

Papworth, J.W. *Ordinary of British Armorial*. With an introduction by J.P.B. Brooke-Little. London: Heraldry Today, 1985. Facsimile of An Alphabetical Dictionary of Coats of Arms belonging to Families in Great Britain and Ireland; forming an extensive Ordinary of British Armorial by J. W Papworth & A. W. Morant, published by T. Richards (1874).

Phillips, D. *Shareware clip art collection of heraldic charges*. Accessed 2005-02-28. <<http://www.digiserve.com/heraldry/gifart.exe>>

Stewart, William. *Displaying a coat of arms*. 1993. (Student project.)

Taylor, Simon. *3D maps*. 2005. (Student project.)

Tinman, Richard. *Blazoning arms*. 2000. (Student project.)

## Classifying the Chimera

---

*Federico Meschini* ([fmeschini@tin.it](mailto:fmeschini@tin.it))

*Tuscia University*

---

While the *Digital Library (DL)* concept is an extremely vague one, its paradigm, the real implementation, is, if possible, still more elusive.

Since the late 80s and during the 90s, the primary concern and task in the textual digital resources field was really a basic (but not simple) one: how can we put texts in computers? How can we encode, manage and memorize them? Which stuff digital texts are made of?

Other important problems were perceived (visualization, for example) but they were temporarily put aside for practical reasons. The Text/Data relationship is really a Castor and Polydeuces' one: is Text a particular kind of Data or Data a particular kind of Text? The main choice, and long-term winning, in the textual encoding issue was the use of the powerful *Standard Generalized Markup Language (SGML)*, and – more specifically – the rules (or better *guidelines*) established by the Text Encoding Initiative (TEI), which could actually be considered the de facto standard for encoding humanistic texts in digital format.

The transition from SGML to the *eXtensible Markup Language (XML)* was, more than an evolution, a sort of Copernican revolution for two main aspects: the introduction of the new stylesheet technology for displaying an XML file, in particular the powerful *Extensible Stylesheet Language Transformation (XSLT)*, and the great diffusion and development of open-source software able to manage XML documents. This transition has of course also taken place in the TEI: starting with the P4 version of the Guidelines, XML is the technology now used. As a logic result different open-source software tools for implementing a Digital Library with texts encoded in TEI/XML are now available.

These tools are divided into two main categories. In the first group there are software created in hard Information Technology contexts (Database Management Systems, Native XML Databases, Publishing Framework, etc.), and these programs need to be adapted to the specific aims of a digital library<sup>1</sup>. In the second, and this is a new trend, these tools are being developed in academic contexts<sup>2</sup>, created from the scratch, or using programs from the first group as the basic core; in both cases the final function is the utilization with digital cultural resources and overall TEI encoded texts.

Compared to the first group, the logical added value of these tools is that often they provide some specific features for the textual field<sup>3</sup>.

The choice is now much wider than it was just a few years ago, but it could also cause some confusion in selecting the right tool. As the experts of knowledge management well know, too much information, if not well structured and organized, is equivalent to no information at all, thus being completely useless. Each software has its own peculiarities, which should be evaluated and confronted against the characteristics of the texts being encoded and the general needs and aims of the project of which the digital library is part. What is the main aim of a project? It's the visualization one, with perhaps a multiple output feature? It's the research, with some form of advanced textual analysis? It is possible to combine all these aspects? And if somebody has already found a solution for our problems how can we find it over the net?

Trying to find a solution for these problems, or better, trying to share the same problems to have common solutions, during the TEI Members Meeting 2003, there was the first reunion of the TEI Presentation Tools Special Interest Group<sup>4</sup>. The Presentation Tools SIG has two main initiatives: the creation and update of a tool list and of a sample collection of texts for testing.

The first version of the tool list has been presented during the TEI Meeting in Baltimore, last October. This list is actually a digital document encoded with the TEI/XML standard, and in this way it's currently published, using an XSLT stylesheet, in HTML<sup>5</sup>. With a simple structure this list presents the various software in an alphabetical order with a short description and the links to the various implementations. From the descriptions and the links it's possible to have an idea of the distinctive features of each tool, that for example the software *XPhilologic* is very good for full-text search and document retrieval<sup>6</sup> and that *Apache Cocoon* could be implemented so to have an XML framework for format scalable output of the same TEI document<sup>7</sup> And again, with *Anastasia* is possible to have an electronic text/image edition of a medieval manuscript<sup>8</sup> and *eXist* is a powerful native XML database that could be used for queries and researches.<sup>9</sup>

But this is not enough. Perhaps from a list you can obtain some information, but what is really needed (and planned since the beginning) it's an higher level of classification, and this become more necessary as the number of such software is increasing<sup>10</sup>. It's a software written in Java or in Perl? Which are requisites for running it on a computer? It's XML-aware? It allows XSLT transformation? The texts are stored in the file system or in a database? What are its peculiar features? It could be integrated with other software in order to augment the possibilities? It can be customized? It's clear that a simple list cannot answer to all these questions.



Many discussions have been made about the kind of classification to apply to the tool list and in my opinion it should be made using practical rather than theoretical principles, with a sort of empirical and pragmatic observation, including also the links to the most possible numbers of the concrete implementations of these tools, so to highlight the best practices and the particular features of each digital library.

A good way of realizing this classification could be the use of the standard ISO 13250<sup>11</sup> or TopicMaps, and the respective *XML Topic Map (XTM) syntax*<sup>12</sup>. A TopicMap is based on the definition of a general topic, the particular and real occurrences of that topic, and the associations between different topics, thus in my opinion it's the best way to obtain a complete classification, which will include the various aspects, from the most technical, concerning the programming languages used or the technical specification needed, to the functionalities of visualization, text research and analysis.<sup>13</sup>

So what is now a TEI document should be elaborated in a XTM document, detecting, separating, organizing, linking and classifying all the information that now are presented in a linear structure.

Once created, the XTM file representing the TopicMap can be used and navigated in several ways. Being an XML file, it is possible to apply the same technologies used for the TEI texts, but there are also available some dedicated software which can exploit the great potentialities of this standard as, for example, the *Omnigator* from *Ontopia*<sup>14</sup>, or the *TM4J*<sup>15</sup>, a java open-source package expressly developed for creating, manipulating and publishing topic maps.

The TopicMap technology has been presented for the first time related to the TEI during the 2003 meeting, and it's growing in interest from this community, for its possibility of adding a metadata semantic layer to the digital collections<sup>16</sup>. Moreover, thanks to the possibilities of merging different XTM documents each representing a different map, the Presentation Tools TopicMap could be integrated with other map about other subjects, the textual content for example<sup>17</sup> or the documentation of the local views of the DTDs<sup>18</sup>, thus creating the basis for the definition of what could become a "TEI Ontology".

- 
1. E.g., *Apache Cocoon* <<http://cocoon.apache.org/>>, *Apache AxKit* <<http://axkit.org/>>, *eXist* <<http://exist.sourceforge.net/>>
  2. *Anastasia* <<http://anastasia.sourceforge.net/>>, *teiPublisher* <<http://teipublisher.sourceforge.net/docs/index.php>>, *XPhilologic* <<http://barkov.uchicago.edu/xphilo/>>
  3. See for example the *TAPoRware* set for textual analysis <<http://cheiron.mcmaster.ca/~taporware/>> or the *Versioning Machine* <[<\[s/ver-mach/\]\(http://www.tei-c.org/Members/2003-Nancy/mm17.html#tap-sig\)> for the comparison of different versions and editions of the same text.](http://mith2.umd.edu/product</a></li>
</ol>
</div>
<div data-bbox=)

4. <<http://www.tei-c.org/Members/2003-Nancy/mm17.html#tap-sig>>
5. Available on line at <<http://miro.acs.its.nyu.edu/tei/cms/show.php>> .
6. See for example the demo on the *Brown Writer Women Collection* at <<http://barkov.uchicago.edu/xphilo/search.brownwvp.html>> .
7. A good implementation of *Cocoon* with TEI can be found at <<http://www.nzetc.org/>> .
8. See the *Caxtons' Canterbury Tales* at <<http://www.cta.dmu.ac.uk/Caxtons/>> .
9. See the *Digital Quaker Collection* at <<http://esr.earlham.edu/dqc/>> .
10. See the presentations on this subject at ALLC/ACH 2004 <<http://www.hum.gu.se/allcach2004/AP/>> . Among the others: Kumar, Amit et al., *teiPublisher a repository management system for TEI documents* <<http://www.hum.gu.se/allcach2004/AP/html/prop118.html>> ; Matthew Zimmerman, *Using AMP technology (Apache, MySQL, PHP) for XML publication* <<http://www.hum.gu.se/allcach2004/AP/html/prop156.html>> ; Stephen Ramsay, Geoffrey Rockwell, Stéfan Sinclair, *TAPoRware: Simple Portal Tools for Text Analysis* <<http://www.hum.gu.se/allcach2004/AP/html/prop136.html>>
11. <<http://www.isotopicmaps.org/rm4tm/>>
12. <<http://www.topicmaps.org/xtm/1.0/>>
13. For an introduction to TopicMap see Steve Pepper, *The TAO of Topic Maps, finding the way in the age of infoglut* <<http://www.gca.org/papers/xmlleurope2000/papers/s11-01.html>> .
14. <<http://www.ontopia.net/omnigator/models/index.jsp>>
15. <<http://tm4j.org/>>
16. John Bradley, "A Model for Text Analysis Tools" <<http://1lc.oupjournals.org/cgi/content/abstract/18/2/185>>
17. See John A. Walsh, "Topic Maps and TEI-Encoded Literary Texts", <<http://drh2004.ncl.ac.uk/abstract.php?abstract=177>>
18. Stuart Brown, "A Topic Map for the TEI" <<http://www.tei-c.org/Members/2003-Nancy/index.html#SB-abs>>

## Bibliography

*Anastasia*. Accessed 2005-05-19. <<http://anastasia.sourceforge.net/>>

Bradley, John. "A Model for Text Analysis Tools." *Literary and Linguistic Computing* 18.2 (2003): 185-207. Accessed

2005-05-19. <<http://llc.oupjournals.org/cgi/content/abstract/18/2/185>>

Brown, Stuart. "A Topic Map for the TEI." TEI Consortium, 2003. <<http://www.tei-c.org/Members/2003-Nancy/index.html#SB-abs>>

Kumar, Amit, et al. "teiPublisher a repository management system for TEI documents." Paper delivered at the ALLC/ACH 2004 Conference, Göteborg. 2004. Accessed 2005-05-19. <<http://www.hum.gu.se/allcach2004/AP/html/prop118.html>>

Pepper, Steve. "The TAO of Topic Maps, finding the way in the age of infoglut." Paper delivered at the XML Europe 2000 Conference, Paris. 2000. Accessed 2005-05-19. <<http://www.gca.org/papers/xml europe2000/papers/sl1-01.html>>

Ramsay, Stephen, Geoffrey Rockwell, and Stéfan Sinclair. "TAPoRware: Simple Portal Tools for Text Analysis." Paper delivered at the ALLC/ACH 2004 Conference, Göteborg. 2004. Accessed 2004. <<http://www.hum.gu.se/allcach2004/AP/html/prop136.html>>

TAPoRware. Accessed 2005-03-11. <<http://cheiron.mcmaster.ca/~taporware/>>

teiPublisher. Accessed 2005-05-19. <<http://teipublisher.sourceforge.net/docs/index.php>>

Versioning Machine. Accessed 2003-12-09. <<http://mit.h2.umd.edu/products/ver-mach/>>

Walsh, John A. "Topic Maps and TEI-Encoded Literary Texts." Paper delivered at the Digital Resources for the Humanities Conference, Newcastle Upon Tyne. 2004. Accessed 2005-05-19. <<http://drh2004.ncl.ac.uk/abstract.php?abstract=177>>

XPhilologic. Accessed 2005-05-19. <<http://barkov.chicago.edu/xphilo/>>

Zimmerman, Matthew. "Using AMP technology (Apache, MySQL, PHP) for XML publication." Paper delivered at the ALLC/ACH 2004 Conference, Göteborg. 2004. Accessed 2004. <<http://www.hum.gu.se/allcach2004/AP/html/prop156.html>>

---

## The Computed Synoptic Table —Tele-Synopsis for Biblical Research

---

**Maki Miyake** ([mmiyake@dp.hum.titech.ac.jp](mailto:mmiyake@dp.hum.titech.ac.jp))

*Department of Human System Science, Tokyo  
Institute of Technology*

**Hiroyuki Akama** ([akama@dp.hum.titech.ac.jp](mailto:akama@dp.hum.titech.ac.jp))

*Department of Human System Science, Tokyo  
Institute of Technology*

**Masanori Nakagawa**

([nakagawa@nm.hum.titech.ac.jp](mailto:nakagawa@nm.hum.titech.ac.jp))

*Department of Human System Science, Tokyo  
Institute of Technology*

**Nobuyasu Makoshi** ([makoshi@gsic.titech.ac.jp](mailto:makoshi@gsic.titech.ac.jp))

*Global Scientific Information Center, Tokyo  
Institute of Technology*

---

### I. Introduction

While over the last two centuries, 'the synoptic problem' has been one of the controversial subjects in the studies of the *New Testament*, only a few studies so far have attempted to give an objective, statistical explanation of the mutual relationships between the synoptic Gospels, Matthew, Mark and Luke (in abbreviation, Mt, Mk and Lk, respectively) (Conzelmann and Lindemann 45-53). Furthermore, even though a large number of studies have made various assumptions of their genealogical interdependence, there still seems to remain a lack of the computational humanities technology enabling the Gospel researchers to present valid arguments based on a huge amount of biblical text data. As the first step of our study, there is a need to develop some specific applications to automatically collect the thorough data of the lexical usage patterns from the electronic bible (Miyake, Akama, Sato and Nakagawa 2002), thus the web-based biblical software, named *Tele-Synopsis* (<http://nerva.dp.hum.titech.ac.jp/tele-synopsis/parallel>), is designed to gather information of the word usage under various conditions and to help further statistical approach to the origin of the variant texts.

## II. Tele-Synopsis — Web-based biblical software

The basic concept design of *Tele-Synopsis* is founded upon the possibilities of natural language processing (NLP) for mediating Thesaurus creation and Conceptual mapping, dual problematic fields whose key concept is always cognition of 'frame' (Minsky; Winston 211-277). *Tele-Synopsis*, which allows us to manipulate lexical data of parallel and variant texts (Miyake, Akama, Sato, Nakagawa and Makoshi 2004), uses the NA27th version of the texts (Nestle-Aland) and for the parallels, the *Synopsis Quattuor Evangeliorum* by Kurt Aland, recognized as the most reliable parallel synoptic table (PST) to date. This system has a merit to make it possible for users to independently add and remove each sentence so as to customize their own synoptic table by changing the temporary segmentation of pericope, yet the challenges are still left on the optimum solutions available to the users, and so we need a sort of 'TextTiling' algorithm that allows us to break parallel texts into units the most suitable for biblical research.

## III. Segmentation Problem

Although there are traditionally two types of synoptic tables covering a lost source called the 'Q' (Mt and Lk) and Mark (Mt, Lk and Mk) respectively, few trials have been done to produce synoptic tables treating other combinations of two Gospels, such as Mt and Mk, Lk and Mk. This kind of inexhaustiveness is due to the *raison d'être* of the synoptic tables that is to consolidate *Two-Source Hypothesis*, according to which Mk and the 'Q' are the origins of quotations (Kloppenborg et al. and Reader). In addition, we have to note that the two traditional synoptic tables were solely made by using a *Form Criticism* which divided the texts into parts by the arbitrary unities coming from tradition or reduction. It can be recognized that the two traditional synoptic tables 'mesh' the world of the Gospels too roughly (as is the case for the Markan triptych table) or too finely (for the bilateral table of the 'Q'). As long as the problem of text segmentation remains unresolved, any experiment in quantitative text analysis will be still a long way from being realized. For our goal of the scientific examination of the Two-Source Hypothesis, we propose a new statistical method of generating the segmentation criteria of the synoptic Gospels, a sort of 'TextTiling' methodology enabling a computed synoptic table (CST) with an objective segmentation based on objective criteria.

## IV. The Computed Synoptic Table(CST)

The computed synoptic tables (CST) are produced by using the algorithm called Synoptic Patch (Figure 1) that consists of the combination of 1) N-gram calculation, 2) Windowing data gathering and 3) TextTiling method.

### 1) Data from the n-gram model

We calculated for the 3 parallel texts (Mk,Mt,Lk) all the cases of n-gram models, thus made an exhaustive list of the instances where words co-occurred across texts. These overlaps were classified by the four combination patterns (D:Mt-Lk, C:Mk-Lk, B:Mk-Mt, A:Mk-Mt-Lk) (Figure 2), and the longest matched strings of words can be thought of as proofs of cross-citation. Having in view the occurrence probability of N-gram instances, we extracted the overall data under the condition of ( $N > 3$ ) because the significance of the bi-gram data is relatively low. This process will allow us to build a more objective synoptic table to replace the traditional one.

### 2) Data obtained by a windowing method

It is well-known that there has been in the realm of Information Retrieval (IR) remarkable progress owing to the elaboration of what we call vector space model or concept-based IR. This method, that consists of collecting the information about term  $i$  occurring  $n$  times in document  $j$ , allows us to identify a word (or a document) using a  $k$ -dimensional vector representation. Each entry of the vector corresponds to the frequency of each of  $k$  co-occurring words. Then the similarity between documents will be computed by the cosine of the angle between these vectors in a  $k$ -dimensional Euclidian space. Taking into consideration the principle that a context-sensitive word (or string of words) is categorized by the neighbor words appearing within a certain distance from it, we implemented some functions to set up a set of synchronized windows changing in size for each parallel n-gram instance (longest matched strings of words) to be centered in. The rule of the window operation for recording one by one and simultaneously in the parallel texts the frequency data of the co-occurring words is that each window must stop the extension if the border meets that of the previous (when moving leftward) or the next (when moving rightward) pericope.

### 3) Application of TextTiling

Synoptic Patch as a method of partitioning off the texts allows us to calculate at every step of the window extension the correlation coefficient between the word frequency vectors generated from each corresponding window instance. Before

the extending operation, the cosine similarity value remains 1, but as different words are being distributed in the parallel setting, this value begins to decline and continues to fall down until another parallel N-gram instance is met in the window extension (cohesion score graph used in 'TextTiling' (Hearst 33-64)). However, in each pericope, there may be several instances of centered key strings (a series of the longest matching words) that are supposed to produce an overlap of windows and descending similarity curves, so that we computed at each word position the mean of the correlation coefficients obtained from all the pairs of parallel word vectors inside a pericope. The threshold is determined by us at 0.5 to properly resegment the pericope because the traditional synoptic tables with the three Gospels tends to include in each frame many divergent passages making the parallel word vectors nearly non-correlated or sometimes too highly correlated. That is why we fixed the segmentation point by using the threshold for the cohesion score graph instead of selecting, just as Hearst recommends it, the steepest part of the descending curve.

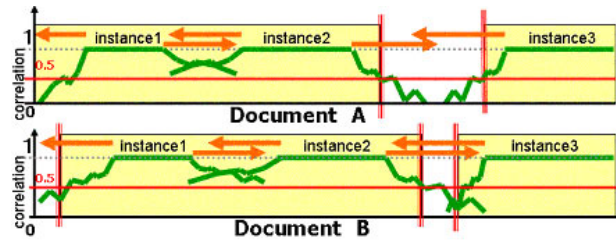


Figure 1

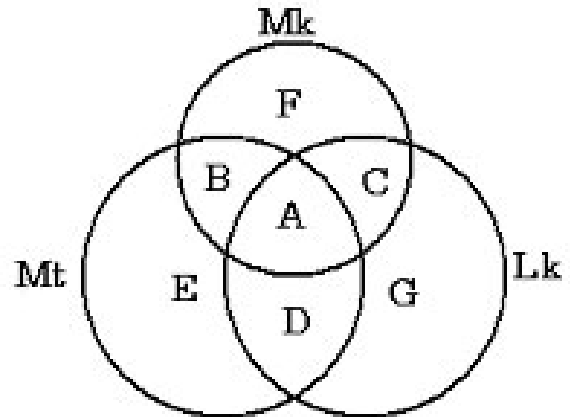


Figure 2

## V. Result and Conclusion

The Synoptic Patch allows us to produce by fulfilling the identical criteria two remaining bilateral synoptic tables allocating Mk and Mt for one and Mk and Lt for the other. The index of difference between the traditional Synoptic Tables (ST) and the Computed Synoptic Table (CST) can be defined by the distribution of the words into the 7 categories as shown in Figure 2. The effects of the new combinations are clearly revealed by the diminution in quantity of some textual overlaps. The ratio of the common parts (A+B+C+D) is 60% in the PST and 42% in the CST (Figure 3). Figure 4 shows the drop in number of the words belonging to the categories A and D whose considerable weights would support the two source hypothesis. It cannot be denied that the new balance between the original parts E, F and G (increasing) and the common parts A+B+C+D (decreasing) will influence the verification regarding the historical formation of the synoptic Gospels. We can instinctively grasp the changing features of the parallels attachment by horizontally comparing the two tables in Figure 5. It will be left for the future investigations to completely evaluate the efficacy of the CST. Further information will be obtained at : <http://nerva.dp.hum.titech.ac.jp/tele-synopsis/synopsis.html> .

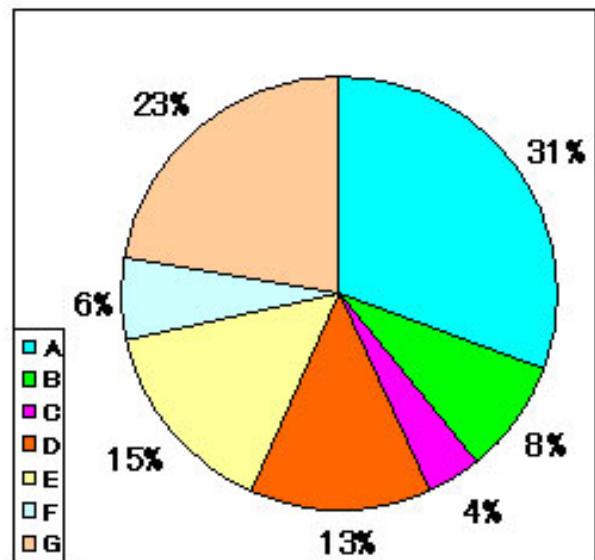


Figure 3

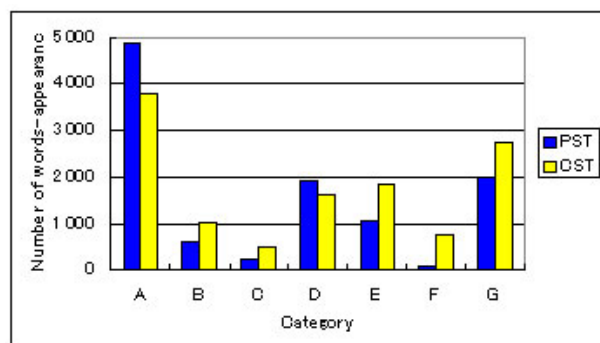


Figure 4

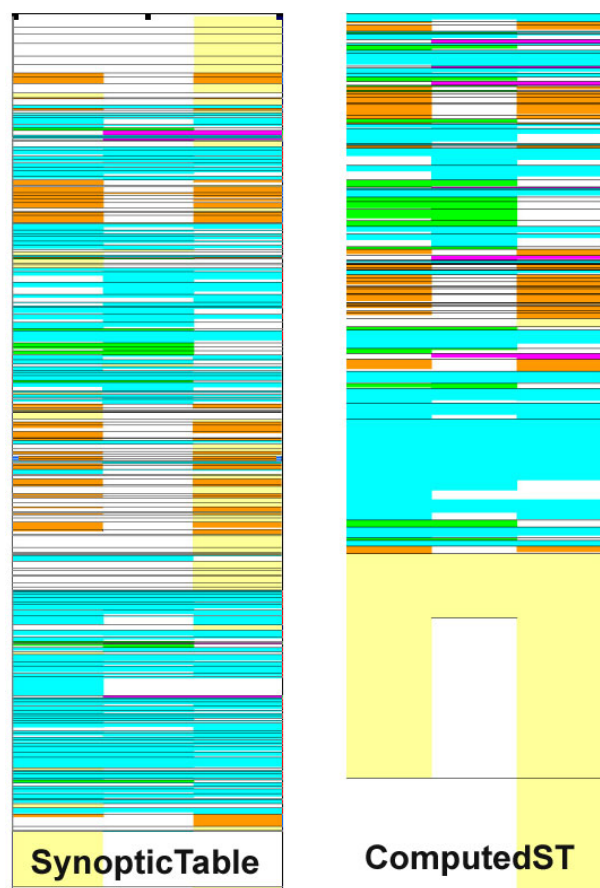


Figure 5

Conzelmann, H., and A. Lindemann. *Interpreting The New Testament*. Trans. Siegfried S. Schatzmann. Peabody, Mass.: Hendrickson Publishers, 1988.

Hearst, Marti A. "Segmenting text into multi-paragraph subtopic passages." *Computational Linguistics* 23 (1997): 15-36.

Kloppenborg, John S. *Q Thomas Reader*. Sonoma, Calif: Polebridge Press, 1990.

Minsky, M.L. *A Framework for representing knowledge*. Cambridge: Massachusetts Institute of Technology A.I. Laboratory, 1974.

Miyake, M., H. Akama, M. Sato, and M. Nakagawa. "Approaching to the Synoptic Problem by Factor Analysis." *Proceedings of the Institute of Statistical Mathematics* 48.2 (2002): 327-337.

Miyake, M., H. Akama, M. Sato, and M. Nakagawa. "Tele-Synopsis for Biblical Research." *Proceedings of the IEEE ICALT*, 2004. 931-935.

Nestle, Erwin, and Kurt Aland, et al., eds. *Nestle-Aland Novum Testamentum Graece*. 26th ed. Stuttgart: Deutsche Bibelstiftung, 1979.

## Bibliography

Winston, Patrick Henry, and Berthold Horn. *The Psychology of Computer Vision*. New York: McGraw-Hill, 1975.

Aland, Kurt. *Synopsis of the Four Gospels*. 9th ed. Stuttgart: German Bible Society, 1989.

## El Trabajo Final de Carrera en Filología: Perspectivas Hacia un Nuevo Horizonte

---

**Isabel Clara Moll Soldevila** (*imoll@uoc.edu*)

*Universitat Oberta de Catalunya*

**Laura Borràs** (*lborras@uoc.edu*)

*UOC/Hermeneia*

**Roger Canadell** (*rcanadell@uoc.edu*)

*Universitat Oberta de Catalunya*

---

**E**sta comunicación se inscribe en el marco de la investigación llevada a cabo durante los últimos cuatro años por el grupo de investigación *Hermeneia: Estudios Literarios y Tecnologías Digitales*. En este período se ha priorizado una de las líneas de investigación considerada clave en el contexto docente y académico en que el grupo ha llevado a cabo su actividad — que es el de la Universitat Oberta de Catalunya (UOC) —, nos referimos a la observación, el análisis y la evaluación de los métodos de enseñanza virtual en un contexto universitario. Dicho grupo no únicamente ha estado pendiente de la evolución de algunas de las asignaturas que se impartían en la licenciatura, tomando como referencia el plan de estudios de filología catalana de la UOC especialmente adaptado para un entorno virtual de aprendizaje, sino que además ha considerado clave la implantación de asignaturas concebidas únicamente para ser impartidas en dicho entorno virtual a partir de materiales que explotan de una manera real el concepto de trabajo hipertextual. El trabajo de licenciatura de filología catalana (TFC) aunque no es en esencia una de estas asignaturas sí que constituye un banco de pruebas para los estudiantes que están al punto de concluir un ciclo de aprendizaje universitario desafiador, inquietante, pero también completo y gratificante. En este sentido, muchos de los estudiantes que tras numerosos semestres de estudio se matriculan en el TFC de filología son personas alfabetizadas en el uso de las nuevas tecnologías en mayor o menor grado, familiarizados con un tipo de enseñanza a distancia y virtual, conscientes de la eficacia de este tipo de enseñanza, especialmente sensibilizados por los retos que les plantea la sociedad de la información. Pero sobre todo, están motivados por poner en práctica no sólo algunos de los conocimientos prácticos y técnicos que han ido adquiriendo a lo largo de la licenciatura sino también por explotar al máximo una

concepción nueva del conocimiento y de su difusión a través de las posibilidades que les brindan las nuevas tecnologías.

Nuestra intención es explicar el funcionamiento de una asignatura singular dentro del plan de estudios por su especificidad formativa al final de un ciclo para después mostrar algunos de los resultados más significativos que se han gestado a la luz de esta asignatura y que están, además, umbilicalmente vinculados al proyecto de investigación *Hermeneia*. El TFC es una asignatura obligatoria de doce créditos que curricularmente está situada al final del recorrido académico que plantea la licenciatura de filología catalana de la UOC. En su momento fue diseñada y concebida con unas características especiales con respecto al resto de las que ofrece dicha licenciatura en la UOC. De entrada es una asignatura que se prolonga durante dos semestres académicos puesto que entre sus objetivos se cuentan los de proponer interrogantes, ejercitar habilidades y avanzar actitudes en el momento de afrontar un trabajo de investigación a nivel de licenciatura, así como estimular a los estudiantes para que al final del proceso presenten públicamente un trabajo de investigación original e inédito. El proceso que supone esta asignatura y, a su vez, el esfuerzo que deben realizar los estudiantes requiere, por lo tanto, la distribución de los créditos requeridos en dos semestres, preferiblemente consecutivos. En esta asignatura también está prevista la actuación de dos figuras docentes: la del tutor de contenidos o director del TFC y la del consultor metodológico o asesor externo y formal del trabajo. La presencia de estos dos 'agentes' representa una diferencia notable respecto a la organización de las demás asignaturas. Por no mencionar el papel más bien secundario que adquiere el aula virtual en este proceso o la necesidad de establecer un encuentro presencial al final de todo el proceso para que el estudiante pueda mostrar, presentar, justificar y defender el proyecto ante un evaluador externo, así como el resto de los compañeros, profesores, consultores de la licenciatura. En esta sucinta descripción hemos querido resaltar las diferencias que separan el TFC del planteamiento de otras asignaturas y que la convierten en una peculiar 'pieza' que encaja en el marco de la licenciatura de filología catalana de la UOC.

Uno de los aspectos esenciales de la asignatura y que resulta imprescindible destacar es el hecho de que el grado de 'libertad' que experimenta el estudiante en esta asignatura no es comparable a ninguna de las asignaturas previas que ha debido superar. Entre el material de la asignatura existe un menú de proyectos de investigación posibles suficientemente amplio para que el estudiante pueda escoger el marco al cual debe ceñir su experiencia investigadora. Este menú es diverso y transversal en ocasiones, aplicado o teórico en otras, pero lo que prima en estas opciones de que dispone el estudiante es la potenciación de su capacidad imaginativa, sus intereses y sus inquietudes y la posibilidad de concretar todo ello en una propuesta asequible que le permita superar la asignatura. Desde mi experiencia

como consultora de la asignatura puedo afirmar que hay pocos casos en los que el estudiante no haya invertido tiempo y esfuerzos suficientes para obtener trabajos de un alto nivel académico: buenos ejemplos de tesis de licenciatura.

Aún así, queremos centrar una parte de nuestra exposición en la presentación de algunos de los trabajos que han explotado hasta sus últimas consecuencias las posibilidades que les brindaban las nuevas tecnologías. Son trabajos que han estado gestados en el área de la literatura y las nuevas tecnologías. Cada uno desde un enfoque teórico y práctico diferente y partiendo de áreas literarias diversas (la poesía, la literatura infantil y juvenil o la temalogía) han conseguido resultados destacables este área de confluencia que hemos señalado. La explicación y el análisis de estos trabajos obedece a varios factores entrelazados. Por una parte, cabe destacar que la tutora o directora de las tres propuestas que presentaremos es, a su vez, la directora del grupo de investigación *Hermeneia*, y en este sentido, la concepción y la génesis de estas iniciativas de investigación quedan insertadas en un terreno que aún podemos llamar experimental. Además, nuestra implicación estos casos particulares responde a una doble motivación la de consultora del TFC y la de investigadora del grupo *Hermeneia*. (Imágenes 1 y 2)

área de la literatura y que hayan sido realizados por estudiantes que han cursado una determinada asignatura de Teoría Literaria y Literatura Comparada que se ofrece desde hace varios semestres en la licenciatura. Será también necesario, en su momento, mencionar que la persistencia y el cambio de concepción en relación a la investigación y difusión en literatura fueron clave para salvar los temores iniciales de estos estudiantes al formato digital. En cualquier caso, la posibilidad de profundizar en estos ejemplos concretos de la asignatura de TFC nos permitirá explorar las posibilidades no siempre evidentes de adentrarnos en el fructífero terreno en el que confluye la literatura, los métodos de investigación y difusión a nivel de licenciatura universitaria y las nuevas tecnologías.



Imágen 1



Imágen 2

Finalmente, es importante señalar una serie de coincidencias que abordaríamos con más detalle en la comunicación, como el hecho de que los trabajos que presentaremos pertenezcan al

## An Examination of the Rhetorics of Digital Scholarship and the Emerging Digital Monograph

---

*Elli Mylonas (elli\_mylonas@brown.edu)*

*Brown University*

---

A variety of rhetorics have been applied to digital compositions: the "rhetoric of hypermedia" (e.g. Landow), "rhetoric of multimedia" (e.g. Liestol), and "rhetoric of new media" (e.g. Chun, Manovich) are all terms that have been used. This work has often focused on the construction of a digital work at the level of nodes and links, more than on its narrative techniques. Landow's rhetoric of arrivals and departures, for instance, looks at the relationships between individual nodes, and their effect on one another via the rhetoric of the link. Bernstein, in his discussion of structural patterns in a hypertext also focuses on relatively smaller groups of nodes; while Bernstein's patterns are clearly intended to be composed into larger structures, or to provide components in the analysis of a larger structure, he does not elaborate how this composition or analysis works at the level of a whole work. Liestol and Fagerjord are more concerned with the narrative construction of digital publications. They are mainly studying the digital documentary, an expository medium which is similar in form and intent to scholarly hypertexts. Students of digital fiction have analyzed larger structures and discussed how meaning emerges from them (e.g. Walker), but there have not been many discussions of overall rhetorical structuring in scholarly web publications.

Projects or publications on the web with academic research subjects tend to fall into several discrete types. A partial list follows:

- Conventional linear monographs in digital form, differing from print monographs only in their medium of publication. These digital publications will not be discussed in this paper.
- A publication or dissemination of a primary source. Text-bases or archives are the main representatives of this class. They are intended for several levels of expertise, and readers can search through them on their own terms.
- A collection of primary and secondary sources meant to be explored by diverse audiences for more than one purpose. These sites have an encyclopedic, expository nature, and tend to surround primary source material with secondary sources.

- A collection of primary or secondary sources that explicitly represent a particular point of view, or publish a particular phase of research. These sites, with their combination of secondary and primary material are the prototypes of the *digital monograph*.
- A collaborative, annotatable, space that may only be accessible to a small group of authors. These sites are composed of collections of texts or other sources. They resemble laboratories, where research is on-going and benefits the participants more than the external reader.

Collaborative, emergent projects are not meant to read by third parties as much as by the participants. They are exploratory and communicative spaces for their participants, and as such, the rhetoric they exhibit is often one of annotation and conversation. Wikis provide a mode of publication that encourages organic growth and collaboration. *The Ivanhoe Game* also shares this modality, although it isn't a means of publication.

Although any collection or publication that has been consciously selected inherently reflects a bias and represents a particular point of view, some web-based projects are more explicit than others about their motivation and their purpose. Scholarly projects that have as their goal an edition or a collection of texts generally identify the guiding principles for their selections. Examples of such projects are text and image archives, like the WWP ( <<http://www.wwp.brown.edu>> ) and *The Empire that Was Russia* ( <<http://www.loc.gov/exhibits/empire>> )

When a scholar embarks on a project whose goal is to elucidate a particular topic through primary and secondary materials, biases become harder to tease out. Often, a project that originates as a research topic turns into a general expository publication whose goals are to collect and disseminate primary sources, and to contextualize them with secondary sources. Such a publication is inherently user-centered. However, because the intent of a researcher or student cannot be foreseen in advance by the creator of the publication, the project is often designed to allow multiple access and navigational modes, so that it can accommodate a wide variety of uses. The digital medium encourages this, and invites an author to attempt to provide many things to many people: syllabus, guided tour and resource collection, for example. Any authorial voice or point of view is overwhelmed by the encyclopedic omniscience and protean presentation of the publication.

Some publications avoid multi-linear, multi-access presentation even as they follow this recipe of primary sources surrounded by a matrix of contextualizing material. Instead, they impose a very rigid form, as in some instructional materials, where it is not possible to deviate from a prescribed path beyond a limited set of carefully scripted choices, if any. This can serve



to hide bias or point of view, by making it harder to question or challenge the publication.

Finally, there are digital publications that resemble the print monograph in that they prioritize a set of research results and have an identifiable authorial voice. Primary source materials coexist with authored text, but they are juxtaposed in support of the argument that is being made. Like the publications described above, these publications may also allow a reader multiple navigational strategies, but they privilege a single narrative thread. The interface and navigational structure are designed to reflect the argument, and the principles used in selecting primary source materials are clear. Readers familiar with the site or its material may only want to consult some primary sources, so it is possible to navigate directly to identifiable nodes, but this kind of navigation is clearly secondary. One of the best-known web based monographs is Thomas and Ayers, 2003. One of the earliest is Kolb, 1994.

The authored, digital monograph allows a scholar to do things that are difficult in a print publication. It is possible, for example, to present a coherent argument while weaving several different threads and approaches to a problem; it is also possible to include supplementary material that would require a distracting digression in a traditional publication. Conventional articles are constrained by length limits and by the rhetoric of the page to making one point. A digital monograph allows an author with single argument to present multiple threads of argumentation and discussion which will augment one another, but which are not necessarily interdependent.

This paper will use examples from projects the author has participated in, as well as others to discuss emerging features of the digital monograph, and to compare them with other types of scholarly, digital publication. It will look at interface and interaction design, as well as at the information design underlying the publication.

## Bibliography

Bernstein, Mark. "Patterns of Hypertext." *Proceedings of the Ninth ACM conference on Hypertext and hypermedia, Links, Objects, Time and Space (Pittsburgh, PA)*. NY, NY: ACM Press, 1998. 21-29.

Bernstein, Mark. "Structural Patterns and Hypertext Rhetoric." *ACM Computing Surveys* 31:4es, Article 19 (December 1999).

Chun, Wendy Hui Kyong. *The Rhetoric of New Media (MC0150, Brown University)*. Senior Seminar course; no further information available.

Fagerjord, Anders. "Rhetorical convergence: studying web media." *Digital media revisited: theoretical and conceptual*

*innovation in digital domain*. Ed. G. Liestol, A. Morrison and T. Rasmussen. Cambridge, MA: MIT Press, 2003. 293-326.

*The Ivanhoe Game*. Accessed 2002-03-31. <<http://www.speculativecomputing.org/ivanhoe/index.html>>

Kolb, David. *Socrates in the Labyrinth*. Watertown, MA: Eastgate Systems, 1994.

Landow, George. *Hypertext 2.0*. Baltimore: Johns Hopkins University Press, 1997.

Liestol, Gunnar. *Essays in Rhetorics of Hypermedia Design*. Oslo: Department of Media and Communication, University of Oslo, 1999. (Doctoral Dissertation)

Thomas, William G., and Edward L. Ayers. "The Differences Slavery Made: A Close Analysis of Two American Communities." *American History Review* (December 2003). <<http://www.vcdh.virginia.edu/AHR/>>

Walker, Jill. "Together and Tearing Apart: Finding the Story in afternoon." *Proceedings of the 10th ACM Conference on Hypertext and Hypermedia: Returning to Our Diverse Roots (Darmstadt)*. NY, NY: ACM Press, 1999. 111-117.

Walker, Jill. "Fiction and Interaction: How Clicking a Mouse Can Make You Part of a Fictional World." Diss., Department of Humanistic Informatics, University of Bergen, 2003.

## Testing EAD Encoding in the *Texas Archival Resources Online* (TARO) System with Textual Analysis Techniques

---

**Vidya Narayan** ([narayanv@ischool.utexas.edu](mailto:narayanv@ischool.utexas.edu))

*School of Information, University of Texas at  
Austin*

**Patricia Galloway** ([galloway@ischool.utexas.edu](mailto:galloway@ischool.utexas.edu))

*School of Information, University of Texas at  
Austin*

---

**E**lectronic archival finding aids encoded in Encoded Archival Description (EAD) are transported across networks and rendered into HTML for display on the browser. Considering the time, effort and money involved in marking up the finding aids, has the markup been used for retrieval purposes? Has the multilevel hierarchical nature of finding aids been used for searching? A few online EAD tag based retrieval systems that process queries look for occurrences of the search term in the corresponding EAD tag, but do not seem to address subject- or topic-based queries. This study explores the possibility of using the content of specific EAD tags for subject retrieval purposes. We studied the consistencies, commonalities and discrepancies in usages of various critical tags across repositories participating in the *Texas Archival Resources Online (TARO)* project. These usages were compared to EAD tagging guidelines as well as TARO guidelines. We identified the <abstract>, <scopecontent> and <bioghist> tags as good representatives of the finding aid from standard archival descriptive practice and examined their content for a sample of repositories within TARO. The content of these tags was processed using text processing techniques to further study and arrive at possible similarity metrics to identify similar finding aids. We feel this would help evaluate EAD as an information retrieval tool within TARO and if our experiments help conclude that EAD can be effective as such a tool (or can be made effective by better descriptive practice), then the prospect of creating a highly interconnected web of finding aids exploiting the hierarchical nature of EAD is possible.

This study was conducted on 1226 EAD encoded finding aids from nine archiving institutions which are part of TARO. Our study was conducted in three phases. First, we verified the usage of EAD tags across repositories within TARO with an

aim of determining if there exists a core set of tags within these finding aids. This part of the study was motivated by the underutilization of the Dublin Core tags as reported by Shreves, Kirkham, Kaczmarek and Cole and Ward. From this part of our study we arrived at a core set of 27 EAD tags from the entire EAD tag library comprising 146 tags. These 27 EAD tags form a superset of the tags deemed mandatory by ISAD(G) as well as the EAD tagging guidelines of other archiving institutions. Additionally, we observed the varied usage of the hierarchy of these tags within these finding aids and very limited usage of tags to achieve electronic linking between documents.

In the second part, we studied if these finding aids have been encoded according to standard archival descriptive practice (i.e. if the text within these EAD tags was appropriate). This was achieved through text processing involving extraction of the text from the specific tags and processing these to arrive at a vocabulary. We conducted this study on the part of the finding aids corresponding to the *University of Texas Alexander Architectural Archive (UTAAA)* and *University of Texas Benson Latin American Collection (UTLAC)* repositories. Comparisons, of the vocabularies of the <abstract> tag between two different repositories indicate that the vocabularies for the said repositories are quite distinct. We found that the content of different tags has different word counts and correspondingly different vocabulary sizes. Additionally, we observed that up to 65% of the total word count in each of the three tags studied (<abstract>, <bioghist> and <scopecontent>) represents the vocabulary, thus indicating that significant information is embedded in the textual content of each of these tags.

In the third step, using the vocabularies obtained, we represented these finding aids as vectors in the vocabulary space. In such a vector representation of finding aids, we compared finding aids using a cosine similarity in conjunction with Term Frequency-Inverse Document Frequency (TF-IDF) weighting. The TF-IDF scheme weights rarely used words higher than commonly used words, and also accounts for the size of the document. We then clustered these finding aids with an online clustering tool (wCLUTO) using the agglomerative clustering algorithm. The agglomerative clustering groups finding aids based on the similarity of content, resulting in a tree of documents. The lowest levels of the tree correspond to individual finding aids and the highest levels of the tree correspond to the entire sets of finding aids. Our study focused on low-level clusters, which are of particular interest to archivists, as these clusters address the descriptive material embedded in the various EAD tags. To determine the similarities between finding aids, we extracted vocabularies for individual tags like <abstract>, <scopecontent> and <bioghist> and clustered the finding aids based on the similarity of textual content with respect to these individual tags. Further, we combined the similarity relations between

finding aids, based on these individual tags, to build a space that encompasses the content similarity for a combination of tags. Our clustering results on individual and combination of tags are in agreement with the classification provided by the curators of the *UTAAA* repository.

We conclude from our study that if finding aids are marked up according to standard archival descriptive practice then they yield meaningful content-based clusters of similar finding aids. Further, we were able to demonstrate; i) the ability of forming 'neighborhoods' of similar finding aids using either individual tags or a combination of tags, and, ii) that the 'neighborhoods' were different for different tags or combination of tags. From this idea of a 'neighborhood' of finding aids, we propose a searchable interface for a repository of finding aids by means of the EAD tags. This search facility, we think, enhances the prospect of creating a web of similar and, thus, interconnected finding aids, which, in turn would facilitate research in the field of archives and help researchers form cliques by common research interests and goals.

Our study demonstrates the ability to apply the text processing techniques from the field of information retrieval to the field of archives with a goal of enabling EAD encoded finding aids transition to the digital world and be visible in the realm of online documents and be accessible to researchers.

## Bibliography

- Shreves, S.L., C. Kirkham, J. Kaczmarek, and T.W. Cole. "Utility of an OAI Service Provider Search Portal." *Proceedings 2003 Joint Conference on Digital Libraries*. Los Alamitos, CA: IEEE Computer Society, 2003. 306-308.
- Ward, J. "A Quantitative Analysis of Unqualified Dublin Core Metadata Element Set Usage within Data Providers Registered with the Open Archives Initiative." *Proceedings 2003 Joint Conference on Digital Libraries*. Los Alamitos, CA: IEEE Computer Society, 2003. 315-317.

## The *LICHEN* Project: The Linguistic and Cultural Heritage Electronic Network

**Lisa Lena Opas-Hanninen**

(*lisa.lena.opas-hanninen@oulu.fi*)

University of Oulu

**Jean Anderson** (*j.anderson@arts.gla.ac.uk*)

University of Glasgow

**Ilkka Juuso** (*ilkka.juuso@ee.oulu.fi*)

University of Oulu

**Tapio Seppänen** (*tapio@ee.oulu.fi*)

University of Oulu

The international, interdisciplinary and multilingual LICHEN project focuses on the languages and cultures of the northern circumpolar region, that is the region north of the 55th parallel. Its underlying assumption is that language and culture are as important to the survival and well-being of populations as more obvious ecological, social and health issues. We believe that the creation of a digital portal giving access to written and spoken texts in the languages of the region will further its well-being.

Faced with minority languages, governments in the recent past have pursued policies of assimilation. This has applied to indigenous languages in Canada, to Gaelic and Scots in Scotland, and to Finnic minority languages in the Circumpolar region: Meänkieli and Swedish Finnish in Sweden; the Kven language in Norway; Viena Karelian, Olonets Karelian and Vepsian in Russia; the Võro and Seto languages in Estonia; and Livonian in Latvia.

LICHEN aims to collect and disseminate information about the languages spoken in the circumpolar region in order to promote the linguistic confidence and self-image of their speakers. It will promote cultural awareness among the peoples of the North, facilitating cross-cultural communication between them in an age of rapid global change. LICHEN will create communication between research units in order to promote discussion on the common needs of research on the minority languages of the North. We are doing this by:

- creating an electronic framework for the collection, management, online display, and exploitation of corpora of the languages of the circumpolar regions;
- creating a website with information on these languages and the peoples speaking them (the LICHEN website will be launched in January 2005);
- creating a virtual learning environment for teaching the linguistic and cultural heritage of the circumpolar region;
- carrying out a pilot project on the Meänkieli and Kven languages;
- identifying research on topics of immediate importance and common interest;
- setting up an inter-institutional doctoral research programme.

LICHEN has existing resources and work has begun. We have Meänkieli tapes totalling about 150 hours and Kven language tapes of about 100 hours for our immediate use. These tapes are now being digitized. We have access to both the structure and contents of the *Scottish Corpus of Texts and Speech* (SCOTS) at the University of Glasgow, currently totalling 0.5 million words of spoken and written Scots and Scottish English. We have the nucleus of a research team based on the English and Finnish Departments and the Department of Electrical Engineering at Oulu and the English Language Department and SCOTS project at Glasgow. This team has considerable linguistic and computing expertise.

Our first aim in year 1 is to complete the technical specifications for the electronic framework through consultations between language and computing staff team members. We are also consulting other people working in the field of corpus building at meetings and conferences, and by email. In addition to housing the data, the system will accommodate management, administration and programs for concordancing and searching the data. The ultimate goal of the development of the computing tools is a shell which can be adapted to any language. For many languages at risk there is a need both to preserve existing materials and develop new ones. During this year, we will design and implement a prototype of this shell. A longer term goal is to provide an interface to the tools which allows the end user to define or rename all functions in their own language.

We will continue work on the Meänkieli and Kven recorded language material. We will work generally on the problem of languages without standard written forms, starting with Kven and Scots (a worldwide problem as people endeavour to record languages before they vanish). As an initial solution to the problem of written forms, it is proposed that several Kven speakers should be asked to transcribe a short passage and the results compared. Scottish Language Dictionaries will be consulted here. We will investigate the feasibility of working through community groups in minority language areas. In

addition to harnessing local knowledge, we hope that a policy of local workshops will stimulate skills development and job creation.

A poster presentation at ACH/ALLC 2005 will enable us to publicise the project to other minority language scholars and to enlist the considerable expertise of the conference participants in the discussion of the design of the online corpus tools. It is our intention that the functions of the corpus shell should include all the basic requirements for a corpus builder and for a corpus user in an easy-to-use environment. We know the end users will include many who are not technically sophisticated and would not have other avenues for finding advice on digitization or access to an Internet platform to share their materials. Our idea of 'basic' requirements for online use include corpus browsing, word and phrase searches, wildcard searches, concordancing; for corpus building we will include guidelines on recording, digitization, copyright and data protection. We would welcome this opportunity to discuss our proposed designs with the expert community of ACH/ALLC.

## Bibliography

Anderson, J., et al. "The SCOTS Corpus." *Models and Methods in the Handling of Unconventional Digital Corpora*. Houndsmills: Palgrave Macmillan, Forthcoming.

*Institute for the Languages of Scotland*. <<http://www.arts.ed.ac.uk/celtscot/institutelanguagesscotland/>>

*Kven bibliography*. <<http://www.ub.uit.no/baser/kvensk/>>

*Linguistic Atlas Projects*. <<http://us.english.uga.edu/>>

*The Linguistic Data Consortium*. <<http://www.ldc.upenn.edu/>>

*Meänkieli information*. <<http://modersmal.skolutveckling.se/meankieli/index.html>>

*MediaTeam Oulu research group*. <<http://www.mediateam oulu.fi/brief/?lang=en/>>

*The Newcastle Electronic Corpus of Tyneside English*. <<http://www.ncl.ac.uk/necte/>>

Opas, L.L., and F.J. Tweedie. "Review of Michael P. Oakes, Statistics for Corpus Linguistics." *Literary and Linguistic Computing* 14.4 (1999): 541-543.

Palander, M., L.L. Opas-Hänninen, and F.J. Tweedie. "Neighbours or enemies? Competing variants causing

differences in transitional dialects." *Computers and the Humanities* 37.4 (2003): .

Ruija kvenmuseum. <<http://museumsnett.no/alias/HJEMMESIDE/vadsomuseet/kven/>>

Ruijan Kaiku newspaper. <<http://www.ruijan-kaiku.no/>>

Scottish Corpus of Texts and Speech. <<http://www.scottishcorpus.ac.uk/>>

TAPoR tools. <<http://tapor.humanities.mcmaster.ca/home.html>>

The Text Encoding Initiative. <<http://www.tei-c.org.uk/>>

Thule Institute. <<http://thule.oulu.fi/>>

University Centre for Computer Corpus Research on Language. <<http://www.comp.lancs.ac.uk/computing/research/ucrel/>>

Winsa, Birger. "Language attitudes and social identity. Oppression and revival of a minority language in Sweden." *Applied Linguistics Association of Australia Occasional paper* 17 (1998).

## Multicultural Issues on the TEI's Horizon: the Case of Tibetan Texts

*Linda E. Patrik* ([patrikl@union.edu](mailto:patrikl@union.edu))

*Union College*

Initially developed by scholars familiar with European and American manuscripts, the TEI is now being pressed into service for the encoding of non-western texts. Will the TEI develop as a closed fraternity of western computer programmers and editor/scholars, or will it emerge as an open, egalitarian community of global scholars, who are interested in applying digital technology to the world's literature? On the one hand, there are practical issues faced by encoders of non-western texts, such as the lack of reliable Unicode for all language scripts and the economic difficulties associated with training non-western scholar/editors in TEI. On the other hand, there are hermeneutic issues raised by TEI going global which are prompted by the general understanding of texts as information rather than as the materially embodied corpus of a culture. Some of the specific issues raised by Tibetan texts in this regard will be used to examine the theoretical tailoring of TEI without presuming that non-western texts will wear TEI comfortably.

Tibetan texts present several challenges to the TEI. Among the practical issues are these: as an endangered language, Tibetan has no government support for its preservation or for its study by the international community of digitally trained scholars. As a minority language, Tibetan waits on a Unicode version guaranteed to represent all of its scripts. Despite these practical problems, because the Tibetan cultural heritage is text-based, it is an obvious candidate for TEI. Even if TEI cannot restore Tibetan texts to their original cultural function, the preservation and encoding of Tibetan manuscripts and transcripts can at least keep one of the most distinctive Asian cultural traditions on 'life support'.

There are, however, other problems raised by the modern encoding of traditional Tibetan texts—problems that are hermeneutic in nature, pertaining to our underlying understanding of what a text is and what needs to be encoded in a text. *Informatics*, as used by Haraway and Hayles, is a theoretical and operational paradigm for treating a text as abstract information that can be taken out of its material medium and its cultural context. The TEI aims to describe a text so that it can be searched, stored and circulated digitally without a loss of its most meaningful features, and in doing so TEI favors the

information in the text that can abandon its material medium and perhaps even its cultural role. But information—i.e., 'informing readers'—is not what many Tibetan texts were meant to provide; the texts were meant to transform readers, helping them become enlightened. It is not just that Tibetan texts are sacred; they are practice texts that demand specific kinds of mental preparation, specific kinds of bodily handling and gestures linked to their materiality, as well as ongoing religious commitment from the readers.

With regard to readers' mental preparation, traditional Tibetan libraries contain many 'restricted' texts, which are not given wide circulation because of the original esoteric meaning of these texts. Certain advanced texts were 'classified', so to speak; they were meant only for meditator/scholars who met certain conditions: they had achieved the right qualifications for understanding the meaning of the texts, had received permission from their teacher to study or use the texts, and had made a commitment to practice the meditation methods described in the texts. These advanced Vajrayana texts are, to this day, not made accessible by traditional Tibetan teachers to any reader for the asking. (In some sense, this problem of Tibetan esotericism is similar to the esotericism of the TEI itself: the TEI tags are a hidden code, developed by, and meant for, a fairly limited group of advanced practitioners of XML.) The TEI, which is a digital publishing tool as much as it is a tool for textual analysis, may make it easier to bring formerly restricted Tibetan texts into public circulation. Although encoding these texts would help Tibetan scholars and advanced practitioners analyze the texts, encoding also makes it more likely that unqualified readers will 'break into' some of the advanced, restricted Tibetan texts. With Tibetan culture in danger, will its textual resources be pillaged by unqualified, spiritual treasure-seekers, and will TEI become a tool for such digital plunder?

A second issue raised by traditional Tibetan texts concerns the ritualistic manner of their reading. Because many of these texts are practice texts, which lay out the particular steps that a meditator would follow in his or her daily meditation practice, the traditional method for reading these practice texts is a training style. It is unlike the reading style used for most western texts, because it requires precision of pronouncing each word (at least in one's head), accurate counting of the chants requiring repetition, and accompanying visualizations of a detailed nature. Within the Tibetan tradition, many manuscripts are not meant to be read through once from start to finish, but are meant to be scripts for daily meditation practices, involving ritualized bodily and mental gymnastics.

To encode a Tibetan practice text as though it is to be read straight through, in the way that most western texts are read, would neglect its most important function, which is to train the reader in meditation. The solution is not to trivialize Tibetan

meditation practices by encoding loops for repeated chants or by encoding inserted graphical images for visualizations, for this would not respect accomplishments expected of the reader/practitioner. It is the reader/practitioners who must contribute the repetition of chants or the visualization of a meditation deity to their reading of the practice text, and a software program that supplied these would sabotage the training that the text instantiates.

Finally, a third issue concerns whether the TEI codes can distinguish between different audiences for the encoded manuscript: can TEI tags be designed to discriminate between a reader who is practicing meditation with the text and a reader who is reading the text for standard western research purposes? A reader who has no interest in meditating or in Tibetan beliefs may read the text as a work of literature, philosophy or history; this reader would benefit from TEI codes that mark the structure and bibliographic details of the text. The TEI tags would allow this first kind of reader to read the text 'western style'. But a second kind of reader, who is actively engaged in using the text in the traditional Tibetan way as a meditation practice text, may be looking for TEI encoding that will map out the most important clues for how to succeed at the meditation practice. Is this a reasonable project for TEI encoding?

The TEI's application to multilingual, multicultural projects is not a simple, uniform expansion but is a hermeneutic exercise in acknowledging and facing its horizons. Textual practices that are unproblematic within one's own culture may be problematic in another culture, because the farther one travels from one's own culture, one finds that texts function in radically different ways. These multicultural horizons need not be limitations on the TEI, but they should be kept within the broad, panoramic view of what the TEI is trying to accomplish.

## Bibliography

Anderson, Deborah. "Unicode and Historic Scripts." *Ariadne* 37 (2003). Accessed 2005-03-03. <<http://www.ariadne.ac.uk/issue37/anderson/>>

Bia, Alejandro, and Manuel Sanchez-Quero. *The Future of Markup is Multilingual*. . Accessed 2005-03-03. <<http://www.hum.gu.se/allcach2004/AP/html/prop119.html>>

Dreyfus, Georges B.J. *The Sound of Two Hands Clapping: The Education of a Tibetan Buddhist Monk*. Berkeley, California: University of California Press, 2003.

Hayles, N. Katherine. *How we became posthuman*. Chicago: University of Chicago Press, 1999.

Thurman, Robert A.F. *Essential Tibetan Buddhism*. San Francisco: HarperCollins, 1995.

## Automatic Discovery of NLP Resources on the Web

---

*Viktor Pekar* (*v.pekar@wlv.ac.uk*)

*CLG, University of Wolverhampton*

*Richard Evans* (*r.j.evans@wlv.ac.uk*)

*CLG, University of Wolverhampton*

---

The World-Wide Web has become a popular vehicle for disseminating and obtaining research and educational resources. In Natural Language Processing (NLP), for example, the specialist community has accumulated a large amount of NLP resources which it has developed over years, including software (part-of-speech taggers, parsers, various corpus analysis tools) and data (evaluation corpora and datasets, frequency lists, glossaries, gazetteers). Most of these resources are freely available, which helps other researchers to save effort in developing them and allows for direct comparisons with previous work. However, finding these resources on the web is not straightforward. Traditional keyword search is often too costly in terms of time and effort. One can look up collections of web links on the topic, such as the ACL/NLP Universe. Unfortunately, as they are manually compiled and maintained, these collections quickly fall out of date and have limited coverage.

The purpose of our work (ESRC grant RES-000-23-0010) is to design a system which will mine the Internet for pages on NLP resources, extract relevant facts from them and produce a publicly accessible database with this information. The task needs to be addressed by a combination of technologies from the areas of language processing and information retrieval. While these technologies have long been investigated in isolation, recent research started to focus on integrating them in complex information systems to be deployed for real-world tasks (e.g., Petasis 2003). The main contribution of this paper is a range of solutions we develop for effective and efficient integration of the technologies. We here present the design of the system under development, focusing on interoperability of its components.

### 1. Domain crawler

Initial inspection of existing web resources showed there are two most useful sources for their discovery, manually prepared collections of links and email announcements on specialized lists. To find these pages, a web crawler was implemented

which (1) seeks out large collections of URLs of interest on the web and (2) selectively downloads pages mentioned in them that may be relevant to the domain. This first step produces an intermediate corpus of potentially useful documents.

## 2. Format normalization

All the retrieved documents come from extremely diverse sources. For the purpose of further processing, their formatting has to be made uniform. To achieve this, the following pre-processing was performed:

- (a) Since email messages typically appear as plain text, the text layout in them (headings, subheadings, itemized lists, emphasized text, etc) was automatically recognized and corresponding HTML tags were inserted.
- (b) Character encoding was normalised to ensure that it is uniform throughout the corpus.
- (c) HTML errors were detected and corrected using the HTML Tidy tool.

## 3. Text filtering

To identify relevant documents among those downloaded, the text filtering component implements an interface to *Rainbow*, a freely available text categorization toolkit. It classifies all the downloaded documents into two categories: relevant and irrelevant ones. As training data, the system uses around 100 pages describing various NLP resources and around 900 irrelevant pages, randomly picked from the output of the crawler at a pilot run and manually classified. After testing a range of parameters in the standard categorization procedure (learning algorithm, feature selection parameters, various tokenization options, etc.), the most optimal categorization scheme was determined (F-measure=0.97).

## 4. Term extraction and gazetteer acquisition

The term extraction component is responsible for the acquisition of the most important terms and named entities (person and organization names, dates, names of software and data resources) in the domain. The list of terms it produces is used to improve the tokenization necessary for text categorization and to construct domain gazetteers. This component looks for the most frequent words and word sequences in the domain corpus, and applies a range of pattern matching rules to filter out errors and recognize particular semantic types of terms and named entities as well as their abbreviations. The term lists are later revised by an expert to ensure their quality.

## 5. Language identification

One characteristic of the downloaded documents is that many of them contain text in two or more languages (e.g., the description of resources for languages other English). The purpose of the language identification component is to remove non-English paragraphs from a document. It uses a character-based 3-gram model of the language identity of the text, which allows the correct recognition of the language of small text snippets. The component puts special tags around the paragraphs identified as non-English so as to exclude them from further processing.

## 6. Named entity recognition

The Named Entity (NE) recognition component identifies various semantic types of proper nouns, a step preceding information extraction. NE recognition is carried out by a combination of methods, which include:

- (a) Common NEs (person names, locations, and dates) are recognized using the GATE system (Cunningham et al. 2002). Its output is customized to the application domain with the help of a set of post-processing rules (e.g., geographical NEs such as oceans and mountain ranges are irrelevant for the domain, so NE tags are removed from such words).
- (b) NEs specific to the domain are further recognized using (1) gazetteers automatically acquired from the domain corpus, (2) a set of transducers (rules for semantic annotation of text).
- (c) A newly proposed method, which exploits text layout to learn NER rules from already annotated NEs, was applied to improve the coverage of the NER component.

The NER component is run before language identification and text categorization. Removing all proper nouns from text before recognizing its language helps to reduce errors caused by the presence of foreign names in it. Text categorization profits from substituting unique proper nouns by their semantic category labels, whereby all semantically similar NEs are mapped to the same feature.

## 7. Coreference resolution

Given that in a particular document, there may be different ways to refer to NEs (e.g., full names, abbreviations, definite descriptions), it is important to link such variants into coreference chains. The component creates coreference chains for the NE types recognized by the NER component by applying a set of pre-defined rules firing orthographic cues (Bontcheva et al. 2002).



## 8. Information extraction

The information extraction (IE) task consists of identifying relations between recognized NEs, which are later used to fill complex templates (Table 1 describes the template to be filled along with the example fillers). Previous research has developed a range of IE methods for tasks that require filling one template per document (e.g., Kushmerick et al. 1997, Freitag 1998). Here, we opt for an IE method similar to the one proposed by (De Sitter & Daelemans 2003). This method learns two distinct machine learning classifiers from an annotated corpus: one operating on the level of sentences and one on the level of words. First, the sentence-level classifier scans the document for sentences that potentially contain template fillers. After that, the word-level classifier attempts to precisely pinpoint the filler instance in the relevant sentences by looking at the local context of each of its words. Thereby, the context of a word occurrence is represented through features corresponding actual text tokens, all semantic and HTML tags appended on them, and part-of-speech tags.

Field	Example filler
Name	CLAWS part-of-speech tagger
Area	PoS tagging
Creator	UCREL
Licence	Commercial, in-house service
Platform	UNIX
Prog_language	
Req_applications	
Nat_language	English
URL	< <a href="http://www.comp.lancs.ac.uk/ucrel/claws/">http://www.comp.lancs.ac.uk/ucrel/claws/</a> >

Table 1

One difficulty with applying IE to the domain corpus is that very often a particular document does not contain information about all the fields of the template (e.g., the field for natural language remains unfilled for language-independent tools like annotation software). They should be distinguished from documents which do not fill the template fields because they have been erroneously classified as relevant at the text filtering stage. To draw this distinction, a special verification step is applied. It is based on estimating the probability of every field being filled for relevant and irrelevant documents in the gold standard corpus and comparing the corresponding probability vectors with a similar vector prepared for each newly processed document.

A prototype incorporating these components has been implemented and in subsequent work, we will continue to

perform user-focused tuning of this tool to enhance it with a view to deployment in the research domain. Although the initial stages of the BiRD project have addressed IE in the field of computational linguistics, it will be interesting to evaluate the system in application to new domains.

## Bibliography

- Bontcheva, K., M. Dimitrov, D. Maynard, V. Tablan, and H. Cunningham. "Shallow Methods for Named Entity Coreference Resolution." *Chaînes de références et résolveurs d'anaphores, workshop TALN'2002* (2002).
- Cunningham, H., D. Maynard, K. Bontcheva, and V. Tablan. "GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications." *Proceedings of ACL'02* (2002).
- De Sitter, A., and W. Daelmans. "Information extraction via double classification." *Proceedings of the ECML/PKDD'03 Workshop on Adaptive Text Extraction and Mining (ATEM'03)* (2003).
- Freitag, D. "Information Extraction From HTML: Application of a General Learning Approach." *Proceedings of AAAI'98* (1998).
- Kushmerick, N., D. Weld, and R. Doorenbos. "Wrapper Induction for Information Extraction." *Proceedings of IJCAI'97* (1997): 729–737.
- Petasis, G., V. Karkaletsis, G. Paliouras, I. Androutsopoulos, and C.D. Spyropoulos. "Ellogon: A new text engineering platform." *Proceedings of LREC-02* (2002).

## *The Robert Graves Diary* (1935-39): a TEI Application using an XML Database (eXist)

---

**Chris Petter** ([cpetter@uvic.ca](mailto:cpetter@uvic.ca))

University of Victoria

**Elizabeth Grove-White** ([grovewhi@uvic.ca](mailto:grovewhi@uvic.ca))

University of Victoria

**Linda Roberts** ([l-roberts@shaw.ca](mailto:l-roberts@shaw.ca))

University of Victoria

**Spencer Rose** ([srose@uvic.ca](mailto:srose@uvic.ca))

University of Western Ontario

**Jessica Posgate** ([jposgate@uvic.ca](mailto:jposgate@uvic.ca))

University of Victoria

**Jillian Shoichet** ([shoichet@uvic.ca](mailto:shoichet@uvic.ca))

University of Victoria

---

**O**ur paper will cover both academic and technical development of the Graves Diary Project (1935-39). The prototype can be found at <http://www.tapor.uvic.ca:8080/cocoon/graves/> .

### Academic Development

**R**obert Graves' 1935-39 diary is part of the prized Graves collection in the UVic Libraries' Special Collection. The diary has been used extensively by biographers and scholars, but it has remained inaccessible to a wider readership until recently. The Robert Graves Trust in Oxford owns the copyright to the diary and in 2001 they agreed to allow the University of Victoria Libraries to publish the diary as both an electronic and a print edition. Beryl Graves, Robert's widow, transcribed the diary into text and a copy of this text was deposited with the University of Victoria. In addition the Trust encouraged Chris Petter to scan an annotated version of the transcript, prepared by Karl Goldschmidt, Graves' long time secretary. William Graves, Robert's eldest son by Beryl, offered to contribute notes that he kept on the Deyá portions of the diary.

Robert Graves (1895-1985) is a major twentieth-century English poet, novelist and essayist. After surviving the First World War and subsequent shell shock, he married, studied at Oxford and

began to publish poetry. In 1926 he met Laura Riding, the American poet whose work he had admired from afar. She became an enormous influence on him, and on his writing, and their intense working relationship lasted for over ten years. They founded the Seizin Press together, and in 1929 they moved to Deyá, Majorca. The novels that made Graves famous — *Goodbye to All That*, *I Claudius* and *Claudius the God* — were written in this period. The diary is an important document illuminating their life together and that of the little coterie of writers and artists they gathered around them.

The diary's permanent project team consists of Elizabeth Grove-White (English Department) who is responsible for the introductory material; Chris Petter (Library), project manager and Linda Roberts M.A., who is responsible for encoding, abstracting and annotation. Spencer Rose, and later Martin Holmes of UVic's Humanities Computing and Media Centre have developed the interfaces. Elizabeth was successful in landing a two year SSHRC project grant for the diary for 2004-2006. Dr. Patrick Quinn kindly contributed monthly abstracts for 1935 and 1936 diary entries.

### TEI Development

Work began in 2002 when Chris Petter was granted a study leave from the Library. The manuscript was digitized and an index created which links the file title to the date. Chris traveled to the University of New Brunswick Text Centre and then to Oxford. At UNB Chris was able to restructure the text files into day entries within month divisions. In Oxford, Sebastian Rahtz advised on using the TEI.corpus.dtd and an XML database to present the diary. The reason for this advice was because of the structural difficulties of the diary with its 115 enclosures and numerous letter logs. Chris also set up databases which could store information on the names, places and titles mentioned in the diary. These included the annotations of Karl Goldschmidt and the notes contributed by William Graves.

### Markup

A guiding principle of the Graves diary markup procedure is to approximate the original document as closely as possible, so that the character of Graves' 'diary style' is preserved along with its content. Fortunately, XML (Extensible Markup Language), with its capacity to convey emendations such as deletions (crossed out) and supralinear additions allows us to produce an authentic version which reflects to some extent the immediacy of the diary mss. It has been necessary to work constantly with the mss in order to identify and adjust any changes made in the transcript which diverge from the copy text, including paragraphing, spelling and punctuation. Any exceptions will be accounted for in the editorial notes. The

markup process allows us to include annotations for names, places, titles, foreign words, and emendations, as well as notes and editorial comments.

## Technical Implementation: Web Interface Development

Work on the web interface began in the fall of 2002, and has since become a platform for testing client-side xml processing in the rendering of XHTML documents using XSL stylesheets. Spencer Rose's contribution to this project, through the Humanities Computing and Media Centre, has involved transforming updated TEI-conformant XML documents into a simple and transparent web interface that is intuitive and useful for researchers.

The first prototype developed by Spencer Rose in 2003 made use of client-side XML processing, but was expanded to accommodate more complex XML markup. These xml processing capabilities became available with advanced web browser software. Some desirable features of client-side XML processing included the offloading of processing from the server to the client, and the direct access of XML files for customizable display. However, unlike server-side XML-to-XHTML transformations, client-side processing depends on the compatibility of the web browser to parse and render using XSL stylesheets, which, until recently, had been an unstable feature of standards-compliant browsers.

## Web Prototype

The interface design involved two phases. The first phase was to build a static web display that allowed for easy browsing of the diary text. The second phase would allow users to perform complex search querying of the XML documents. For this prototype, the interface design involved a number of separate components. Of these components, some might be considered common to most web development projects such as using CSS and javascript to web-enable the site; others required special work. The static components of the site design included the general web design, XHTML layout and styling using XSL rendering and Cascading Stylesheets. The dynamic components involve using javascript for client-side interactivity. These components are brought together to form a document that is web-enabled.

## XSL Templates and XHTML

*The Graves Diary* xml documents strictly conform to the *Text Encoding Initiative* guidelines and therefore use standard tags

and attributes that describe typographic and analytic structures of the text. Attention to detail in the XML markup was reflected in the detail of the XSL-Transformed representation such that the diary's wide range of styling features — all encoded using TEI elements — were reproduced in the transformed XHTML document. As well, the interactive features of the interface — including image scan and spot-of-reference pop-ups, as well as other dynamic display elements — were developed using client-side javascript.

## XML Indexing System

*The Graves Diary* contains numerous enclosures — letters, poems, photographs — clippings that are components of the transcription. As with each diary entry, each enclosure has a separate digital scan that is indexed in XML documents. As well, each entry and enclosure also contains numerous biographical, geographical and bibliographical references that link to an external reference database. Because of this complex cross-indexing of media, reference information and enclosures, an important design issue was deciding on a suitable indexing system that linked these components in a coherent display.

One of the projects greatest innovations was the creation by Spencer Rose of two modular XML index files: one file cross-indexes the collection of digital image scans of the diary (including enclosures) with the main diary files; another XML file lists reference entries identified with reference locations in the diary text. Both of these external XML files originated in different file formats and needed to be transformed into XML documents. These XML files could then be included with the diary markup in the XSL templates, and as well made the creation of image and file index displays straightforward. Finally, XML pointer files for the diary entries were also used to isolate the XSL references in the document header from the actual document. This has the benefit of removing the diary files from a specific stylesheet reference.

## eXist XML Database Late Breaking Development

The present phase of this interface project is to make the transcribed *Graves Diary* documents searchable online. For this, the implementation of the Open Source native XML database system eXist (<http://exist-db.org/>) has shown a promising start — with at least the proof of concept being established in a working prototype.

The eXist search engine makes use of an extended XPath query language called XQuery to search elements in a document. XPath is an established document syntax that is integral to XSL in that it defines the elements of XML documents for stylesheet

transformations. eXist's enhanced querying includes basic XPath expressions to search through the nodal structure of the XML document, but it is also capable of keyword searches on XML elements and attributes, as well as queries on the proximity of search terms and regular expressions. Analyses of nodal relationships (e.g. parent-child relations between elements) are also possible with eXist. One feature of the eXist search engine is that, for a wide range of XPath expressions, it uses stored index files that reference the structure of the XML document nodes. Information can then be retrieved without accessing the collections documents directly. This improves the speed and efficiency of information retrieval.

*The Graves Diary* eXist database is still in the process of development, with the rendering and placement of enclosures (some multi-page) alongside their digital images proving to be a challenge for Martin Holmes (Humanities Computing). In the meantime, the markup of the diary text and the creation of abstracts for each month by graduate students continues under the supervision of Linda Roberts. The project is scheduled for completion by July 2006.

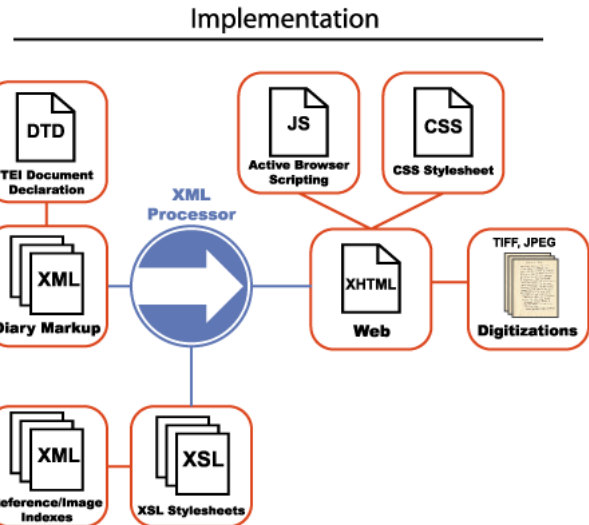


Figure 2

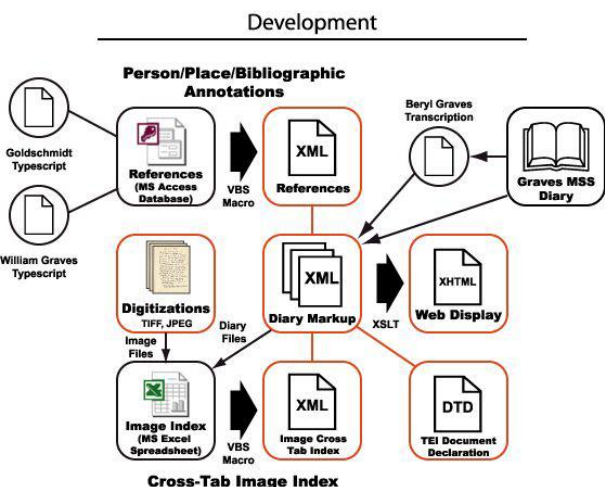


Figure 1

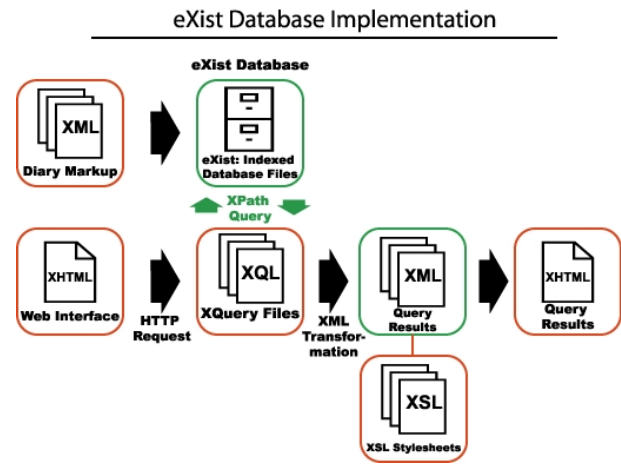


Figure 3

## Bibliography

Meier, Wolfgang. *eXist: An Open Source Native XML Database*. Accessed 2002-11. <<http://exist-db.org/webdb.pdf>>

# An Encoding Model for Librettos: the Opera Liber DTD

Elena Pierazzo ([pierazzo@ital.unipi.it](mailto:pierazzo@ital.unipi.it))

University of Pisa

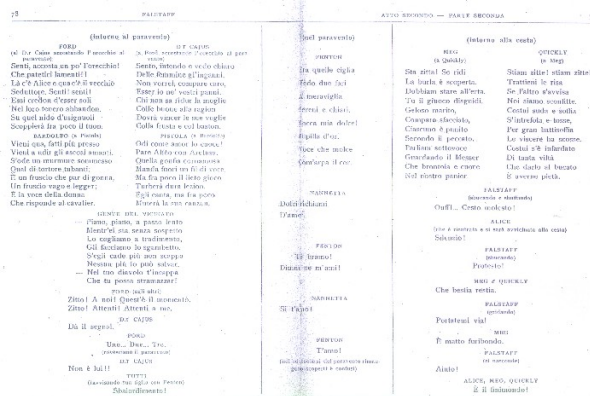
Opera librettos are a very peculiar literary genre. Often considered an ancillary part of the opera, merely the plot through which the music can express its power and its beauty or the pretext for singers to show their capabilities and the potential of their voices, the libretto is a little studied aspect of the literature.

A considerable number of web sites are currently presenting collections of librettos in several formats (doc, pdf, html, gif, txt). However, they generally do not cite their source or even which version of the text they are based upon; furthermore, in most cases they do not respect editorial traditions of the libretto.

Librettos have some peculiar structural characteristics: they can be considered a subcategory of drama texts, but they distinguish themselves from the non-musical drama texts mainly in two ways:

- the presence of *Concertato* sections
- the extreme fragmentation of the versification.

The *Concertato* is a musical term that is passed on in the librettos tradition to mean normally a scene or part of a scene performed simultaneously by different characters, each singing different texts, including several cues and stage directions. The number of simultaneous sequences can range from a minimum of two, to a maximum of seven/eight, as in the following example taken from Falstaff (music by Giuseppe Verdi and libretto by Arrigo Boito).



Pages 77-78 of the Falstaff libretto

In the libretto the versification is extremely fragmentary: as for the drama, verses and stanzas are usually split according to the different cues; furthermore, *fin de siècle* librettos admit different metres that can change at any moment, even within a cue.

An important point is that usually the libretto that is printed and distributed to the public can be markedly different from the one that is sung on the stage. In the score, verses and words are adapted to the musical progression and for that reason they can be stretched, repeated, modified, cut and added. The libretto is often conceived as a support for the spectator; in the libretto, indeed, portions of text suppressed in the score, stage directions, comments, notes that have no match with the performed opera, can help the spectator to follow the plot. All these peculiarities need to be seriously taken into consideration before starting any encoding. Firstly, this is because the librettos' printing tradition has fixed some conventions to represent the different characteristics. Second, the public to which a digital collection of librettos is addressed will expect its habits to be taken into account.

In the last two years a research project named "L'Opera prima dell'Opera" (*The text/literary source before the staging of Opera*) has carried out the creation of a digital library of librettos called *Opera Liber*, freely available on the Net (currently at <http://80.19.150.245/operaliber/> but will be soon transferred to <http://www.operaliber.it>). *Opera Liber* is a portal for the study and the documentation of the Italian librettos for the period 1870 - 1920, including works of the main Italian composers such as Verdi, Puccini, Leoncavallo, Mascagni, Ponchielli and many others. The main resource of the web site is represented by the collection of texts, available both for reading and for linguistic querying. The texts have been encoded in XML TEI format and are managed and queried using the native XML database eXist. The *Opera Liber DTD* is a customization of the TEI DTD P4, fully documented on the web site, and it is constituted by a mixed base set (verse and drama) and additional tag sets such as figure, transcriptions of primary sources, linking, and

names and dates. Some customizations of the DTD have been made, following the prescription of the Chapter 29 of the *Guidelines for Text Encoding and Interchange* (Sperberg-McQueen & Burnard).

In creating the encoding model the main problem was to find a correct encoding for *Concertatos*. The Concertato can surely be considered a sort of structural division, even if not at the same level of usual structural divisions (such as acts and scenes, encoded by the TEI <div> elements). A milestone approach that was also considered would miss the consistency of the *Concertato* sequences. For that reason we decided to create a new element <sequences> that will include a number of <sequence> elements, according to the number of columns in the printed form.

We decided also to consider as source physical copies of librettos, and not so called *ideal copies* and that because it is often difficult to determine the belonging of a copy to a particular edition or issue. Publishers, in fact, usually printed a large amount of librettos, storing unsold copies, just changing the front matters to fit the libretto they have in their repositories to a particular *mise en scène*, sometimes mixing copies from different printings. Furthermore, some of the copies we have considered for encoding contain manuscript notes or dedications. For all these reasons we settled on recording the provenance of the encoded copies, creating the element <copyStmnt> (and the child elements <settlement> and <repository>) inside the <sourceDesc> element.

Another problem was given by the encoding of the name of characters. One of the peculiar characteristics of 'classic' drama (from Greek tradition till the beginning of the twentieth century) is the so called *agnition*, i.e. a character that is believed to be a certain person, is recognised to be someone else, often determining the unravelling of the plot. That means that a character may have two names, the supposed and the real one, but it is not two persons and that's why the possibility of using two nested <persName> was refused. We decided, instead, to create a new attribute (called *alias*) for the <persName> element to record supposed or virtual names, reserving the *reg* attribute for real names.

The different kinds of metre have not been semantically encoded because in many case the difficulty of understanding the rationale invites caution. For that reason the metric divisions have been marked only in really obvious cases, while in other cases only the physical appearance of the verse has been encoded by the usage of the <hi> element. In such way we have recorded the presence of:

- particular indentations, normally, but not always, representing a changing of metre;
- inverted commas, normally representing not sung verses;

- dashes normally representing the alternation of voices in choral singing.

A number of minor implementations of the DTD have been operated, fully documented in the web site.

The web site *Opera Liber* collects the experience of two years' work in the field of encoding opera librettos and offers itself as a point of reference for analogous experiences.

## Bibliography

*Opera Liber*. <<http://80.19.150.245/operaliber/>>

Sperberg-McQueen, C.M., and L. Burnard, eds. *TEI P4: Guidelines for Electronic Text Encoding and Interchange*. Text Encoding Initiative Consortium, 2002. Accessed 2004-10-09. <<http://www.tei-c.org/P4X/>>

Verdi, Giuseppe, and Arrigo Boito. *Falstaff*. New York: G. Schirmer, 1963.

## SVG Visualization of TEI Texts

Wendell Piez ([wapiez@mulberrytech.com](mailto:wapiez@mulberrytech.com))  
 Mulberry Technologies, Inc.

### SVG Visualization of TEI Texts

One of the more interesting benefits of XML technology for text processing has been the 'network effects' we get from using different XML technologies together. For example, XSLT proves to be suitable for a great range of tasks beyond simply the routine formatting of texts for display in a browser or on the page (the job for which it was designed): the investment we make in learning XSLT to generate reading versions of our XML texts also pays off many times over in enabling us to perform other kinds of tasks such as extra-schema validation, heuristic analytics of the markup or the text itself, and even (up to a point) querying. Likewise, it proves easy to produce a wide range of different kinds of output to represent the results of these operations. An XML application such as SVG proves to be a straightforward target for a transformation from XML data. The resulting SVG graphics can be anything. For example, graphs and bar charts of information captured in numerical data sets and represented in XML are easy to create using XSLT/SVG. But so are more arcane kinds of depictions of source datasets or their features, including using SVG as a display format for 'maps' of a document's structure.

This basic architecture, XML + XSLT -> SVG, has been demonstrated repeatedly in both the commercial and academic sectors in recent years (see Bibliography; several applications by the author demonstrating the use of XSLT to create SVG graphical depictions of various kinds are included (Piez 2000, 2002, 2003a, 2003b). There is nothing particularly innovative at this point (late 2004) about this inexpensive and powerful method of creating graphics. What has been explored perhaps less deeply is what can be done with stylesheets generating graphical depictions of specifically *literary* works, leveraging descriptive tagging of the 'pure' kind (that is, tagging that has been designed to reflect documents' logical organization, without any particular renditions in mind). Not only are the structures and features of such works of intrinsic interest to students of literature; they can also serve as a diverse and heterogeneous testbed for prototyping techniques of rendition and visualization that could be used on other sources or indeed, on other kinds of XML data. These techniques would be widely applicable both to works of narrative or discursive prose and to more highly structured literary texts such as verse and drama.

Earlier demonstrations of this approach make it clear that we are now, with the maturation of XML technologies and the increasing support of SVG in readily available tools (the *Mozilla* development team has lately been implementing SVG for their browser, and Adobe continues work on the technology as well), in a position where we can perform these operations on a larger scale. One of the features of the architecture is that a family of documents marked up consistently with the same tag set (say, TEI) should be processable with the same stylesheet. The marginal effort required to create a graphic depiction of a new text, consequently, is negligible when that text's tagging conforms to a known and supported usage pattern (preferably valid to a known DTD). In theory, it should be possible to generate an entire library of graphics to represent a library of texts, all with a single stylesheet.

The poster I am proposing for ACH/ALLC 2005 will present the results of a set of experiments testing these ideas, applying stylesheets (both extant and new) on a variety of texts from the *Women Writer's Project* at Brown University (with their kind permission and collaboration). This will have the twofold purpose of exploring what kinds of visual representation of these structures are most revealing, as well as testing to what extent single stylesheets or small families of stylesheets can be used across a document repository, to draw interesting and revealing comparisons among texts. (It is quite possible that per-document "tuning" of the presentation logic will be necessary, through a customization layer, for best results; but until we have tried the technique on a range of texts, we will not know the extent to which stylesheet reuse is practical. This extent may also vary between different stylesheets used to create different sorts of graphics.)

Stylesheets developed for this poster will also be contributed to the *WVO (Women Writers Online)* project, and made available to the wider TEI community.

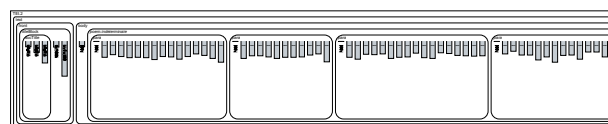


Figure 1: Aphra Behn, "A Pindaric Poem to the Reverend Doctor Burnet" (1689). An example of a free verse form.

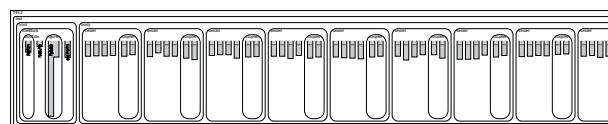


Figure 2: Catherine Clive, "The Case of Mrs. Clive" (1744). An example of a work in prose.

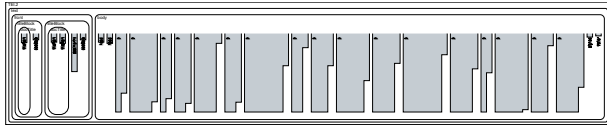


Figure 3: Mary Sidney, Countess of Pembroke. "The Doleful Lay of the Fair Clorinda" (1595). An example showing a regular verse form (sestets containing couplets).

---

## Regelbasierte Suche in Textdatenbanken mit Nichtstandardisierter Rechtschreibung (Rule Based Search in Text Databases with Non-Standard Orthography)

---

### Bibliography

- Birnbaum, David J. "Analyzing and visualizing the structure of medieval encyclopedic works with XML-related technologies." Paper delivered at the Extreme Markup Languages 2003, Montreal. August 2003.
- Cagle, Kurt. *SVG Programming: The Graphical Web*. Berkeley, CA: Apress, 2002.
- Eisenberg, J. David. *SVG Essentials*. Sebastopol, CA, USA: O'Reilly, 2002.
- Mangano, Sal. *The XSLT Cookbook*. Sebastopol, CA, USA: O'Reilly, 2002.
- Mansfield, Philip A., and Darryl W. Fuller. "Graphical Stylesheets: Using XSLT to Generate SVG." Presented at XML 2001. 2001. On line at <http://www.idealliance.org/papers/xml2001/papers/html/05-05-02.html>
- Piez, Wendell. *The Sonneteer: A demonstration of structured form*. Accessed 2005-04-13. <http://sonneteer.xmlshoestring.com>.
- Piez, Wendell. "SVG By Way of XSLT." Tutorial delivered at Extreme Markup Languages 2001, Montreal. August 2001.
- Piez, Wendell. "Visualizing XML document structure using XSLT and SVG." *interChange, the journal of ISUG (the International SGML Users' Group)* (December 2003): n. pag. On line at <http://www.xmlshoestring.com/xml499/visualizingxml>
- Piez, Wendell. "XSL: Characteristics, Status and Potentials for the Humanities." Presented at ALLC/ACH 2000, Glasgow. July 2000. On line at <http://www.idealliance.org/papers/xml2001/papers/html/05-05-02.html>
- Tennison, Jeni. *Beginning XSLT*. Birmingham, UK: Wrox Press, 2002.

*Thomas Pilz* ([pilz@informatik.uni-duisburg.de](mailto:pilz@informatik.uni-duisburg.de))  
*University of Duisburg-Essen*

*Prof. Dr. Wolfram Luther*  
([luther@informatik.uni-duisburg.de](mailto:luther@informatik.uni-duisburg.de))  
*University of Duisburg-Essen*

*Prof. Dr. phil. Ulrich Ammon*  
([ammon@uni-duisburg.de](mailto:ammon@uni-duisburg.de))  
*University of Duisburg-Essen*

*Prof. Dr.-Ing. Norbert Fuhr*  
([fuhr@uni-duisburg.de](mailto:fuhr@uni-duisburg.de))  
*University of Duisburg-Essen*

---

In this paper we describe our interdisciplinary project in support of the conservation of cultural heritage, especially for the German reception of Nietzsche. We present a rule based fuzzy search-engine which allows retrieval of text data independently of its orthographical realization. The rules used are derived from statistical analyses, historical works, linguistic principles and professional administration. Our web based tool aims at experts as well as interested amateurs. In addition to its present features, further functions are currently worked out that include automatic rule derivation and a finer result classification via a generalized Levenshtein similarity measure.

Dans cette note, nous décrivons notre projet interdisciplinaire concernant l'édition électronique de la réception allemande des idées et de l'œuvre de Nietzsche. Nous avons centré un point d'intérêt sur la création d'un moteur de recherche accessible dans le Web. Celui-ci permet la recherche floue, phonétique et par troncation nécessaire au traitement de la plupart des textes numérisés écrits avant la réforme de l'orthographe en Allemagne en 1901/02. Le logiciel est basée sur un algorithme qui déduit pour chaque nom, verbe et adjectif toute orthographe possible selon un système de principes ou règles linguistiques cités dans la littérature historique ou dérivés en collaboration avec des



spécialistes. Plusieurs autres options sont prévues y compris une dérivation automatique des règles et une classification des résultats basée sur une mesure de similarité généralisée de Levenshtein.

Im Kontext eines Digitalisierungsprojekts zur Nietzsche-Rezeption aus den Jahren 1865-1945, das seit mehreren Jahren in Duisburg in Zusammenarbeit mit dem Nietzsche-Kolleg in Weimar verfolgt wird [BM02, BM03], beschäftigt sich das von der Deutschen Forschungsgemeinschaft geförderte RNSNR-Projekt mit der Erforschung und Entwicklung eines linguistischen Regelsystems, einer Transformationsmethodik und zeitabhängiger Filter zur Unterstützung der Suche in Textdokumenten in nichtstandardisierter Rechtschreibung.

Es wurde bereits eine Java-basierte Suchmaschine erstellt, welche es durch einen neu entwickelten phonetischen Regelsatz ermöglicht, auf Texten, die mehrere hundert Jahre vor der Rechtschreibvereinheitlichung des Jahres 1901 verfasst wurden, eine Suche mittels orthographisch genormter Schlagwörter durchzuführen (vgl. Abbildung 1) [P03]. Durch Einführung eines Abstands begriffs [ZD96] sind verschiedene Stufen der Ähnlichkeit realisiert. Außerdem erlaubt der Algorithmus durch einen zusätzlichen speziellen Regelsatz auch die Suche nach Wörtern, welche durch OCR-Software fehlerhaft erkannt wurden. Die Suchmaschine ist in das online-verfügbare HTML-basierte Nietzsche-Archiv integriert.

Mit der regelbasierten Suche verfolgen wir einen anderen Ansatz als viele große Wörterbuchprojekte. Indem nicht mit statischen Wortlisten gearbeitet wird, erhoffen wir uns eine höhere Trefferquote, besonders bei Texten mit stark variierender Schreibung. Zusätzlich wird der Arbeitsaufwand durch manuelle Eintragung von Wort-Relationen vermieden. Andererseits hoffen wir durch Grundlagenforschung, besonders in den Bereichen der Phonem-Graphem-Struktur des Deutschen, der unscharfen Suche und der Ähnlichkeitsmetriken, einen Wortabstands begriff zu definieren, der sowohl eine größtmögliche Differenzierung unterschiedlicher als auch Zusammenfassung äquivalenter Wörter ermöglicht [A98].

Neben der Anwendung als Suchmaschine sind auch Einsatzpunkte im Vergleich oder der temporal-lokalen Einordnung von Texten denkbar. Zentraler Betrachtungszeitraum sind für uns die Jahre 1700-1900. Eine spätere Ausweitung des Regelsatzes auch auf frühere Zeitabschnitte ist durchaus möglich.

Im Einzelnen verfolgt das Projekt die folgenden Ziele:

- Entwicklung von Zeit- und Ortsfiltern für phonetische Regeln, Revision der Regeln aus der Textbasis und aus statistischen Analysen, Nutzung eines Kontrollwörterbuchs gegen Homonymhäufung.

- Entwicklung eines neuen adäquaten Abstands begriffs auf der Basis eines modifizierten graphematischen und phonetischen Levenshtein-Ähnlichkeitsmaßes, Berücksichtigung typischer Erfassungsfehler, Entwicklung von Unschärfeskalen.
- Integration der Suchmaschinen in das Nietzsche-Projekt und in andere Systeme wie das Deutsche Rechtswörterbuch oder das Projekt Deutsch Diachron Digital, Entwicklung von Regelsätzen und Erweiterung der Suchmaschine auf (früh-)neuhochdeutsche Archive.

Hauptsächliche Arbeitspunkte sind zur Zeit

- eine Verbesserung des Tools in Hinsicht auf Effizienz
- Grundlagenforschung zum regelbasierten Ansatz
- Untersuchungen zur Levenshtein-Distanz
- ein Vergleich regelbasierter mit Wörterbuch-basierter Suche
- eine Einbringung der Suchmaschine in andere Projekte.

Mittelfristig wird eine verbesserte Realisierung mit einem Java-Frontend, einem Web-Server und einer modernen XML-basierten Archivlösung angestrebt [FGG02], die auch in vergleichbaren Digitalisierungsprojekten Anwendung finden kann.

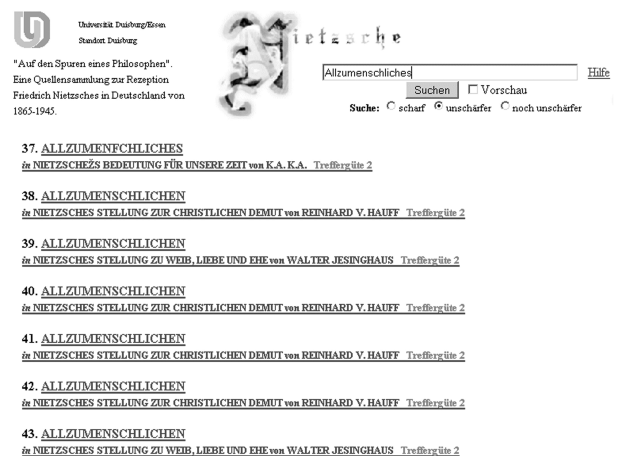


Abbildung 1: Webinterface zur unscharfen Suche

## Arbeitsmethodik

Die zu Grunde liegenden Texte der bisher bearbeiteten Nietzsche-Rezeption reichen teilweise bis in das Jahr 1865 zurück. Dadurch, dass in diesen Dokumenten Zitate noch deutlich älteren Ursprungs verwendet werden, treten zusätzlich Formen auf, die bereits zu dieser Zeit obsolet waren.

Die Verwendung eines historischen Wörterbuchs wurde zunächst nicht ins Auge gefasst, da ein solches zwar einen Teil der auftretenden Wörter effizient zu indizieren vermag, eine

zumindest annähernd vollständige Erkennung allerdings nicht ermöglichen kann. Dies beruht auf der enormen Divergenz prinzipiell möglicher Schreibungen, die mit dem Alter der Texte immer weiter ansteigt. Es ist ja gerade Merkmal der Texte vor der Rechtschreibvereinheitlichung von 1901/02, dass — zumindest prinzipiell — jede Schreibung möglich war, auch wenn diese nicht immer realisiert wurde.

Allerdings sind diese Allographie durch die Zugehörigkeit von Graphemen zu bestimmten Lautklassen, die sogenannte Graphem-Phonem-Korrespondenz, beschränkt: Jedes Phonem der deutschen Sprache lässt sich nur durch eine endliche Menge von Graphemen realisieren. Unter der Annahme, dass ein heutiges Wort sich in seiner Lautstruktur prinzipiell nur unerheblich von seinem historischen Gegenstück unterscheidet, kann dieses mittels Variation seiner Phonem-Realisierungen rekonstruiert werden. Diese Annahme lässt sich immerhin für das Neuhochdeutsche — und damit für einen Zeitraum von rund 600 Jahren — bestätigen, wenn auch mit einer in der historischen Tiefe abnehmenden Sicherheit.

Durch eine Untersuchung, mittels welcher Grapheme damals eine der wahrscheinlich heute noch gültigen Aussprache ähnliche Lautung zu erreichen war, können Repräsentationsfehler vermieden werden.

Grundlage der schon entwickelten Suchmaschine sind somit phonetische Regeln, welche die in dem betrachteten Zeitraum möglicherweise auftretenden Schreibweisen nachbilden. Durch Kombination weniger Regeln ergibt sich bereits eine erstaunlich realistische Transformationsvariation. Betrachtet man nun noch die weiteren Variationsmöglichkeiten abhängig von der Initial-, Medial- oder Finalstellung der Grapheme, so empfehlen sich keine Arbeitsmethoden, die wie bei vergleichbaren Realisierungen auf zentraler Verwendung eines Wörterbuchs basieren.

Die zu jener Zeit vorkommenden Allographie konnten dafür aus Arbeiten von Rechtschreibreformern wie Adelung und Schottelius indirekt abgeleitet werden. Indem diese forderten, dass Schreibung A vermieden und durch Schreibung B zu ersetzen sei, belegen sie die Existenz der Phonemrealisierung A. Die vor allem im 16. und 17. Jh. aufkommenden Normierungsbestrebungen und die endgültige Übernahme des Hochdeutschen für den gesamten deutschsprachigen Raum haben glücklicherweise eine Vielzahl gut dokumentierter linguistischer Arbeiten hervorgebracht.

Eine eingehende Untersuchung der Aussprache des heutigen Standarddeutschen konnte weitere produktive Regeln hervorbringen. Das grundlegende Schreibprinzip alphabetischer Schriftsysteme "Schreibe, wie du sprichst, und sprich, wie du schreibst!" (phonologisches Prinzip der Rechtschreibung) [K86] hält hierbei damals wie heute die tatsächlich anwendbare Gesamtzahl der Allographie in einem überschaubaren Rahmen.

Durch eine modulare Erweiterung des Regelsatzes um spezielle OCR-Probleme betreffende Produktionen, etwa die Fehlererkennung des <s> in Fraktur durch <f>, wurde der Suchmaschine weitere Funktionalität verliehen.

Auf Basis beliebiger HTML-Dokumente werden in der bereits existierenden Version mittels des verwendeten Regelsatzes zwei Varianten jedes Schlagwortes gebildet 'unschärfer' und 'noch unschärfer'. Die erste verwendet nur einen Teil der Regeln, welche die erwartungsgemäß häufigsten Unterschiede betreffen. Die zweite Variante berücksichtigt alle Produktionen inklusive möglicher OCR-Fehler. Diese Unterteilung wurde aufgrund der resultierenden Homonymhäufung getroffen: Durch die Schlagworttransformationen fallen umso mehr Wörter zusammen, je umfangreichere Regeln verwendet werden. Diesem Phänomen wird mit einem Kontrollwörterbuch zu begegnen sein. Aus den drei Repräsentationen jedes indizierten Wortes sowie aus dessen Position innerhalb des entsprechenden Dokumentes werden mittels eines JAVA-Programms die Tabellen in einer *MySQL*-Datenbank mit Daten versehen.

## Hauptarbeitspunkte

**B**eim Anlegen der Wort-Tabellen für ein Dokument erscheint es naheliegend, neben dem Wort auch eine phonetische Realisierung desselben abzulegen und dann bei der Suche nur wenige ‚nahe‘ Wörter zu berücksichtigen. Allerdings ist es äußerst schwierig, eine korrekte phonetische Realisierung zu einem vorgegebenen Wort zu finden. Es müssen daher Bewertungsmethodiken entwickelt werden, die bestimmen, welche Wörter zu dem Suchwort passen. Dabei können Gesetzmäßigkeiten der Phonetik und Graphematik, aber auch der Wahrnehmungspsychologie wertvolle Hinweise liefern.

Wenn wir bei der Suche Transformationen regelbasiert berechnen sowie relevante Regeln *on the fly* für einen konkreten Text auswählen oder generieren und fakultativ validieren, gelangen wir zu einem schlanken, anpassungsfähigen und letztlich auch tragfähigeren Werkzeug, das die Verwendung von Wörterbüchern (mit mehr oder weniger 'modernen' Einträgen) auf ein Minimum begrenzt und damit die Abhängigkeit von der Vollständigkeit des Wörterbuchs aufbricht. Zusätzlich sollen Regeln zu OCR-Fehlern für eine Suche dazugeschaltet oder aber ganz abgeschaltet werden können. Wichtig für unseren Ansatz ist dabei eine effiziente Verwendung einer weiterentwickelten Levenshtein-Distanzfunktion. Um eine klare Trennung zwischen OCR-Fehlern und Allographen zu ermöglichen, sollen diese anders gewichtet werden als Abweichungen phonetisch naher Schreibweisen. Berücksichtigung bei der Gewichtung finden sollte die Anzahl der angewendeten Regeln, um die Schreibung zu erreichen, wie auch ihre Relevanz. Dabei geht allerdings die

Symmetrie einer Distanzfunktion verloren, da die Ableitungen i.a. nicht umkehrbar sind.

Die Suchmaschine behandelt Anfragen in folgender Art und Weise: In einem Vorverarbeitungsschritt werden Sprache, Zeit und Ort der zu suchenden Dokumente bestimmt, woraus sich die anzuwendenden Regelsätze ergeben. Die Suchterme einschließlich etwaiger Wildcards werden dann durch Anwendung der Regelsätze und unter Berücksichtigung einer parallel zu entwickelnden verallgemeinerten Levenshtein-Ähnlichkeitsmetrik [C&D, 2000] in die internen Suchbedingungen übersetzt; dabei können durch Vorgabe eines Schwellwertes und / oder Ausnutzung eines Kontrollwörterbuchs unwahrscheinliche Varianten ausgeschlossen werden, bevor die eigentliche Suche durchgeführt wird. Die Suchergebnisse werden dann nach absteigender Ähnlichkeit geordnet ausgegeben.

Im Rahmen der nächsten Arbeitsschritte soll die existierende Suchmaschine bezüglich Retrievalqualität und Funktionalität verbessert werden. Um eine hohe Anzahl relevanter Dokumente zu finden, also den Recall zu erhöhen, sollen möglichst alle Flexionsformen und Schreibvarianten eines Suchwortes bei der Suche berücksichtigt werden. Hierzu müssen durch Anwendung der entsprechenden Regelsätze alle Varianten eines Anfragewortes erzeugt werden, mit denen dann im Dokumentenbestand gesucht wird. Da der Regelbestand sehr dynamisch ist, können nicht, wie sonst insbesondere in experimentellen Information-Retrieval-Systemen üblich, die Dokumente schon beim Einfügen in die Datenbasis entsprechend indexiert werden, sondern die Expansion der Suchwörter muss zum Retrievalzeitpunkt erfolgen. Um die Antwortzeiten trotzdem gering zu halten, müssen noch entsprechend effiziente Verfahren implementiert werden.

Durch diese Vorgehensweise können sehr viele Dokumente gefunden werden, wovon aber auch viele nicht relevant sind. Nur durch eine entsprechende Rangordnung der Retrievalantworten kann eine hohe Präzision des Suchergebnisses gewährleistet werden. Liegen zu einem Suchbegriff Wörterbucheinträge vor, so sollen diese Angaben bei der Suche mit berücksichtigt werden. Schreibweisen, die auch im Wörterbuch auftauchen, erhalten dann ein höheres Gewicht als andere Varianten.

Zur Erweiterung der Suchfunktionalität soll die Suchmaschine um gängige Suchoperatoren erweitert werden. Bei der Eingabe von Einzelwörtern soll Trunkierung erlaubt werden, und mehrgliedrige Begriffe sollen mit Hilfe von Kontextoperatoren (Wortabstandssuche) spezifiziert werden können. Mehrere Suchbedingungen sollen wahlweise durch Boolesche Konnektoren verknüpft oder in Form einer linearen Anfrage als Menge von möglicherweise gewichteten Bedingungen spezifiziert werden können. Hier muss die Retrievalfunktion dann die Gewichtungen eines Dokumentes bezüglich der

einzelnen Suchbedingungen passend verrechnen. Hierzu können wir uns an die Definition der Semantik der von uns entwickelten Anfragesprache XIRQL anlehnen [FG01].

## Bibliografie

- Ammon, U. *Variationslinguistik/ Linguistics of Variation/ La linguistique variationelle*. Tübingen: Niemeyer (Sociolinguistica 12), 1998.
- Biella, D., E. Dyllong, H. Kaiser, W. Luther, and Th. Mittmann. "Wege zur digitalen Erfassung der Nachwirkung Nietzsches in Deutschland von 1865-1945. Ein Arbeitsbericht zum Duisburger Retrodigitalisierungsprojekt." Kolloquium "Vom Umgang Nietzsches mit Büchern zum Umgang mit Nietzsches Büchern", Weimar 23.09.-25.9.2002, erscheint in einem Sammelband.
- Biella, D., E. Dyllong, H. Kaiser, W. Luther, and Th. Mittmann. "Edition électronique de la réception de Nietzsche des années 1865 à 1945." *Proceedings of ICHIM03, Paris*, . 8.-12. Sept. 2003.
- Camps, R., and J. Daudé. "Improving the efficacy of approximate personal name matching." *Proceedings of the 8th International Conference on Applications of Natural Language to Information Systems (NLDB'03)*. 2003. Accessed 2005-04-13. <<http://www.lsi.upc.es/dept/techreps/ps/R03-9.ps.gz>>
- Fuhr, N., N. Gövert, and K. Großjohann. "HyREX: Hypermedia Retrieval Engine for XML." *Proceedings of the 25th Annual International Conference on Research and Development in Information Retrieval*. New York: ACM, 2002. 449. Accessed 2005-04-13. <<http://www.lsi.upc.es/dept/techreps/ps/R03-9.ps.gz>>
- Fuhr, N., and K. Großjohann. "XIRQL: An XML Query Language Based on Information Retrieval Concepts." *ACM Transactions on Information Systems* 22 (2004): 313-356. Accessed 2005-04-13. <<http://www.lsi.upc.es/dept/techreps/ps/R03-9.ps.gz>>
- Keller, R. *Die Deutsche Sprache und ihre historische Entwicklung*. Hamburg: Helmut Buske Verlag, 1986.
- Pilz, Th. *Unschärfe Suche in Textdatenbanken mit nichtstandardisierter Rechtschreibung am Beispiel von Frakturtexten zur Nietzsche-Rezeption*. Staatsexamensarbeit, Universität Duisburg-Essen, 2003.
- Zobel, J., and P. Dart. "Phonetic String Matching: Lessons from Information Retrieval." *Proceedings of the 19th International Conference on Research and Development in Information Retrieval (SIGIR'96)*. Ed. H.-P. Frei, D. Harman, P. Schäuble and R. Wilkinson. New York: ACM, 1996. 166-172. Accessed

2005-04-13. <[http://www.lsi.upc.es/dept/tech\\_reps/ps/R03-9.ps.gz](http://www.lsi.upc.es/dept/tech_reps/ps/R03-9.ps.gz)>

## ***The Walt Whitman Archive: Archivist-Scholar Collaboration in Description and Representation***

---

***Kenneth Price*** ([kprice2@unl.edu](mailto:kprice2@unl.edu))

*University of Nebraska-Lincoln*

***Katherine Walter*** ([kwalter1@unl.edu](mailto:kwalter1@unl.edu))

*University of Nebraska-Lincoln*

***Terence Catapano*** ([thc4@columbia.edu](mailto:thc4@columbia.edu))

*Columbia University*

***Daniel Pitti*** ([dpitti@virginia.edu](mailto:dpitti@virginia.edu))

*University of Virginia*

---

### **Topic**

**A**rchivists, librarians, scholars, and technologists are collaborating to build *The Walt Whitman Archive*, an emerging digital thematic research collection that sets out to make Whitman's vast work easily and conveniently accessible to students, researchers, and the general public. This massive undertaking is complicated in part by the physical dispersal of Whitman's manuscripts, which are located in more than seventy repositories in the United States, the United Kingdom, and in France. The sheer volume of materials produced by Whitman has required the current project team to narrow its focus, initially at least, to a more manageable subset of the manuscripts - namely, the poetry manuscripts.

While network and computer technologies have made it possible to build a virtual archive, the intellectual and technical complexities in creating and maintaining the collection in accordance with archival and scholarly standards require an unusual and close collaboration across professional communities and among multiple institutions. Several institutions are intimately involved in the project including the University of Nebraska-Lincoln, the Institute for Advanced Technology in the Humanities at University of Virginia, the New York Public Library, the Harry Ransom Center at the University of Texas at Austin, and Duke University. The project as a whole has contributed to a clearer understanding of technical issues relating to standards that are still in development, and especially to the integration of those standards.

The poetry manuscripts, which have been the most recent focus of the Whitman project team, are scattered in over thirty repositories. At the 2002 ALLC/ACH conference in Tübingen, Germany, a preliminary report on the project entitled *Ordering Chaos: A Virtual Archive of Whitman's Manuscripts* was presented by Mary Ellen Ducey, Andrew Jewell, and Kenneth M. Price. Subsequently, the Whitman EAD project team has successfully created *An Integrated Guide to Walt Whitman's Poetry Manuscripts* using Encoded Archival Description (EAD). XSLT stylesheets are used to harvest information from various repositories' finding aids so as to create an integrated finding aid with links back to the original versions. As the project comes to closure, participants have found it both exciting and revealing on many levels. Representing the different community perspectives (scholar, archivist, librarian, and technologist), the speakers will explore both the opportunities and the challenges of working together and will discuss the implications of such collaboration for the future of each profession.

## Organization

**K**enneth M. Price, Co-Director of the *Walt Whitman Archive* and recently named Co-Director of Digital Research in the Humanities at the University of Nebraska-Lincoln, will describe the widely distributed Whitman manuscripts, the complex history and publication of Whitman, and the history and objectives of the *Archive*. He will discuss the reasons behind the decision to use item-level EAD as a means of bibliographic and editorial control and of user access. Our project is demonstrating the power of EAD to pull together dispersed collections and create a single, scholarly-oriented view or collocation of the materials. We are also addressing an unresolved issue in digital scholarship, namely how best to integrate description and transcription (EAD and Text Encoding Initiative [TEI] files).

Katherine L. Walter, Chair of Digital Initiatives & Special Collections and, with Price, Co-Director of Digital Research in the Humanities at the University of Nebraska-Lincoln, will describe collaborative efforts to provide integrated descriptive access to Walt Whitman's poetry manuscripts. In some cases, our EAD files are based upon encoding previously done by the holding repositories themselves; in other cases, we have created EAD files based upon paper records. We invariably add scholarly information to records, and we offer this additional information back to the individual repositories. Whitman scholarship is complicated by the fact that the poet only occasionally titled his manuscripts, and when he did, he often used a title different from that employed in any of the six distinct editions of *Leaves of Grass*. The project is ordinarily able to identify manuscripts that puzzle non-specialists, and

we also supply date range, uniform title, and Whitman work IDs within the files.

Terence Catapano, Librarian at Columbia University, will discuss the complementary use of current technical standards, in particular EAD, EAC, METS, MODS, and TEI. It is still an open question how these overlapping standards - created by various communities - can be best integrated and used effectively in this kind of highly detailed collection. The *Whitman Archive* is sufficiently large, ambitious, and visible to make it a good case study for testing the integration of metadata standards. We have made significant progress in the use of TEI and EAD; we have recently begun to employ METS in relation to EAD and TEI; we plan soon to work on METS in relation to MODS as well. The periodical printings of Whitman's poetry will be used to research the use of MODS and its integration with the other metadata standards. One challenge is to figure out what role each standard is to have, and how they are to interrelate. For example, descriptive metadata resides in EAD (its primary purpose), in TEI headers (a secondary purpose), and in METS. Which of the three has the authoritative data, and which data should be derived from this authoritative source?

Daniel Pitti, Associate Director of the Institute for Advanced Technology in the Humanities of the University of Virginia, will moderate the session and, in conclusion, will discuss the implications of collaboration for the future of digital scholarship. Digital thematic research collections are valuable resources being developed through collaborations between the library/archival and scholarly communities. One of the points that has been made regarding such collections is that the essential standards shaping the infrastructure for these collections are being developed primarily in the library/archival communities. *The Walt Whitman Archive* is demonstrating that a strong collaboration with equally important contributions from different professional communities working together offers another important model for the future of digital scholarship, and for the development of standards.

# Exhibition: A Problem for Conceptual Modeling in the Humanities

---

**Allen H. Renear** (*renear@uiuc.edu*)

*GSLIS, University of Illinois at  
Urbana-Champaign*

**Jin Ha Lee**

*GSLIS, University of Illinois at  
Urbana-Champaign*

**Yunseon Choi**

*GSLIS, University of Illinois at  
Urbana-Champaign*

**Xin Xiang**

*GSLIS, University of Illinois at  
Urbana-Champaign*

---

## Contents

1. Introduction
2. Exhibition
3. An Example
4. Why Exhibition is a Problem for Modeling
5. Relevant Work
6. False Resolutions
7. Conclusion

## Introduction

There has recently been increased interest within the humanities computing community in formalizing the 'semantics' of document markup (e.g., Sperberg-McQueen et al., Buzzetti, Witt, Renear et al. 2002, Bayerl et al., Dubin et al., Sasaki), or, in an alternative characterization, developing 'conceptual models' that generalize the representation of the textual structures (Cover). We endorse this agenda, which has been a long time in the making (Raymond et al.), but we also wish to draw attention to some difficulties that may be unique to cultural material.

Natural human communication is characterized by what might be termed *plenary semiosis*. Without waiting for formal languages to be provided humans immediately proceed to attempt to say everything they think, and, at least arguably, they generally succeed. The result is that natural communication systems exhibit every imaginable feature that troubles knowledge engineers: fuzzy predicates, modal notions, non-extensional contexts, incompleteness, inconsistency, ambiguity, and so on. But there is an additional complexity as well. Human communication takes place within social contexts that, as linguists and philosophers have been telling us for some time, confound efforts to conceptualize it as sets of assertions only. These two aspects of human communication, plenary semiosis and multiple interacting levels of non-assertional representation combine to produce some of the most difficult, and significant, features of communicative artifacts. We describe one of these features and argue that unless current conceptual modeling systems are extended to accommodate this and other related features those systems will be inadequate for the representation of cultural objects.

## Exhibition

In ordinary linguistic communication we often use a name to refer to something in order to then go on to attribute some property to that thing. However when we do this we do not naturally construe our linguistic behavior as being at the same time an assertion that the thing in question has that name. We do however have a particular cognitive relationship to this latter state-of-affairs; it is just that this attitude is not one of assertion — we *rely* on, or are *committed* to, or *presuppose* that the thing in question has the name we are using to refer to it, but we are not *asserting* that it does.

We refer to this relationship as *exhibition*. We say that the brief document/utterance "*Moby Dick* was written by Herman Melville" *exhibits* the state of affairs that "the name of the author of *Moby Dick* is 'Herman Melville'", but it does not *assert* that state of affairs. What it does assert is that Melville is the author of *Moby Dick*. Although naming is our prototypical example of exhibition in this paper, we believe that exhibition is a widespread and diverse phenomenon.

## An Example

Consider this XML markup, adapted from the *TEI Guidelines* (P4):

```
<bibl>
<author>Edward R. Tufte</author>
<title>Envisioning Information</title>
<pubPlace>Cheshire, Conn.</pubPlace>
```

```
<publisher>Graphics Press</publisher>
</bibl>
```

The *Guidelines* characterize these element types as follows:

- `author`: "... contains the name of the author(s), personal or corporate, of a work ...".
- `title`: "... contains the title of a work ...".
- `publisher`: "...provides the name of the organization responsible for the publication ... of a bibliographic item".
- `pubPlace`: "contains the name of the place where a bibliographic item was published."

Close reading of these definitions reveals that these markup tags convey two quite different sorts of information:

#### Set A

1. Edward R. Tufte authored *Envisioning Information*.
2. *Envisioning Information* was published by Graphics Press.
3. *Envisioning Information* was published in Cheshire, Connecticut.

#### Set B

1. The name of the author of [this book] is "Edward R. Tufte".
2. The name of the publisher of [this book] is "Graphics Press".
3. The name of the place where [this book] was published is "Cheshire, Connecticut".

## Why this is a Problem for Modeling

First, note that the markup is overloaded. The markup tag `author` is used to say that something is a name, *and* it is also used to say that someone is an author (or the author of a particular book). Consider a representation in any commonly used data modeling language, say, RDF's graph-based representation: nodes for individual entities and arcs for binary relationships between them. We would expect a single arc for the assertion represented by a single element — but here apparently a single element must be unpacked into two arcs. TEI specialists are fond of saying that TEI markup is about the text, not about the world the text is about; but we see plainly this isn't always so. And we also note that this overloading crosses a profound and famously troublesome semantic boundary: that between *using* an expression and *mentioning* it.

But the most revealing feature of this analysis is that when we take the union of assertions in sets A and B we will have a model of possible semantic content that, as a whole, is almost certainly incorrect; at least in this sense: it is unlikely that there is any single communicative object whose semantics is correctly modeled by this set of assertions. There are two cases to consider; we present them using some terminology ('expression', 'work') from the *Functional Requirements for Bibliographic*

*Records* (IFLA 1997) and distinguish two senses of 'XML document' (Renear et al. 2003).

1. Consider first an XML document that is understood to be a symbolic expression realizing an intellectual work such as, say, a manual about web design. Such a document will be correctly understood as making the assertions in Set A, but not as asserting any of the assertions in Set B.
2. Now consider an XML document that is a transcription of a source text, that is, a document that is an expression realizing a work which is itself a "theory of the text" (Sperberg-McQueen); that text (expression) being the text of the manual (a work). Such a document would generally be understood as making the assertions in Set B, but not as asserting any of the assertions in Set A.

As a consequence a correct graph model for either case cannot represent the assertions (as assertions) in the other case. However a correct representation of Case 1 could represent Case 2 assertions as exhibitions — if specific expressive devices, qualified arcs say, were available for this representation. This is the extension that we are recommending.

Some clarification of these intricacies may be useful. First note that the cases are not isomorphic: Case 1 asserts the propositions in Set A and exhibits those in Set B, but Case 2, although asserting the propositions in Set B, does *not* exhibit the propositions in Set A. While might be plausibly argued that the propositions in Set B logically imply those in Set A, and so any document that asserts Set B asserts Set A, we would resist this for two reasons: first because the intuitive logic of assertion simply does not seem to require that all logical implications of asserted propositions are themselves asserted; and second because we suspect that a completely correct presentation of Set B, one more in line with TEI doctrine on the textual orientation of markup, would eliminate all commitments to books, authors, and authorship, and that paraphrase would block the logical implications in any case. What one could say however is that in Case 2 the Set A propositions occur in *oratio obliqua*.

We also note, as an illustration of the usefulness of the concept of exhibition, that scholarly transcription into TEI markup can be understood as identifying exhibitions and then re-expressing them as assertions.

## Relevant Work

The rudiments of this problem have already made an appearance in the Semantic Web and Dublin Core communities. However we do not think its significance, at least for cultural material involving human communication, is fully appreciated. Dan Brickley, chair of the W3C Semantic Web Interest Group has noted that the Dublin Core `dc:creator`

element is defined in a way that encourages a similar confusion between names and things (Brickley), not surprisingly, as the definition of `dc:creator` is similar in logical structure to the ones we cite from the *Guidelines*:

"Examples of a Creator include a person, an organization, or a service. Typically, the name of a Creator should be used to indicate the entity." (DCMES 2003)

The varying usage of the `dc:creator` code (sometimes for Creators, sometimes for their names) amongst metadata encoders is now recognized as a serious practical problem for the development of an 'abstract model' for Dublin Core. (Powell).

It is of course in the context of efforts to be absolutely precise and formal that the problem is acute. Jeremy Carroll, co-editor of the W3C Recommendation *Resource Description Framework (RDF): Concepts and Abstract Syntax*, writes in a posting on `w3c-rdfcore-wg`:

I have been looking through the (RDF) primer, particularly looking at the Dublin Core examples (throughout the primer). These seem like perfectly fair examples of how Dublin Core is used. Unfortunately, there are many instances where strings are used to represent people and things rather than themselves. This is not in agreement with the model theory...

(Carroll)

Carroll then goes on to note that given the RDF model theory incorrect implications immediately ensue: in our example for instance, that *Moby Dick* was authored by a string rather than by a person.

## False Resolutions

Three deflationary perspectives on this problem are possible.

One, anticipated from the TEI community, is that TEI encoding always represents the features of the linguistic text only and 'real-world' assertions are either misunderstandings, mistakes, or anomalies. This may be so, although we are skeptical as to whether this stance can be maintained with respect to the full range of TEI applications. But in any event exhibition remains a common feature of communicative artifacts, characteristic of many XML element sets, and of many other systems of symbolic communication. It must be accommodated.

Another approach, this one anticipated from the Semantic Web community, is simply to insist on an unambiguous corrected conceptual representation: one arc for being named "Herman Melville", one for authoring *Moby Dick*. But this resolution fails for the reasons presented in the preceding section. Although this model would be in some sense an accurate representation of "how the world is" according to the document,

it would not represent what is *asserted* by the document. The authorship arc in the corrected RDF graph model will correspond to relationships of exhibition, not assertion; and there is no accommodation for this distinction in the modeling language.

Finally, it is also natural to feel that the phenomenon of exhibition is similar in some respects to the already noted much studied phenomenon of linguistic *presupposition* and to wonder whether exhibition is simply a special case of presupposition (Levinson). Currently we are undecided on this issue but we note that even if exhibition does turn out to be a form of presupposition that would remove neither the difficulty exhibition creates for conceptual modeling, nor its intellectual significance. In fact it would be a rather substantial finding to determine the matter one way or the other.

## Conclusion

The phenomenon of exhibition is not limited to the simple naming examples used above. We believe it is characteristic of communication and communicative cultural artifacts in general. For instance when we title our articles we do not say that the title is a title, although we exhibit it as a title, allowing that inference to be drawn (Renear). Or for a quite different sort of case: consider how morphological distinctions exhibit our commitments to syntactical roles, without actually asserting that the words in question are playing those roles — though indeed we use those words with those particular grammatical and syntactical properties in order to make the assertions we do make.

We conclude that current conceptual modeling projects within the humanities computing community will fail to be adequate for the study of cultural objects if they take the approach of the Semantic Web community and see exhibition as a simple problem of ambiguity or error, rather than defining new constructs to express these distinctive relationships. To be adequate for the humanities, conceptual modeling must be extended to accommodate the data of the humanities.

## Bibliography

Bayerl, P.S., H. Lungen, D. Goecke, A. Witt, and D. Naber. "Methods for the Semantic Analysis of Document Markup." *Proceedings of the 2003 ACM symposium on Document Engineering*. ACM Press, 2003. 161-170.

Brickley, D. *Using Dublin Core Creator*. FOAF Wicki, July 2003. Accessed 2005-03-21. <<http://rdfweb.org/topic/UsingDublinCoreCreator>>



- Buzzetti, D. "Digital Representation and the Text Model." *New Literary History* 33.1 (2002): 61-88.
- Carroll, J. *Dublin Core, the Primer and the Model Theory*. Posting in w3c-rdfcore-wg, May 16, 2002 10:32:42. Accessed 2005-03-21. <<http://lists.w3.org/Archives/Public/w3c-rdfcore-wg/2002May/0040.html>>
- Cover, R. *Conceptual Modeling and Markup Languages*. Cover Pages, January 24, 2001. Accessed 2005-03-21. <<http://xml.coverpages.org/conceptualModeling.html>>
- Dubin, D., C.M. Sperberg-McQueen, A. Renear, and C. Huitfeldt. "A Logic Programming Environment for Document Semantics and Inference." *Literary and Linguistic Computing* 18.2 (2003): 225-233.
- Dublin Core Metadata Element Set. Version 1.1 Reference Description*. DCMI, 2003. Accessed 2005-03-21. <<http://dublincore.org/documents/dces/>>
- International Federation of Library Associations (IFLA). Functional Requirements for Bibliographic Records: Final Report. UBCIM Publications-New Series*. Munchen: K.G.Saur, 1998.
- Levinson, S.C. "Chapter 4: Presupposition." *Pragmatics*. Cambridge: Cambridge University Press, 1983. 167-225.
- Powell, A. *DOAP*. Posting in "Creative Commons Metadata", July 16, 2004:33:48 EDT. Accessed 2005-03-21. <<http://lists.ibiblio.org/pipermail/cc-metadata/2004-July/000421.html>>
- Raymond, D.R., and F.W. Tompa. "Markup Reconsidered." *Technical Report 356*. Department of Computer Science, The University of Western Ontario, 1993. Presented at the First International Workshop on the Principles of Document Processing, Washington DC, October 21-23 1992; an earlier version was circulated privately as "Markup Considered Harmful" in the late 1980s.
- Renear, A. "The Descriptive/Procedural Distinction is Flawed." *Markup Languages: Theory and Practice* 2.4 (2001): 411-420.
- Renear, A., D. Dubin, C. M. Sperberg-McQueen, and C. Huitfeldt. "Towards a Semantics for XML Markup." *Proceedings of the 2002 ACM Symposium on Document Engineering*. Ed. R. Furuta, J.I. Maletic and E. Munson. McLean, VA, November 2002. 119-126.
- Renear, A., H.C. Phillippe, P. Lawton, and D. Dubin. "An XML Document Corresponds To Which FRBR Group 1 entity?" *Proceedings of Extreme Markup Languages 2003*. Ed. B.T Usdin and S.R. Newcomb. Montreal, Canada, August 2003.
- Sasaki, F. "Combining Markup Semantics and Semantic Markup: A Secret Marriage." *Proceedings of ALLC/ACH 2004*. Goteborg Sweden, 2004. 122-125.
- Sperberg-McQueen, C.M. "Text in the Electronic Age: Textual Study and Text Encoding, With Examples from Medieval Texts." *Literary and Linguistic Computing* 6 (1991): 34-46.
- Sperberg-McQueen, C.M., A. Renear, and C. Huitfeldt. "Meaning and Interpretation of Markup." *Markup Languages: Theory and Practice* 2.3 (2000): 215-234.
- Witt, A. "Meaning and Interpretation of Concurrent Markup." *Proceedings of ALLC/ACH 2002*. Tuebingen, 2002.

## L'Autoguidage: une Approche pour le Perfectionnement du Français Écrit en Milieu Minoritaire

---

*Sylvain Rheault (sylvain.rheault@uregina.ca)*  
*University of Regina*

---

### Situation des francophones en milieu minoritaire

Le perfectionnement du français écrit pour les francophones en milieu minoritaire présente des défis considérables. Les francophones du Canada sont dispersés sur un territoire immense et n'habitent pas tous à proximité d'une institution d'enseignement. Plusieurs d'entre eux sont des adultes sur le marché du travail qui souhaiteraient obtenir plus de formation mais sans avoir à retourner sur les bancs d'écoles. D'autres sont de jeunes francophones aux études qui voudraient ajouter à leur programme un cours de perfectionnement. Pour compliquer un peu plus les choses, le bagage linguistique d'un apprenant à l'autre peut présenter des divergences importantes. De plus, comme il s'agit de francophones en milieux minoritaires, il faut tenir compte de l'influence de l'anglais. Le cours de français écrit idéal pour cette clientèle hétérogène et éparse devrait être en mesure de s'adapter aux besoins individuels des apprenants.

L'Internet apparaît comme l'outil capable de relever les défis qui viennent d'être énumérés. Tout le monde connaît déjà les outils éprouvés de l'enseignement électronique à distance, comme l'affichage d'informations et d'hyperliens, les courriels, les forums de discussions, bref tout ce qui permet de simuler une salle de classe en ligne. Cependant, il reste encore le défi d'enseigner à une classe grandement disparate.

Le présent projet de communication pour le congrès présentera une application nouvelle dont le potentiel pourrait utilement servir les types d'apprenants décrits plus haut. Il s'agit de l'autoguidage, soit la capacité, pour un cours informatisé, d'adapter la matière d'enseignement à chacun des apprenants. Avant d'en parler plus avant, il importe de dire quelques mots sur le CAFÉ.

### Le CAFÉ

CAFÉ est l'acronyme de "Cours Autodidactique de Français Écrit". Ce cours, créé par Dupriez, conçu pour l'apprentissage individuel, a vu le jour en 1966. Il a été enseigné et a fait l'objet d'expérimentations dans plusieurs pays de la francophonie ainsi que divers lycées (Paris et régions). À propos de la "formule CAFÉ" il faut savoir que le traitement statistique des réponses aux milliers de QCM expérimentées dans des groupes régionaux permet de mesurer scientifiquement l'utilité probable des questions à poser à tout nouvel apprenant, en fonction de son niveau et de son progrès. La formule CAFÉ permet d'individualiser l'apprentissage, aussi efficacement que rapidement. Sa force est aussi dans ses contenus informatisés (12000 QCM sur les fautes courantes et les difficultés rencontrées en rédaction et dissertation).

Les statistiques obtenues pour la France et le Québec peuvent ne pas convenir à une population francophone vivant en milieu minoritaire. Il faudra donc obtenir des statistiques pour la population visée. Cette tâche sera réalisée au moyen de questionnaires accessibles sur l'Internet. Il faudra aussi créer des QCM (questions à choix multiple) traitant de difficultés spécifiques aux francophones vivant en milieu minoritaire. Ces items sont produits à partir des erreurs les plus fréquentes relevées dans les travaux d'étudiants.

À propos de la méthode basée sur des questions à choix multiple L'enseignement du français, contrairement à d'autres matières comme les mathématiques ou la chimie, doit compter avec les particularités propres à chaque région. On n'enseigne pas le français de la même façon à un anglophone, à un arabophone, etc. Dû aux interférences linguistiques, chaque région de la francophonie a des défis qui lui sont particuliers. Ainsi, les régions d'Afrique centrale doivent compter avec les langues bantoues, le sangö, etc. Au Québec, comme en France, les anglicismes sont fréquents. Les interférences sont particulièrement difficiles à dépister, lorsqu'elles touchent non pas le vocabulaire mais la syntaxe. C'est en adaptant les questions et les statistiques aux particularités linguistiques de chacune de ces régions qu'on peut faire réaliser à chaque utilisateur un maximum de progrès.

Comment le niveau des questions est-il établi? Pour le Québec, un sondage a été effectué dans les cégeps et les universités. Plus de 8000 (75 x 110) questions à choix multiple ont été posées à 12000 étudiants entre 1995 et 1997. Avec 100 répondants ou plus par question, il est possible de mesurer la cote de difficulté, de discriminance et de fiabilité avec précision.

Les améliorations notées démontrent que les apprenants du français pour lesquels un tel cours est élaboré en tirent un avantage considérable. Ils développent une conscience plus aiguë de l'usage, en particulier lorsqu'il peut y avoir des

interférences entre le français et l'anglais, sans avoir à oublier l'une de ces langues pour mieux s'exprimer dans l'autre.

## La méthode de l'autoguidage

Le CAFÉ a d'abord été conçu comme un cours par correspondance. La méthode du cours, basée sur les besoins des apprenants, consiste à donner, dans le corrigé, la règle en jeu avec des contre-exemples pour chacun des autres choix de réponses, l'objectif étant de susciter chez l'apprenant une démarche réflexive où il compare les façons d'écrire en contexte et selon la norme. Les méthodes du CAFÉ sont scientifiquement reconnues au Québec, en France et en Afrique.

### Parcours linéaire



Figure 1: enseignement linéaire

La plupart des étudiants qui débutent à l'université possèdent déjà des compétences en français, ce qui n'est pas le cas pour une discipline comme l'anthropologie par exemple. Traditionnellement, l'enseignement consiste à commencer avec les éléments les plus faciles (niveau -1,0), puis de présenter des éléments de plus en plus difficiles (niveaux -0,8 à 1,0). Pour les étudiants qui ne savent rien, il s'agit du cours idéal. Pour les étudiants qui ont déjà des compétences, la première partie du cours sera un peu pénible, mais, au bout d'un certain temps, ils apprendront des choses nouvelles. Enfin, pour les étudiants très avancés, un cours qui ne leur offre que de la matière déjà apprise sera d'un ennui mortel. Ils risquent de ne pas s'impliquer activement, voire d'abandonner avant la fin.

### Parcours autoguidé

- Bonne réponse
- Mauvaise réponse

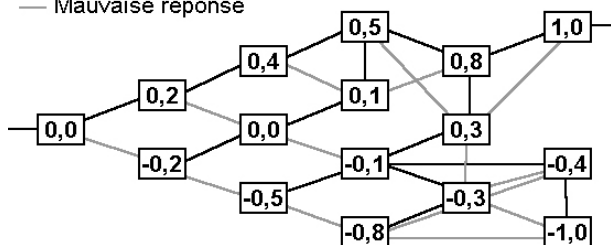


Figure 2: enseignement par fourches

En gros la méthode de l'autoguidage consiste à donner à l'apprenant la matière dont il a besoin. La progression est basée sur la compétence déjà acquise. En simplifiant, on peut décrire ainsi la méthode: le cours commence avec des éléments moyens. Si l'apprenant répond bien, on lui offre une matière un peu plus difficile. Sinon, on lui offre une matière un peu plus facile, jusqu'à ce qu'il soit parvenu au niveau qui lui convienne le mieux. À partir de là, l'apprenant peut vraiment commencer à apprendre. Le niveau est recalculé constamment, à chaque nouvelle réponse. Le système informatisé de guidage choisit comme prochaine question celle qui offre le plus d'intérêt pour l'apprenant, c'est-à-dire celle qui est la plus proche de son niveau atteint. De plus, chacun des apprenants travaille à son rythme, sans ralentir les autres ni se voir bousculé.

### Parcours de l'apprenant moyen

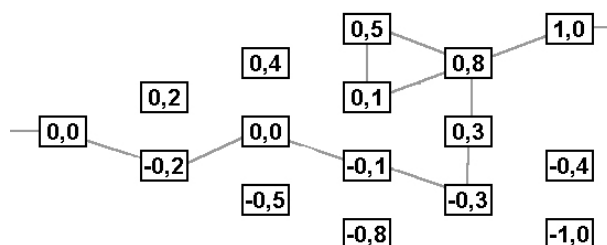


Figure 3: apprenant "moyen"

Un apprenant moyen passe par l'essentiel de la matière et peut passer outre certains items trop faciles.

### Parcours de l'apprenant fort

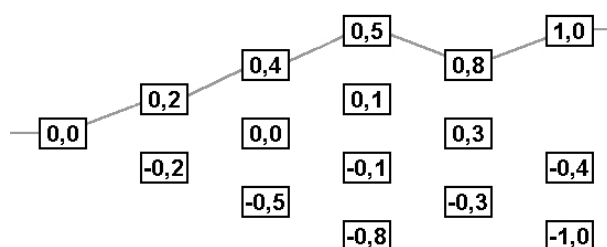


Figure 4: apprenant "fort"

Il serait possible, à un étudiant particulièrement brillant, de n'avoir à répondre qu'à un très petit nombre d'items. C'est qu'il maîtrisait déjà bien la matière qu'on lui propose. En ce cas, le cours est très vite terminé.

## Parcours de l'apprenant faible

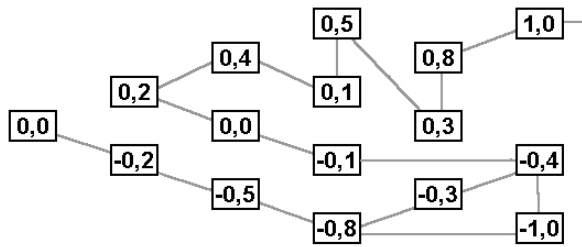


Figure 5: apprenant "faible"

L'apprenant faible sera d'abord guidé vers des questions plus faciles avant d'être ramené à des questions plus subtiles. À la fin, les étudiants forts, moyens et faibles auront tous atteint l'objectif. Le zélé peut foncer, le faible peut prendre son temps.

## Bibliographie

CAFÉ. Accessed 2005-04-12. <[http://www.cafe.edu](http://www.cafe.edu/)>

## National Support for Humanities Computing: Different Achievements, Needs and Prospects

*David Robey* ([d.j.b.robey@reading.ac.uk](mailto:d.j.b.robey@reading.ac.uk))

*Arts and Humanities Research Board*

*John Unsworth* ([unsworth@uiuc.edu](mailto:unsworth@uiuc.edu))

*University of Illinois*

*Geoffrey Rockwell* ([georock@mcmaster.ca](mailto:georock@mcmaster.ca))

*McMaster University*

Different countries vary greatly in the kinds of national structures, services and funding resources they provide for humanities computing. Most if not all of us involved in the field are faced by similar challenges, though in different ways and varying degrees: lack of IT awareness in our subject communities, shortage of technical support, insufficient academic recognition, the need to take advantage of new technological developments in other fields, the need for more, or better, digital data resources and tools, problems of standards, and much more. The purpose of this session is to explore the ways different countries have met and are meeting these challenges: differences in existing provision at national level, whether publicly or privately funded; different national needs and procedures for identifying them; new strategies and initiatives for the future.

Three panellists represent Canada, the UK and the USA, and each occupies a key position in humanities computing in their country. Between now and the conference we shall recruit at least one further panellist, from a non-English-speaking country [our planned fourth contributor has had to withdraw]. Even on their own, however, these three English-speaking countries show striking differences. Each speaker will make a brief presentation in respect of his own country on:

- national-level support structures, services and funding resources;
- new and forthcoming strategic initiatives;
- national needs: how are they or can they be identified, and how far have they not yet been met?

The ensuing discussion will invite relevant input about other countries, seek to identify the different issues and problems

raised by different forms of support, consider how far national needs differ, and explore the scope for new forms of international collaboration in the area.

## **David Robey (Director, AHRB ICT in Arts and Humanities Research Programme): the UK**

**T**he UK has latterly developed a strong system of support by national public funding bodies for humanities computing. Two funding bodies are involved: the Arts and Humanities Research Board (AHRB), and the Joint Information Systems Committee (JISC). Each has provided substantial support for data-creation projects, generally through open competition: out of over £100m spent by the AHRB on research project awards since 1999, about half has been given to projects with some kind of digital output. Some limited support has also been given to the creation of software tools and systems of specifically humanities interest. As far as services are concerned, the AHRB and JISC jointly fund the Arts and Humanities Data Service, to the extent of some £1m p.a., with a brief to provide national support for data creation and preservation, including a national repository for the digital output of publicly funded research. JISC and the AHRB jointly fund Humbul, the humanities hub of the Resource Discovery Network, providing portal access to on-line resources in the UK and internationally. JISC has also funded a new suite of on-line introductory training resources in the use of ICT for humanities researchers who have not progressed beyond the elementary use of word-processing, email and web-browsers (ARIA).

Launched last year, the ICT in Arts and Humanities Research Programme, a major new initiative by the AHRB, is funding two new initiatives: the ICT Methods Network, a centre for the exchange and dissemination of advanced ICT methodologies in arts and humanities research, which will complement the AHDS by focussing on methods, processes and uses of data; and a scheme of ICT Strategy Projects, which will partly gather knowledge about ICT uses and needs in the arts and humanities, and partly develop generic ICT resources and tools in the area. The Programme is also promoting and developing an arts and humanities e-science, or e-research, agenda, taking advantage of the high-profile activities currently taking place under this heading in other domains. All these public supports are funded for the medium term only, until 2007. In 2006 the AHRB will conduct a fundamental strategic review of ICT support provision and the related needs of the arts and humanities research communities, with a view to determining longer-term systems of provision.

The UK therefore has some strong support structures in place, and plans have been laid for better determining future needs. Nevertheless a lot of problems remain at the level of the individual humanities researcher. There is insufficient awareness in the research community of the potential of digital methods and resources; most importantly generic, low-level and specialist technical support are inadequate except in a very small number of institutions. The programme of public activities that is now being funded will be able to do something about the first problem, but it will be a long time before the second one is solved.

## **Geoffrey Rockwell (Project Leader for the Text Analysis Portal for Research project funded by the Canada Foundation for Innovation): Canada**

**T**his discussion will offer a brief history of the major centres and organizations of humanities computing in Canada, with special attention to the emergence of a national society; an overview of the new and coming programs, both undergraduate and graduate, and the types of research positions being created to support them; and a survey of some of the major research projects and how Canadian researchers have worked with provincial and federal funding agencies like OIT (Ontario Innovation Trust), SSHRC (Social Science and Humanities Research Council) and CFI (Canada Foundation for Innovation). Special attention will be given to some current projects including the TAPoR (Text Analysis Portal for Research) project, and a new initiative, iMatter, that is developing the case for a national digital arts and humanities research institute--an initiative which is working within the context of a major transformation of our research council.

The presentation will be aimed primarily at describing strategies that have worked and may be applicable elsewhere, and at describing opportunities for transnational funding. Humanities computing in Canada has developed a national network around a society with annual meetings, we have developed a network of programs, centres and research faculty positions. What is missing?

One relationship that is stronger in other countries is the connection with the library and information science community. Another relationship is with media studies and journalism. A third and important next step is to make strategic alliances that will allow us to develop contextualized knowledge, especially theory, that reflects the pragmatics of digital humanities work. In this we can work with the digital artists, especially those working in university contexts who also do research and

creation. Research/creation, a term drawn from the UK AHRB for a new grant program, is a unique form of academic practice that combines the communicative, the critical and the creative. In other words, it is time to think systematically and together about doing and making, and to think through doing and making. In such an endeavor we can also reach forward into the emerging games studies or interactive arts community which, we believe, will develop along similar lines that weave research and creation together. We are beginning to tell each other stories of what learning and research will look like in the next generation of institutes.

## **John Unsworth (Chair, ACLS Commission on Cyberinfrastructure for Humanities and Social Sciences): the USA**

**I**n the United States, there is no single source of funding for humanities computing activities. The National Endowment for the Humanities has, over the past decade, funded a number of important humanities computing projects, especially in the realm of editorial work, and more recently in the online state encyclopedias. The Institute for Museum and Library Services (now more well-funded than NEH) will also sometimes provide funding for digital scholarly projects, if they include library, archive, or museum participation. The National Science Foundation has been a more difficult source from which to fund humanities computing projects, but some--especially those that have something to do with speech recognition or natural language processing--have gotten funded.

Probably the most consistent source of support for humanities computing in higher education has been the Andrew W. Mellon foundation, which has supported large collaborative projects across a number of disciplines (literature, archaeology, art history, music, linguistics, etc.). Other private foundations have been involved as well, but none so prominently as Mellon. Mellon is also responsible for some of the most significant non-commercial infrastructure projects, where infrastructure refers to shared resources: *JSTOR*, the social science journals project, is a Mellon creation, as is the more recent *ArtStor*, which brings together images useful for teaching and research in art, art history, archaeology, and other areas. Mellon has also funded *Bibliovault*, at the University of Chicago Press (for converting print backlist books to print-on-demand, for university presses), the *TORCH* project at Oxford UP (for delivering university press materials to individual and institutional subscribers in electronic form), and the electronic imprint at the University of Virginia Press, for publishing born-digital humanities scholarship. Mellon's also been a persistent funder of digital library research, for example the

*FEDORA* project (to develop digital object repository architecture for complex digital content), the *Making of America* project, and many others.

While this history is admirable, and the digital library and humanities computing community owes much to the Andrew W. Mellon foundation, the leadership of that foundation is about to change hands, and that makes it an appropriate time to think about how the base of support for this activity might be broadened, among private foundations and government agencies alike. The recent NSF commission on cyberinfrastructure, led by Dan Atkins, produced a report whose recommendations are also an occasion for self-examination and strategic planning in the humanities and social science communities, as we consider what free-rider benefits there might be, for these communities, from the work that will be done in the computational sciences, and as we consider also what work is not likely to be done in those fields that will be important to the humanities and social sciences. To address those questions, the American Council of Learned Societies has assembled its own commission (with funding from Mellon), and that commission has held public meetings and private information-gathering and working sessions throughout 2004. A report from the commission is forthcoming in 2005, and this presenter, as chair of the commission, can offer a preview of the results of that process, and the conclusions of the report.

---

## **TAPoR: Five views through a text analysis portal (COCH/COSH Allied Association Session)**

---

**Geoffrey Rockwell** ([georock@mcmaster.ca](mailto:georock@mcmaster.ca))

McMaster University

**Stéfan Sinclair** ([sgsinclair@gmail.com](mailto:sgsinclair@gmail.com))

McMaster University

**James Chartrand** ([jc.chartrand@mcmaster.ca](mailto:jc.chartrand@mcmaster.ca))

OpenSky Solutions

---

### **A. Session Introduction**

The *TAPoR* project started as a project to create a portal where users could manage texts, tools and then run tools on text. The Alpha version of the *TAPoR* portal nicely demonstrated the potential of this simple workbench paradigm. *TAPoR.2* builds on the individual project paradigm to make the portal useful for research communities. It does this in a number of ways:

1. We have developed a Try It first encounter interface for use by new users, casual users, and just-in-time users. This interface has been developed in close coordination with usability researchers, though it is now going into extensive testing.
2. *TAPoR.2* allows user information to be saved for groups or made public in a fashion similar to community information portals like *del.icio.us* (<http://del.icio.us>) and *CiteULike* (<http://www.citeulike.org>). Some types of information have always been intended for public viewing like the News built into *TAPoR* from the beginning. We have not only extended the sharing model to all types of information managed, but we have added communal editing to selected types of information, especially documentation, with a wiki editing-like interface.
3. We have extended the project paradigm to allow interfaces to be created that can be integrated into other projects and web sites. Thus advanced users can create projects that are styled to look like part of a different project.
4. We have developed a tool developers interface so that tools as web services can be added and documentation quickly entered. We have also used the community building features

of the portal to develop *TA!DA!* or the *TAPoR Developers Association* – a site for the developer community.

5. We have developed *TEA*, the *TAPoR Engine of Association*, which is designed to help the serendipitous exploration of texts, references, links, people, projects and tools. *TEA* combs and visualizes topic maps which associate items across users.

In this session we are going to present the portal from five views that move from a conventional first encounter view of a tool portal to an inverted view of the portal as a research community association engine. These five views will be presented as three coordinated papers.

### **B1. TAPoR: First Encounters**

**Geoffrey Rockwell**

The first paper will demonstrate the first encounter interface, Try It. Woven into this presentation will be a discussion of the usability research and testing that led to this interface hypothesis. It is our hope that this encounter interface will be of use to novices and advanced, but casual, users. It is an interface that doesn't require a portal account so it can be used occasionally and it is optimized for ease of use and successful results.

Rockwell will then demonstrate the basic user account paradigm for people who want to use the portal for sustained text analysis projects. He will demonstrate how from a first encounter once can get a myTAPoR account with which to organize links to texts, organize tools, and manage projects.

### **B2. TAPoR: Developing Encounters**

**Stéfan Sinclair**

The second paper will demonstrate and discuss the Tool Developers interface and the community tools designed to assist developers. In this context Sinclair will discuss the first *TAPoR* “hackers ball” funded by the Social Science and Humanities Research Council of Canada through a grant led by Stéfan Sinclair. He will also discuss the technical design of the underlying tool broker and the data interfaces that allow results to be saved to a Data Bench for use as an input text for a different tool. This component of the presentation will end with a blatant attempt to enlist attendees in *TA!DA!* so we can enrich the tools collection.

The portal must bring together the text analysis community. In particular, the portal must make it as easy as possible for researchers who have existing tools, or want to write new tools — in their preferred programming language — to make the tools available through the portal. Web services provide a standard language and protocol to enable communication between different programming languages, and therefore are a

very appropriate vehicle for connecting text analysis tools together through the portal. Further, most programming languages provide tools to publish existing program code as web services with little or no modification, and little extra setup. In some cases the tools will take an existing program function and create the entire infrastructure needed to make the function available over the internet: the web server, the code to listen for remote requests and translate them into calls to the local program code, and code to package the results up and return them to the original caller.

Text analysis tools provided as web services are easier to combine in simple ('piped') combinations, but can also be combined in very sophisticated arrangements (using scripting) — without requiring that the user learn new programming languages or run through elaborate setup procedures.

### **B3. TAPoR: Community Encounters**

#### **James Chartrand**

The third paper will discuss the underlying technologies deployed in the portal so as to show how the portal can be rethought as a community association engine. We chose *Apache Cocoon* as our web development framework for the portal. *Cocoon* satisfies several of our objectives. *Cocoon* provides a basic portal implementation geared towards custom development. *Cocoon* is open source. Much of *Cocoon* is made up of code donated from large scale software projects; code that has gone through numerous development cycles on large systems. *Cocoon* is actively maintained and supported by hundreds of developers. *Cocoon* is therefore stable, secure, and scalable. In addition, *Cocoon* runs on Java and therefore, can run without modification on *Linux*, *Windows* and the *Mac*, allowing new projects to install the portal with ease.

The portal must provide a uniform and single point of access for text analysis tools, but must also engender an online community of knowledge. We chose Topic Maps for knowledge management because they are adaptable, simple, and standards based. Topic Maps can be thought of as a very rich index. An index that doesn't just point into texts, but can describe relationships between almost any object or idea. In our case, the relationships are between texts, between tools, between texts and tools, between projects, between projects and tools, between projects and users, between users and texts, and so on. Topic Maps also make the portal more adaptable to the needs of other projects outside the text analysis community.

In the context of underlying technologies James Chartrand will demonstrate the portal again, but now from the view-point of how it can be used to develop a research group or project taking advantage of the incorporated technologies. He will demonstrate the deep skinning features that allow users to create views that suit their research, their groups, or their projects. In this context

he will illustrate how the *TAPoR* portal, is, from one perspective, just a web of associations between links, notes, tools, and topics.

## **C. Issues**

**T**here are a number of key issues that underlie all three papers.

- i. Peer review of tools and academic credit. In a panel organized for the ACH/ALLC 2003 in Athens Georgia by Stéfan Sinclair on "Peer Review of Humanities Computing Software" we presented some models for how review of tools could be supported. *TAPoR* as a public portal that gives access to tools elsewhere that run as web services can be site for the review and documentation of software tools. We will present a documentation interface that allows public comments and reviews of tools that could serve some of the need for a peer review system.
- ii. Open source. A popular paradigm for the creation and maintenance of community tools is to release them as open source under one of the various licenses available. We will discuss the way in which the portal as software is open source and the ways individual tools can be made available or protected. Likewise we will discuss the need for authentication for selected texts which cannot be made available openly.
- iii. Humanities software development. The portal must, fundamentally, meet the needs of a research community. Needs which aren't, by definition, yet completely defined as research evolves. To that end, we have adopted an "agile" development process that involves regular meetings and storytelling. This approach has proven extremely effective. We have avoided getting bogged down in over-analysis and excessive documentation, and at same time have been able to adapt development cycles to meet the evolving needs of the project. Adaptability is particularly important for a research project like this where midstream research outcomes can lead to new paths, or close others.
- iv. Stories. The 'story' is the fundamental unit of work in our process. Stories are informal descriptions of how the end-user would like to use the portal. Stories can be written in whatever style makes sense for the user. Stories and other documentation is kept in the *TAPoR* Wiki which is a shared development space. The stories are then broken down by the Open Sky Solutions team into 'tasks' that are assigned time estimates.
- v. Adaptability. An important objective of the project is to enable other projects to adapt the portal and to contribute to its development. We have, therefore, organized the development process around standards that make it straightforward to not only download and install the portal,



but to setup the development environment. Our goal is to ensure continued development of the portal.

## D. Conclusions

The *TAPoR* Portal is fundamentally conceived of and designed to be an extensible, network-based research environment. As such, it has been crucial to devise mechanisms for enriching the portal by allowing developers and users to encounter the portal, use it, and adapt it for others. It is worth emphasizing how this approach differs from the development of text analysis tools of the past, such as *OCP* and *TACT*, that are essentially pre-defined workstation-based programs. *TAPoR*, by contrast, seeks to accommodate unknown and unanticipated resources. Such flexibility requires considerable engineering to ensure compatibility between disparate texts and tools. We will present a model for such flexibility, but recognize that it will need testing and scrutiny to become genuinely useful.

# The Non-Traditional Case for the Authorship of the Twelve Disputed "Federalist" Papers: A Monument Built on Sand?

---

*Joseph Rudman* ([jr20@andrew.cmu.edu](mailto:jr20@andrew.cmu.edu))  
*Carnegie Mellon*

---

## Introduction

This paper discusses the controversy over the authorship of twelve of the "Federalist" papers as seen and studied by over twenty non-traditional authorship attribution practitioners. The "Federalist" papers were written during the years 1787 and 1788 by Alexander Hamilton, John Jay, and James Madison. These 85 propaganda tracts were intended to help get the U.S. Constitution ratified. They were all published anonymously under the pseudonym, "Publius." The general consensus of traditional attribution scholars (although varying from time to time) is that Hamilton wrote 51 of the papers, Madison wrote 14, Jay wrote 5, while 3 papers were written jointly by Hamilton and Madison, and 12 papers have disputed authorship — either Hamilton or Madison.

In 1964, Frederick Mosteller and David Wallace, building on the earlier unpublished work of Frederick Williams and Frederick Mosteller, published their non-traditional authorship attribution study, "Inference and Disputed Authorship: The Federalist." It is arguably the most famous and well respected example from all of the non-traditional attribution studies. It is the most statistically sophisticated non-traditional study ever carried out. There even has been a 40 page paper explicating the statistical techniques of the Mosteller and Wallace study (Francis). Since then, hundreds of papers have cited the Mosteller and Wallace work and over two dozen non-traditional attribution practitioners have analyzed and/or conducted variations of the original study.

These practitioners wanted to test their statistical approaches against the Mosteller and Wallace touchstone study. Mosteller and Wallace set the boundry conditions for the subsequent work — e.g., not using the Jay articles as a control. Their experimental design and overall report is never questioned. Most of these later practitioners do not select or prepare the input text as rigorously as Mosteller and Wallace — whose

own selection and preparation was not as rigorous and complete as it should have been.

## Text Selection

### (1) "Federalist" Papers

This section discusses the way the Federalist papers were originally published (76 in newspapers and 8 in the book compilation) and which editions the practitioners chose for their non-traditional studies — how 84 papers became 85 and how some papers had different numbers in different editions. The effect that the lack of Hamilton and Madison holographs had on the studies is discussed. The choice of edition has the potential of profoundly changing the results of the studies.

Project Gutenberg Etexts are usually created from multiple editions, all of which are in the Public Domain in the United States, unless a copyright notice is included. Therefore we do NOT keep these books in compliance with any particular paper edition, usually otherwise.

(Front Material of Gutenberg Etext #1404)

The compounding problem of down-loading texts via the internet is explicated — e.g., one of the texts includes every variant of every paragraph. It is shown why none of the Federalist studies used a 'valid' text of the Federalist papers. The question, "Does this incorrect input data invalidate the final 'answer?'" is discussed.

### (2) The Control Texts

#### (a) The "Known" Hamilton Sample

This sample cannot contain questionable Hamilton writings. This sample must also fulfill the other criteria of a valid sample — e.g., same genre, same constricted time frame. There also should be a sub-set of this sample set aside for later analysis in order to guard against the charge of cherry picking the style-markers. This is not the same as the Mosteller and Wallace "training sample."

#### (b) The "Known" Madison Sample

In addition to discussing the way the Madison sample was constructed, what was said about the Hamilton sample will be applied here.

Does the lopsided number of Hamilton papers over Madison papers (51 to 14) pose a problem for the studies? Were the Hamilton and Madison control texts from outside the Federalist papers chosen correctly? Why are these "outside" controls not used by most of the other practitioners? This section goes on

to discuss the control problems that arose with the Mosteller and Wallace study and have been perpetuated through the subsequent studies. This section also discusses the other control problems introduced in these studies.

## Text Unediting, De-editing, and Editing

The cumulative effect of NEARLY A THOUSAND SMALL CHANGES [emphasis mine] has been to improve the clarity and readability of the text without changing its original argument.

(Scigliano, lii)

### (1) The "Little Book of Decisions"

In the Mosteller and Wallace study, a "little book of decisions" is mentioned. This "book," originally constructed by Williams and Mosteller, contained an extensive list of items that Mosteller and Wallace unedited, de-edited, and edited before beginning the statistical analysis of the texts — items such as quotations and numerals. Unfortunately, neither Williams and Mosteller nor Mosteller and Wallace published the contents of this "little book of decisions" and only mention five of their many decisions in the published work. [Mosteller and Wallace 7, 16, 38-41] The little book has been lost and cannot be recovered or even reconstructed [Mosteller]. This paper goes on to discuss the many ramifications of the "little book" on their study and the subsequent studies. Also, how the loss of the "little book" casts a shadow of "scientific invalidity" over the Mosteller and Wallace work — i.e., it cannot be replicated. Their "little book" was not used by any of the following studies — making meaningful comparisons moot.

### (2) Other Decisions

This section goes on to list many of the unediting, de-editing, and editing items that need to be considered. It lists several of the mistakes made by the many practitioners and what these mistakes mean to the validity of the studies (e.g.):

- (a) Wrong letters
- (b) Quotes — e.g., 131 words of Federalist 5 are a quote from Queen Ann, 334 words of Federalist 9 are a quote from Montesque
- (c) Footnotes — the author's and the editors'
- (d) Numbers
- (e) Foreign languages
- (f) Spelling
- (g) Homographic forms
- (h) Contracted forms

- (i) Hyphenation
- (j) Word determination
- (k) Disambiguation
- (l) Editorial intervention — internal (e.g., Hamilton on Madison) and external (e.g., from the first newspaper copy editor to present day editors)

## Conclusion

### (1) Acceptance of Results by Non-Traditional Practitioners

Are practitioners (statisticians and non-statisticians) so blinded by the statistical sophistication that the other elements of a valid non-traditional authorship study are ignored?

### (2) Acceptance of Results by History Scholars

Do professional historians accept, deny, or show indifference to the body of work that supports the Mosteller and Wallace study? Why did I spend hours searching for a Mosteller and LAWRENCE study of the Federalist papers?

### (3) Do the multiple flaws in all of these non-traditional studies invalidate the results.

Is the case put forth by Mosteller and Wallace and buttressed by the other non-traditional practitioners nothing but a "Monument" built on sand? What effect does showing the flaws in the Federalist studies have on non-traditional studies in general — i.e., if the best is suspect, what about the rest!

## Bibliography

Adair, Douglass. "The Authorship of the Disputed Federalist Papers." *The William and Mary Quarterly* 1.2 Part I and 1.3 Part II (1944): 97-122 and 235-264.

*Avalon Project*. Yale Law School. 97-122 Ind 235-264. Accessed 13 February 2004, 10:30AM. <<http://www.yale.edu/lawweb/avalon/>>

Bosch, Robert A., and Jason A. Smith. "Separating Hyperplanes and the Authorship of the Disputed Federalist Papers." *The American Mathematical Monthly* 105.7 (1998): 601-607.

Bourne, E.G. "The Authorship of the Federalist." *The American Historical Review* 2.3 (1897): 443-460.

Collins, Jeff, et al. "Detecting Collaborations in Text: Comparing the Authors' Rhetorical Language Choices in the Federalist Papers." *Computers and the Humanities* 38.1 (2004): 15-36.

*constitution.org*. Accessed 9-30-03. <<http://constitution.org/fed/feder00.htm>>

Cooke, Jacob E., ed. *The Federalist*. Cleveland: Meridian Books (The World Publishing Company), 1956.

Davis, George. "RE: Gutenberg edition of Federalist." Private E-mail, 20 November 2003 18:46:51.

Engeman, Thomas S., et al., ed. *The Federalist Concordance*. Middletown, Connecticut: Wesleyan University Press, 1980.

Farrington, Jill. *Analysing for Authorship*. Cardiff: The University of Wales Press, 1966.

Farrington, Michael G., and Andrew Q. Morton. "Fielding and the Federalist." *Department of Computing Science Research Report* (1990/R6).

Forsyth, Richard S. "Stylistic Structures: A Computational Approach to Text Classification." Diss. University of Nottingham, 1995.

Francis, Ivor S. "An Exposition of a Statistical Approach to the Federalist Dispute." *The Computer and Literary Style*. Ed. Jacob Leed. Kent Ohio: Kent State University Press, 1966. 38-78.

Fung, Glenn. "The Disputed Federalist Papers: SVM Feature Selection via Concave Minimization." *Proceedings of the 2003 Conference on Diversity in Computing*. Atlanta, Georgia, 2003. 42-46.

Fung, Glenn. *CS 635 Project*. Spring Semester 1999. Accessed 2004-11-09. <<http://www.cs.wisc.edu/~gfung/GSVMFP.ps>>

Fung, Glenn, and Olvi L. Mangasarian. "The Disputed Federalist Papers: SVM Feature Selection via Concave Minimization." Paper delivered at the CSNA 2002 Conference, Madison, Wisconsin. 15 June 2002.

Hamilton, Alexander, et al. Ed. Robert Scigliano. *The Federalist: A Commentary on the Constitution of the United States*. New York: The Modern Library (Random House), 2000.

Hart, Michael. "RE: Gutenberg edition of Federalist." Private E-mail, 21 November 2003 12:59:08.

Hilton, Michael L., and David I. Holmes. "An Assessment of Cumulative Sum Charts for Authorship Attribution." *Literary and Linguistic Computing* 8.2 (1993): 73-80.

Holmes, David I., and Richard S. Forsyth. "The Federalist Revisited: New Directions in Authorship Attribution." *Literary and Linguistic Computing* 10.2 (1995): 111-127.

- Khmelev, Dimitri V., and Fiona J. Tweedie. "Using Markov Chains for Identification of Writers." *Literary and Linguistic Computing* 16.3 (2001): 299-307.
- Kjell, Bradley. "Authorship Determination Using Letter Pair Frequency Features with Neural Network Classifiers." *Literary and Linguistic Computing* 9.2 (1994): 119-124.
- Kjell, Bradley, et al. "Discrimination of Authorship Using Visualization." *Information Processing & Management* 30.1 (1994): 141-150.
- Martindale, Colin, and Dean McKenzie. "On the Utility of Content Analysis in Author Attribution: The Federalist." *Computers and the Humanities* 29 (1995): 259-270.
- McColly, William, and Dennis Weier. "Literary Attribution and Likelihood-Ratio Tests: The Case of the Middle English Pearle-Poems." *Computers and the Humanities* 17 (1983): 65-75.
- Merriam, Thomas. "An Experiment with the Federalist Papers." *Computers and the Humanities* 23.3 (1989): 251-254.
- Mitchell, Ann F.S., and Clive D. Payne. "A Conservative Confidence Interval for a Likelihood Ratio." *Journal of the American Statistical Association* 66.336 (1971): 861-866.
- Mosteller, Frederick, and David L. Wallace. "Notes on an Authorship Problem." *Proceedings of a Harvard Symposium on Digital Computers and their Applications*. Cambridge, Massachusetts: Harvard University Press, 1962. 163-197.
- Mosteller, Frederick, and David L. Wallace. "Inference in an Authorship Problem. A Comparative Study of Discrimination Methods Applied to the Federalist Papers." *Journal of the American Statistical Association* 58 (1963): 275-309.
- Mosteller, Frederick, and David L. Wallace. *Applied Bayesian and Classical Inference: The Case of the Federalist Papers*. New York: Springer-Verlag, 1984.
- Pennebaker, James W. "The Federalist." Unpublished preliminary work.
- Pennebaker, James W. "[no title]." Private E-mail, Wednesday 09 July 2003, 14:45:34.
- Pennebaker, James W. "[no title]." Private E-mail, Wednesday 09 July 2003, 15:32:59.
- Piaia, Jesse. "[For Frederick Mosteller]." Private E-mail, Tuesday 22 July 2003, 10:57:38.
- Piaia, Jesse. "[For Frederick Mosteller]." Private E-mail, Tuesday 22 July 2003, 11:48:04.
- Project Gutenberg*. Accessed 2003-09-30. <<http://promo.net/pg/>>
- Rokeach, Milton, et al. "A Value Analysis of the Disputed Federalist Papers." *Journal of Personality and Social Psychology* 16.2 (1970): 245-250.
- Roland, Jon. "RE: The Federalist on constitution.org." Private E-mail, 11 September 2003, 10:24:36.
- Rudman, Joseph. "Unediting, De-Editing, and Editing in Nontraditional Authorship Attribution Studies: With an Emphasis on the Canon of Daniel Defoe." *Papers of the Bibliographical Society of America* 99:1 (March 2005).
- Sarndal, Carl-Erik. "On Deciding Cases of Disputed Authorship." *Applied Statistics* 16.3 (1967): 251-268.
- Stamatatos, E., N. Fakotakis, and G. Kokkinakis. "Text Genre Detection Using Common Word Frequencies." *COLING 2000: Proceedings of the 18th International Conference on Computational Linguistics*. 2000. 808-814.
- Stamatatos, E., N. Fakotakis, and G. Kokkinakis. "Computer-Based Authorship Attribution Without Lexical Measures." *Computers and the Humanities* 35 (2001): 193-214.
- Tankard, Jim. "The Literary Detective." *BYTE* 11.2 (1986): 231-238.
- Tweedie, Fiona J., S. Singh, and D.I. Holmes. "Neural Network Applications in Stylometry: The Federalist Papers." *Computers and the Humanities* 30.1 (1996): 1-10.
- Wachal, Robert Stanley. *Linguistic Evidence, Statistical Inference, and Disputed Authorship*. Dissertation, University of Wisconsin, 1966.
- Waugh, Sam, Anthony Adams, and Fiona Tweedie. "Computational Stylistics Using Artificial Neural Networks." *Literary and Linguistic Computing* 15.2 (2000): 187-197.
- Yang, Albert C.C., et al. "Information Categorization Approach to Literary Authorship Disputes." *PHYSICA A* ().

---

## Interface Design

---

**Stan Ruecker** (*sruecker@ualberta.ca*)

*University of Alberta*

**Stéfan Sinclair** (*sgsinclair@gmail.com*)

*McMaster University*

**Stephen Ramsay** (*sramsay@uga.edu*)

*University of Georgia*

**Milena Radzikowska** (*mradzikowska@mtroyal.ca*)

*Mount Royal College*

**Alan Galey** (*agaley@uwo.ca*)

*University of Western Ontario*

---

Unsworth suggests that humanities computing has progressed during the last thirty years through three phases, beginning with the design of text analysis tools, moving to the design of online digital collections, and now returning to the development of a new generation of tools for working with online materials. While many of the first generation of text analysis tools focused on the analysis of a single text, the emphasis in many of the new tools is on data mining and visualization of results across sets of documents or even entire collections.

There are several distinct research agendas involved, including interest in making digital materials more accessible (Lyman), in providing new affordances for doing synthetic research at the level of the interface (Ruecker), and in using visualizations of digital information not only for analysis but also for further access to subsequent information (Flanders).

This session brings together three papers on visualization and interface design. First is Ruecker, Sinclair, and Radzikowska's "The Aesthetic Function: The role of visual communication design in interface research", which examines the role of graphic design in interfaces to digital collections and visualization interface research. Their conclusion is that

Careful attention to the details of graphic presentation can have a significant impact on the perceived value of a digital collection, the function of a visualization system, the research results available from analysis of visualizations, and the dissemination of findings both within the academic community and for the larger public audience.

Second is Alan Galey's "'Alms for Oblivion': Bringing an Electronic New Variorum Shakespeare to the Screen". Galey emphasizes the importance of W3C advanced standards

compliance for the delivery of academic collections online. Interfaces to collections like the eNVS call for digital adaptations that can make traditional scholarly apparatus more manageable:

The value of the eNVS interface lies in reinventing such fundamental scholarly mechanisms as the textual collation line, the commentary footnote, and the annotated page — three structures from which the variorum derives both its archival power and, for many print users, its aura of cognitive overload.

The third paper is Ramsay's "Mining Shakespeare", which discusses the results and the implications of using the *StageGraph* and *D2K* software systems for semi-automatically mapping scene changes in Shakespeare's plays. Ramsay comments not only on the results of this study, but also on the implications of data mining in humanities scholarship:

Though it may be used to support or refute hypotheses, data mining is far more useful in the service of the broad humanistic mandate to find new and insightful ways of looking at textual artifacts.

These presentations provide a cross-section of current research on visualization in the humanities, moving, as does the research itself, between intricate detail and broad theoretical principles.

### Bibliography

Flanders, Julia. "Text analysis and the problem of pedantry." Paper delivered at CaSTA 2004: The Face of Text. 3rd conference of the Canadian Symposium on Text Analysis, McMaster University, Hamilton, Ontario. November 19-21 2004. 2004.

Lyman, Eugene W. "In pursuit of radiance: Report on an interface developed for the Piers Plowman Electronic Archive." Paper delivered at CaSTA 2004: The Face of Text. 3rd conference of the Canadian Symposium on Text Analysis, McMaster University, Hamilton, Ontario. November 19-21 2004. 2004.

Ruecker, Stan. *Affordances of Prospect for Academic Users of Interpretively-tagged Text Collections*. PhD. Dissertation, Edmonton: University of Alberta, 2003.

Unsworth, John. "Forms of Attention: Digital Humanities Beyond Representation." Paper delivered at CaSTA 2004: The Face of Text. 3rd conference of the Canadian Symposium on Text Analysis, McMaster University, Hamilton, Ontario. November 19-21 2004. 2004.

## The Aesthetic Function: The role of visual communication design in interface research

Stan Ruecker, Stéfan Sinclair, Milena Radzikowska

Since the professionalization of various humanities disciplines in the latter part of the nineteenth century, humanities scholars have been primarily occupied with the interpretation and analysis of existing cultural artifacts, such as texts. The expertise and artistry required to produce the material form of the objects are generally outside the scope of a humanities education; print-making, for instance, is a separate craft. However, since the rise of personal computing and graphical interfaces, many humanities scholars have been empowered to create the interface through which their materials can be studied: the proliferation of digital collections such as *Perseus* and *Rossetti* bears witness to this phenomenon. The distinctions between author, critic, editor and publisher have blurred. Significantly though, the knowledge and perspective of artists and designers have either largely been ignored by the digital humanities scholar, or else have contributed in a manner that has not been subject to direct analysis. Similarly, the growing interest in visualization systems for the humanities is another research area where design issues are relevant (Bradley and Rockwell). The significance of the visual is sufficiently evident in all of these cases that aesthetic factors become intrinsically woven with issues of functionality. Research interests in graphic design and presentation find a new relevance and weight, not only as a contributing factor in the design of computer interfaces and visualization systems, but also as an area of study in their own right.

We address the issue of graphic design contributions to visualization research by making reference to several recent interface design research projects at the University of Alberta and McMaster University. The emphasis is on the functional differences between early and later prototypes, including an analysis of what Frascara calls "the aesthetic function of design". We argue that aesthetic function is a composite that includes attracting viewers, holding their attention, and compelling their trust and respect. Design, in other words, is of utmost importance to the value and legitimacy of scholarly digital content.

Our first example is from the *TouchGraph* representation of XML data, where initial designs included boxes around each of the individual text items. Although this form of display is commonly used with topic maps, it is also unreasonable for several reasons: it draws the reader's eye to locations that are not particularly meaningful; it introduces unnecessary clutter; and it misappropriates a grouping affordance for a single element, which does not require grouping.

A second example is a prototype system for blocking and reading plays, called *Watching the Script* (Ruecker et al.). The

interface design has gone through three distinct stages, beginning with a white parallelogram, moving to a four-colour square, and ending with a very attractive full-colour combination of stage, playback controls, and large coloured dots that clearly indicate character positions. Attention to the details of the graphic design is in this case intrinsically related to the details of the functions of the system, such as the location and movement of characters and text. However, the list of additional qualities would not be complete without acknowledging that part of the attraction of the most recent iteration is the aesthetic function. In its extreme form, this value can result in forms of interface that are in some senses autotelic — they can become an end in themselves for some users, who find their attractions sufficient to make the system worth further attention, outside the context of any particular research task.

The connection between graphic design and academic research also has implications for the ongoing need for improved communication between the academic world and everyone else. Several strategies are required at different levels, including public information campaigns, academic contributions to popular media, and a more significant presence of the academic in the community. One potential role that design has to play is in visually rewarding the reader of research results. However, there are barriers to be overcome, not least of all within the academy. It might even be argued that there is an anti-aesthetic subtext in certain research areas, since effort to engage readers through visual appeal (and its related functionality) might be understood as devaluing more essential research outcomes. However, Pujol points out that the visual qualities of professional design are one of the key signifiers by which we distinguish the individual voice from the institutional. If someone hand letters a sign to advertise a garage sale, we understand the sale as an amateur activity. If that same person employs graphic design skills and produces a glossy poster, we may interpret the same event, at least until we arrive at the site, as the establishment of a new retailer.

Karvonen takes this line of reasoning even further in her study of the relationship between trust and design. The cohort for her project was Scandinavian, with participants from both Finland and Sweden. The long-standing cultural awareness of design quality in those countries is probably a factor in her findings that people tended to find that web sites with a clearly professional design quality were rated as being more trustworthy than more vernacular sites. It would be indefensible to suggest that a professional standard of visual communication design could contribute to the perceived reliability of research results, since there are other, more important indicators that are applicable. However, it is not outside the domain of the possible that graphic quality is potentially a contributing factor not only in the evaluation of research results, but, particularly in the areas of visualization and information design research, also in the results obtained from user study. Careful attention to the

details of graphic presentation can have a significant impact on the perceived value of a digital collection, the function of a visualization system, the research results available from analysis of visualizations, and the dissemination of findings both within the academic community and for the larger public audience.

## Bibliography

Bradley, John, and Geoffrey Rockwell. "What scientific visualization teaches us about text analysis." *Consensus ex machina? ALLC-ACH 94 abstracts*. Paris, 1994. 203-204.

Frascara, Jorge. *User-centred Graphic Design*. London: Taylor and Francis, 1997.

Karvonen, Kristiina. "The beauty of simplicity." *Proceedings on the 2000 conference on Universal Usability*. November 2000. 99-99.

Pujol, Monica. "Design as a social practice." Public lecture. Trans. Jorge Frascara. Department of Art and Design, University of Alberta, 1 Nov 2001.

Ruecker, Stan, Eric Homich, and Stéfan Sinclair. "Watching the Script of Synge's Playboy of the Western World." Paper delivered at the COCH/COSH. Congress 2004. Winnipeg: University of Manitoba. 2004.

Unsworth, John. "Forms of Attention: Digital Humanities Beyond Representation." Paper delivered at CaSTA 2004: The Face of Text. 3rd conference of the Canadian Symposium on Text Analysis, McMaster University, Hamilton, Ontario. November 19-21 2004. 2004.

## "Alms for Oblivion": Bringing an Electronic New Variorum Shakespeare to the Screen

Alan Galey

In this paper, I describe the development of a digital interface for the *Electronic New Variorum Shakespeare (eNVS)*, and explore certain historical and technical issues that bear upon our design strategies. With the considerable burdens of content development and encoding resting on others' shoulders, *NVS* co-general editor Paul Werstine and I have been free to focus exclusively on interface design – an area of humanities computing that, I would suggest, has not kept pace with advancements in web browsers and third-party design standards. Although many computing humanists might see an *eNVS* edition as a document, however complex, we have instead approached it as a data object, where the organizing logic is that of object-oriented programming, not hypertext. The value of the *eNVS* interface lies in reinventing such fundamental scholarly mechanisms as the textual collation line, the commentary footnote, and the annotated page – three structures from which the variorum derives both its archival power and, for many

print users, its aura of cognitive overload. With these issues in mind, I will argue that in order to bring electronic editing projects like the *eNVS* to the screen, humanists must also be information architects, who think past documents to embrace the principles of object-oriented and standards-compliant programming and design. Conversely, the historical section of this paper will show that the programmers must also be humanists, who understand the cultural and bibliographical histories of the interface traditions in which they work.

Scholarly opinion differs on the present value and future viability of Shakespeare variorum editions, print and electronic, but tends to agree that, in any form, they rank among the largest information-management projects in Shakespeare scholarship. On one side of the debate, Richard Knowles goes so far as to call variorums "the memories of the profession" (43), though he also stresses that variorums, like all memory, incorporate a principle of selection in their management of heterogeneous masses of data. Maurice Hunt takes issue with the traditional perception of the variorum edition as "the still point in the turning world of texts, a text which would arrest, and even reverse, the processes of textual change and corruption" (62, quoting McGann 93) – a view that consigns these editions to the "tombs/tomes" of an "obsolete modernism" (Hunt 62). Instead, Hunt contends, the variorum structure anticipates postmodernist values in the heterogeneity of its apparatus, which conveys the indeterminacy of the Shakespeare text more than any other kind of edition. Yet for some scholars on the other side, projects like the *NVS* are more about the past than the future, amounting to "admission[s] of failure" and monuments to unachieved textual stability (Rhodes and Sawday 11; Bristol 101). In one instance of pointed criticism, John Lavagnino claims that the uncategorized nature of variorum commentary renders it un-digital in advance, and "not productively open to flexibility of display" (201). He concludes with a call for improvements in display technology (203), which our project echoes and in part hopes to answer. As this range of thought indicates, the challenges facing an electronic variorum are not purely technical, and require a level of interface design that accounts for the historical issues at stake.

This paper has four sections:

1. History: the variorum interface in print
2. Possible futures: the web browser as design platform
3. Examples of the *eNVS* interface
  - (a) Textual apparatus
  - (b) Page/screen layout(s)
  - (c) Annotations
4. Conclusion: alms for oblivion

## 1. History: the variorum interface in print

This section will provide a brief outline of the design challenges we have inherited from eighteenth-century editors. Interface issues have dominated the variorum's historical role in Shakespeare studies since Samuel Johnson first applied the format to Shakespeare in 1765. Part of our research mandate is to reinvent the complex layout of the Shakespeare variorum, which has remained largely unchanged – and unloved, many Shakespeareans would say – for over two centuries. As recent scholarship on editing's cultural history shows, the reservations expressed by Bristol and Lavagnino are as old as the Shakespeare variorum itself (see DeGrazia 209-14, and Gondris). I will confine my focus here to the historical problem of too much (Shakespearean) information, which casts a long shadow over any interface design.

## 2. Possible futures: the web browser as design platform

The *eNVS* is an interface with a 200-year history, and with an eye to the present moment of standards-friendly design in the wake of the so-called browser wars between Microsoft and Netscape. If Bristol is correct that the *NVS*'s goals exceed the limits of print (101), and if Lavagnino is correct that complex digital commentary is insufficiently served by hypertext alone (198-200), we might conclude that an adequate *eNVS* interface demands advancement beyond traditional, HTML-era design. I will briefly summarize the case for advanced browser-based interfaces, with particular reference to the W3C's standardization of key web technologies such as CSS, XML, and the DOM – and, most importantly, the implementation of these and other standards in 'postwar' open-source browsers such as Mozilla and Firefox.

## 3. Examples from the *eNVS* interface:

### 3.a Textual apparatus

Known by such tongue-in-cheek epithets as the "band of terror" or "barbed wire" that runs beneath the text (Thomas Berger and Edmund Wilson, quoted in Rasmussen 211), the traditional collation of variants offers the most obvious candidate for a digital reconception. Where many electronic editions at best display variants by way of linked parallel texts (swapping one print interface for another), and at worst simply recode the collation line as a text string, the *eNVS* interface instead generates a properly machine-readable apparatus by means of object-oriented scripting. This allows us to expand the collation line, textually and graphically, into the textual history it compresses and encodes.

### 3.b Page/screen layout(s)

As Gondris has shown, the Shakespeare variorum page inherited from the eighteenth century constitutes a critical structure that promotes some habits of thought and suppresses others. The consequences of rearranging it will therefore reach beyond readability and convenience – important enough issues in themselves – to impact the production of meaning. This is one of the most challenging aspects of the *eNVS*, not least because of the computer screen's orientation toward vertical scrolling. Again, an object-oriented interface enables multiple layout options without generating redundant files.

### 3.c Annotations

The question of how best to display electronic annotation remains a central debate in electronic editing. It is also a central concern in our project, since the variorum's primary content is not its playtext, but its notes. But as Lavagnino has pointed out, it is difficult for digital interfaces to improve upon – or even match – the cognitive elegance of the print reader's glance from text to footnote (198-9). This section will demonstrate how our note design works with the page layouts, and with the complex archive formed by the network of *NVS* annotations.

## 4. Conclusion: alms for oblivion

The paper will conclude with a restatement of the argument for closer integration of textual studies and web programming in the practice of electronic editing, especially in projects like the *eNVS*. Much of the energy that might advance interface design in humanities computing is presently devoted to digitization and tagging, in response to the archival impulse still strong in the humanities. As an invitation to discussion, I will close by reflecting on Knowles's quotation of Shakespeare's *Troilus and Cressida*, which he uses to make the point that all scholarship risks becoming "alms for oblivion, ... good deeds past, which are devoured / As fast as they are made, forgot as soon as done" (3.3.141-4, quoted in Knowles 39). Shakespeare variorums are akin to all electronic preservation formats in that they attempt, in Knowles's words, "to guard against oblivion," even as they are subject to it. The *eNVS* seeks to preserve scholarship into the future by increasing its accessibility and relevance in the present.

## Bibliography

Bristol, Michael D. *Shakespeare's America, America's Shakespeare*. London: Routledge, 1990.

De Grazia, Margreta. *Shakespeare Verbatim: The Reproduction of Authenticity and the 1790 Apparatus*. Oxford: Clarendon P, 1991.



Gondris, Joanna. "All this Farrago': The Eighteenth-Century Shakespeare Variorum Page as a Critical Structure." *Reading Readings: Essays on Shakespeare Editing in the Eighteenth Century*. Ed. Joanna Gondris. Madison, WI: Fairleigh Dickinson UP, 1998. 123-39.

Hunt, Maurice. "New Variorum Shakespeares in the Twenty-First Century." *Yearbook of English Studies* 29 (1999): 57-68.

Knowles, Richard. "Variorum Commentary." *TEXT* 6 (1994): 35-47.

Lavagnino, John. "Two Varieties of Digital Commentary." *Textual Performances: The Modern Reproduction of Shakespeare's Drama*. Ed. Lukas Erne and Margaret Jane Kidnie. Cambridge: Cambridge UP, 2004. 194-209.

McGann, Jerome J. *A Critique of Modern Textual Criticism*. Charlottesville, VA: UP of Virginia, 1983.

Rasmussen, Eric. "Richly Noted: A Case for Collation Inflation." *Arden: Editing Shakespeare*. Ed. Ann Thompson and Gordon McMullan. London: Arden Shakespeare-Thompson Learning, 2003. 211-8.

Rhodes, Neil, and Jonathan Sawday. "Paperworlds: Imagining the Renaissance Computer." *The Renaissance Computer: Knowledge Technology in the First Age of Print*. Ed. Neil Rhodes and Jonathan Sawday. New York: Routledge, 2000. 1-17.

## Mining Shakespeare

### Stephen Ramsay

"[T]he computer," writes Susan Hockey in her 2000 book *Electronic Texts in the Humanities*, "is best at finding features or patterns within a literary work and counting occurrences of those features" (Hockey 66). For many areas of inquiry, such finding and counting is eminently useful. Word-frequency analysis in the context of computational linguistics, concordance generation as a prelude to the study of word usage, and the various search functions that have now become an ordinary part of the task of research in so many disciplines represent clear examples of the utility of computational tools. For scholars engaged in the task of literary critical interpretation, however, such finding and counting can seem beside the point. As Hugh Craig put it:

The leap from frequencies to meanings must always be a risky one — Lower-level features are easy to count but impossible to interpret in terms of style; counting images or other high-level structures brings problems of excessive intervention at the categorization stage, and thus unreliability and circularity.

(Craig 103)

The risk of which Craig speaks is not merely a matter of interpretive caution. In most cases, low-level features simply

don't assert themselves in any obvious way into the broad, complex patterns upon which literary critical interpretation depends.

Data mining provides a suggestive set of methods for bridging the gulf between low and high. It has its roots in a number of statistical techniques with venerable histories in digital humanities (in particular, the use of factor analysis in the study of literary and philosophical texts)<sup>1</sup>, but introduces an exploratory dimension far more conformable to the elaborate task of prompting meaningful critical insight. Data mining techniques operate on low-level features, but use a variety of statistical and logic-programming methods to discern broad complex patterns in the data set (such as classifications, categorizations, and prediction models) that are not conceived in advance. In other words, data mining lets us ask what's interesting about an apparently disparate set of low-level features without having to form any concrete expectations in advance.

This paper presents research on the structure of Shakespearean drama and its relation to genre categorization using two programs: *StageGraph* and *D2K*. *StageGraph* generates directed graph visualizations of scene changes and character movements from XML representations of plays (figure 1). It can also use the generated graphs to produce matrices of individual graph properties (e.g. degree number, number of cycles, diameter, chromatic number). While such graphs and matrices provide fertile ground for interpretive reflection, the utility of the graphs is greatly enhanced when the generated properties are themselves 'mined' for broad patterns and features, and the results presented in the form of an open-ended visualization.

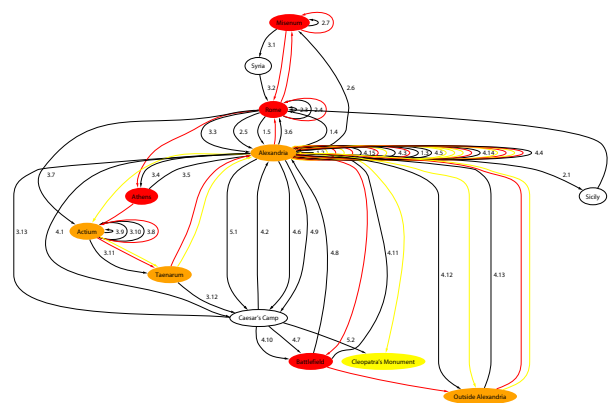


Figure 1

The author and a research assistant at the National Center for Computing Application used the *D2K* software to conduct naive Bayesian analysis and decision tree generation (two standard data mining techniques) on the graph property matrices to try to see if the low-level structural features of Shakespeare's plays assemble themselves into clusters that correspond to the traditional genre categories of comedy, tragedy, history, and

romance. We then used a technique pioneered by one of the project participants that combines concept-tree clustering with shaded similarity matrices to generate a visualization of the degrees of similarity among Shakespeare's plays in terms of genre (figure 2). The results are extremely suggestive, in that the visualization not only groups the plays broadly into traditional generic categories, but also contains anomalies that correspond to some of the more influential insights into Shakespearean genre (for example, Susan Snyder's argument that *Othello* possesses the basic structure of comedy appears to be confirmed by the data mining algorithm).

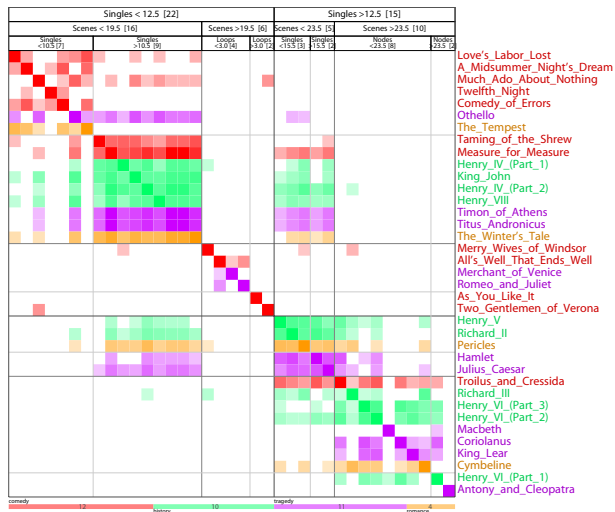


Figure 2

Most interesting of all, however, are the anomalies that represent neither traditional classification nor the product of exegetical insight. For it is here, I believe, that data mining and visualization present the most promise for text analysis practitioners in literary study. Though it may be used to support or refute hypotheses, data mining is far more useful in the service of the broad humanistic mandate to find new and insightful ways of looking at textual artifacts.

## Bibliography

- Bradley, John, and Geoffrey Rockwell. "Watching Scepticism: Computer Assisted Visualization and Hume's Dialogues." *Research in Humanities Computing* 5 (1996): 32-47.
- Burrows, J.F., and D.H. Craig. "Lyrical Drama and the 'Turbid Montebanks': Styles of Dialogue in Romantic and Renaissance Tragedy." *Computers and the Humanities* 28 (1984): 63-86.
- Craig, Hugh. "Authorial Attribution and Computational Stylistics: If You Can Tell Authors Apart, Have You Learned Anything About Them?" *Literary and Linguistic Computing* 14 (1999): 103-13.

Hockey, Susan. *Electronic Texts in the Humanities*. Oxford: Oxford University Press, 2000.

Snyder, Susan. *The Comic Matrix of Shakespeare's Tragedies: Romeo and Juliet, Hamlet, Othello, and King Lear*. Princeton: Princeton University Press, 1979.

Wang, Jun, Bei Yu, and Les Gasser. "Concept Tree Based Clustering Visualization with Shaded Similarity Matrices." *Proceedings of the 2002 IEEE International Conference on Data Mining (2002)*. 2002. 697-700.

- cf. J. F. Burrows and D. H. Craig's studies of Romantic and Renaissance tragedy and John Bradley and Geoffrey Rockwell's use of cluster analysis for the study of Hume's *Dialogues*, op cit.

## Academic Libraries and Information Communities: New Models for Supporting Digital Scholarship

*Christine Ruotolo (ruotolo@virginia.edu)*

*University of Virginia*

Since the first forays into digital library activity in the early 1990s, academic libraries have been intimately involved in the application of technology to the teaching and research missions of the universities they support. Libraries offer increasingly sophisticated technology training to faculty, and provide critical systems infrastructure and programming expertise to support digital scholarship. As more and more faculty wish to produce born-digital scholarship, they expect libraries to supply the basic technology support services that their academic or campus computing units are often unable to provide. And libraries have rushed bravely into the breach, becoming ad hoc publishers, software developers, and instructional designers. Partly as a result of this electronic outreach activity, libraries are accumulating a critical mass of digital materials not governed by any explicit selection policy. This trend threatens to accelerate, as grant-funded humanities computing projects of all shapes and sizes are turning to libraries to aggregate, disseminate and preserve their materials over the long term. Libraries supporting digital scholarship must therefore manage both technological development and collection development in a way that serves scholars' needs and maximizes the effectiveness of limited resources.

The University of Virginia Library, like many of its peer institutions, is currently developing a large institutional repository for the digital content it acquires, along with tools for ingesting and disseminating that content. But by its very nature, our repository will have rather rigid requirements with regard to standards and format. Even when the repository development is complete, our faculty and their collaborators will continue to produce scholarly ephemera - born-digital writings, archival materials, teaching resources, and experimental tools - that fall outside the repository's collections parameters but which the Library will nonetheless be expected to help sustain and organize in some way. At the same time, the Library will need to encourage the use of its digital resources beyond the core group of early adopters by making it easy for users to find and exploit relevant materials in their subject area.

In order to meet these challenges, the UVa Library has developed an "information communities" model to complement its digital repository initiatives. We hope this model will allow us to identify common needs, establish priorities, and minimize redundant digitization and tool-building efforts. Ideally, a robust information community will foster scholarly communication in all its diverse forms; it will encourage innovation and spark new areas of cross-disciplinary and cross-institutional research.

At its core, the information community involves an equilateral relationship between people, collections, and tools. The people are scholars, students, researchers, information professionals, and (for public institutions) the community at large. These people serve many roles - they are producers and consumers of digital content; they are authors, editors, commentators, selectors, and publishers. The collections are conceived as broadly as possible - primary and secondary scholarship, both formal and informal, across all media types: text, images, maps, datasets, audio and video. The tools can be divided into several sub-categories: communication tools, like rosters, mailing lists, wikis, and virtual conferencing tools; analytic and interpretive tools, such as software for textual collation, or for the manipulation of statistical data; authoring tools that support standard markup schemes; and hybrid tools, such as collaborative editing tools or peer review tools that combine the communication and authoring functions. This triangular relationship facilitates sharing - sharing scholarly materials and sharing tools for accessing and analyzing those materials. The community can foster the formal or informal scholarly exchange of ideas, in the form of new publication as well as conferences, seminars and online discussion.

The University of Virginia has several active information community prototypes. Two are of particular interest. *The Tibetan and Himalayan Digital Library (THDL)*, founded by David Germano in the Religious Studies department, developed organically from the need to allow a small but globally dispersed group of Tibetan experts to collaborate effectively, in order to build a comprehensive online archive of Tibetan materials. The site includes materials across all media types - literary and religious texts, dictionaries, image sets, gazetteers, maps, statistical datasets, timelines and time-based media. The tools necessary to support this community are highly specialized and include Tibetan, Nepali, and Chinese fonts, along with the input tools necessary to generate materials in these languages, geo-referencing and interactive mapping tools, collaborative editing tools, and tools for the multilingual transcription of audio and video materials.

*The American Studies Information Community*, a two-year grant-funded pilot project, sought to capitalize on a substantial but unintegrated body of American Studies-related digital materials produced by the Library, its electronic centers, and affiliated research units. Federated search and browse interfaces

were developed to create easy access to these widely dispersed materials. New core collections centering on a locally relevant topic (the Lewis and Clark expedition) were digitized, and faculty used these in class both as teaching resources and as seed materials for digital student projects. The community also published a collaboratively edited database of electronic resources in the field, which will help users locate materials for research and teaching but will also identify gaps in the field that might indicate opportunities for future digital projects.

Based on our experiences at Virginia, we've identified some key factors contributing to the success of an information community or similar initiative. The community needs at least one dedicated, charismatic faculty leader who can organize and motivate colleagues across institutional boundaries; this person should ideally be at a career point where he or she can take chances on speculative projects that might fail. The community should be bold and seek out innovative, ambitious ideas, despite the current budget-conscious institutional inclination toward caution. An information community is most effective when centered around "keystone" collections of value to the scholarly community, and should ideally assemble a body of primary materials large enough to support meaningful analytical study. A scholarly community needs a sense of purpose and a clear set of organizational principles based upon its purpose. It needs a stable institutional home, with sustainable funding. Finally, it needs to consider all of its potential constituencies and try to anticipate their particular needs.

## Bibliography

*The American Studies Information Community*. University of Virginia Library. Accessed 2005-04-04. <<http://infocomm.lib.virginia.edu/amst/>>

*The Tibetan and Himalayan Digital Library*. University of Virginia Library. Accessed 2005-04-04. <<http://iris.lib.virginia.edu/tibet/>>

---

## Modeling Diachrony in Dictionaries

---

*Susanne Salmon-Alt* ([salt@atilf.fr](mailto:salt@atilf.fr))

ATILF-CNRS

*Laurent Romary* ([romary@loria.fr](mailto:romary@loria.fr))

Loria-Inria

*Buchi Eva* ([eva.buchi@atilf.fr](mailto:eva.buchi@atilf.fr))

ATILF-CNRS

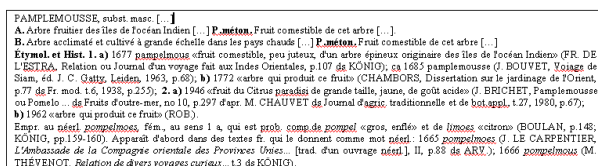
---

## Introduction: The variety of lexical structures

Lexical data appear in a wide variety of forms. These can range from basic morpho-syntactic structures (Romary et al.) intended to be used in language engineering applications to important editorial projects that cover multiple levels of lexicographic description: morphological information, syntactic constructs, sense related information (definitions, examples, usage notes, etc.) or historical information. Entries can also vary in their internal organization. Among other factors, the fundamental choice between an onomasiological (concept to word) and a semasiological (word to concept) representation directly impacts on the internal structure of entries, as well as on the possible choice of descriptors attached to them. From a computational point of view, this situation prevents the design of one single data structure that would fit all the possible needs, whereas one would like to be able to have uniform access to similar information across heterogeneous lexical sources. This has been the source of strong debates, leading for instance to the ubiquitous "Print Dictionaries" chapter of the *TEI (Text Encoding Initiative)* that tries to combine structured and unstructured views of lexical entries. Still, we want to show in this paper that it is possible to apply coherent modeling principles to deal with this variety of structures while providing a precise account of complex sub-components such as diachronic information as they appear in dictionaries with wide lexical coverage. Besides, we want to show that such modeling principles can guide the possible evolution of the TEI towards a more flexible data for the concrete representation of dictionaries.

## Diachronic information in dictionary entries

We consider diachronic information along the lines of its modern, large acceptance as "a word's biography" (Baldinger). As such, it covers both etymological information in a restricted sense — tracing out origin and primitive significance of a lexeme in its source language — and historical notes about successive changes of form and meaning once it entered into the target language. This type of information can for instance be found in the *Oxford English Dictionary* (*OED*), in the *Deutsches Wörterbuch* (*DWB*) or in the *Trésor de la Langue Française* (*TLF*), for which Figure 1 illustrates the organization of diachronic information within the micro-structure of a lexical entry (*pamplemousse*): here, the Etymol. et Hist. section, separated from the synchronic description of the lexeme, consists of two parts, the first one being dedicated to the lexeme's history within the target language (modern French) and the second one to etymology proper, i.e. origin and word sense in the source language (Dutch).



PAMPLEMOUSSE, subst. masc. [...] **A.** Arbre fruitier des îles de l'océan Indien [...] **P.p.méton.** Fruit comestible de cet arbre [...] **B.** Arbre acclimaté et cultivé à grande échelle dans les pays chauds [...] **P.p.méton.** Fruit comestible de cet arbre [...] **Etymol. et Hist.** 1. a) 1677 pamplemousse <fruit comestible, peu juteux, d'une autre espèce: originaire des îles de l'océan Indien (FR. DE L'ESTRA, Relation ou Journal d'un voyage fait aux Indes Orientales, p.107 de KONIG); ca 1685 pamplemousse (J. BOUVEY, Voyage de Siam, éd. J. C. Gally, Leiden, 1963, p.68), b) 1772 <arbre qui produit ce fruit> (CHAMBORS, Dissertation sur le jardinage de l'Orient, p.77 de Fr. mod. 16, 1938, p.255), 2. a) 1946 <fruit du Citrus paradisi de grande taille, jaune, de goût acide> (J. BRICHEL, Pamplemousse ou Pomelo... de Fruits d'outre-mer, no 19, p.297 d'apr. M. CHAUVET de Journal d'agric. traditionnelle et de bot. appl., t.27, 1980, p.87), b) 1962 <arbre qui produit ce fruit> (ROB). Empr. au néerl. pampelmoes, fém., au sens 1 a, qui est prob. comp. de pompel <gros, entée> et de limoes <citron> (BOULAN, p.148; KONIG, pp.159-160). Apparaît d'abord dans des textes fr. qui le donnent comme mot néerl. 1665 pampelmoes (J. LE CARPENTIER, L'ambassade de la Compagnie orientale des Provinces Unies... trad. d'un ouvrage néerl.), II, p.88 de AKL.); 1666 pampelmous (M. THÉVENOT, Relation de divers voyages curieux... t.3 de KONIG).

Figure 1

## Historical Notes

The main objective of the historical notes is to provide (earliest) written testimony for each of the senses — and possibly different usages of a sense — with respect to the synchronic description of the entry. Therefore, temporal information and quoted source text associated with bibliographical references play a central role in this section. Whereas the *OED* and the *DWB* realise the projection from synchronic sense organization explicitly by subordinating historical notes under sense description, the diachronic part of the *TLF* takes up each of the four synchronic senses by a sense identifier, a date, a quotation and bibliographical reference. The latter might be complex in case of use of secondary literature. One may also notice that differences in word spelling led to two testimonies for sense 1a. Despite the very strict application of the sense projection principle — which is far from being applied systematically throughout the dictionary, as mentioned for example in Hausmann et al. — one may however notice that the synchronization has not been made

explicit, for example through the use of the same sense identifiers within the synchronic and diachronic sections.

## Etymological Notes

The etymology section of dictionaries is concerned with the origin and development of the lexeme before entering into the target language. As a central task, it informs about one or more etymons and determines the etymological class (inheritance, loan word, word generation) for the oldest sense of the lexeme under consideration. As a consequence, it is not directly related to individual senses in the modern stage of the considered language. In the example, the etymon for the oldest sense of *pamplemousse* (1a), is the Dutch *pampelmoes*, itself being a word generated via composition from *pompel* and *limoes*. Although there have been attempts to formalize further etymological notes (cf. etymological formulas, Ross), they are generally not subject to well defined organisation principles, at least in current dictionaries. Additionally to core information about etymon, etymological notes may indicate bibliographical sources of the etymological hypotheses and discuss other related issues (phonetic evolution, concurrent hypotheses, confidence statements, secondary etymons, testimony of etymons, intermediate states etc.).

## A representational model for diachronic information

In the following sections, we apply the main modeling principles of the LMF (Lexical Markup Framework) project within ISO committee TC 37/SC 4 to outline the structure of diachronic information in dictionary entries. Those principles (Ide & Romary) allow one to combine a meta-model, which informs the main agreed upon practices within a given field, with data categories, corresponding to elementary information units attached to the nodes of the metamodel. In the case of lexical structures, a metamodel is itself the combination of a core metamodel (a simple structure organizing a lexical entry with form related information and a hierarchy of senses) and lexical extensions, seen as additional modules attached to the core meta-model. In our case, we will consider what kind of lexical extensions are needed for both etymological and historical information.

## A Lexical Extension for Etymological Structure

We propose a basic lexical extension for etymological notes (*Etymology*, cf. Figure 2), i.e. a structure that

accounts for the description of links to etymons. The *Etymology* component may occur at most once for a given lexical entry, under the assumption that lexical entries are purely polysemous, excluding homonyms, given that the difference between both is made upon historical criteria, cf. *adresse*<sup>1</sup> (*TLF*). This etymological information is further structured by means of *Etymological Unit* and *Etymological Link* components. *Etymological Unit* components are word forms playing the role of etymons. As such, they might be characterized by any existing data category defined for the description of lexical entries, i.e. lemmata and inflected forms (*language*, *orthography*, *sense*, *part-of-speech*, *inflectional information* etc.). Two points have to be noticed: first, the coverage of *language* should be extended to more fine-grained geographical and diachronic variants as those currently available from the ISO 639 series. Second, depending on available resources, all or part of this information could be recovered by a pointing mechanism. *Etymological Link* components stand for the etymological relation between linguistic units. A link is basically characterized by an *etymological target* and an *etymological source*, i.e. pointers to external resources, including lexical entries of the current dictionary and etymological units previously described. Etymological links are typed by the *etymological class* (loan word, inheritance etc.). They may additionally bear information about the bibliographical source, confidence level or other type of notes. The full paper will show how this data structure accounts for different types of etymological notes in current dictionaries, including cases of concurrent, popular, secondary and multiple etymons.

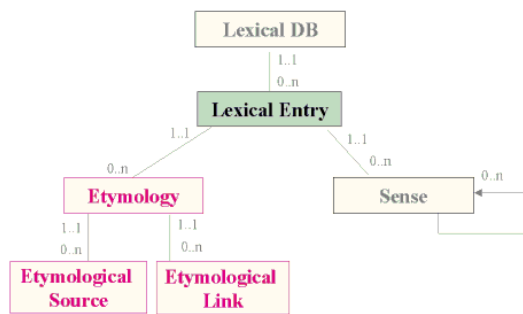


Figure 2

## A Lexical Extension for Historical Notes

The modeling of historical notes can actually be seen from two complementary and somehow sequentially organized perspectives. Firstly, we have identified that historical notes

are organized as a hierarchy of sense like objects, which leads to the simple historical extension depicted in Figure 3.

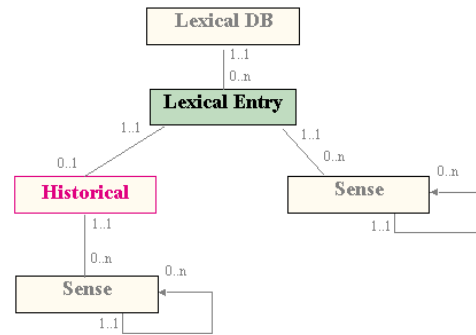


Figure 3

This extension takes up the sense component that already exists in the core LMF meta-model, while further characterizing it with specific dating (/date/) and bibliographic (/bibliography/) information. Such an extension accounts for the situations where there is no a priori editorial coherence between the sense organization in the lexical entry and its possible counterparts in the historical notes as encountered in, e.g., the *TLFi* or the *OED*. In that case, we can see that we keep open the possibility to actuate links (/synchronic reference/) between components of the historical notes and senses in the main entry. If we want to model a more controlled editorial project, we suggest to move from the previous extension to an integrated view (cf. Figure 4), which directly anchors historical descriptions on the corresponding senses. Doing so, it would always be possible to externalize the corresponding information, to derive an autonomous representation conformant to Figure 3.

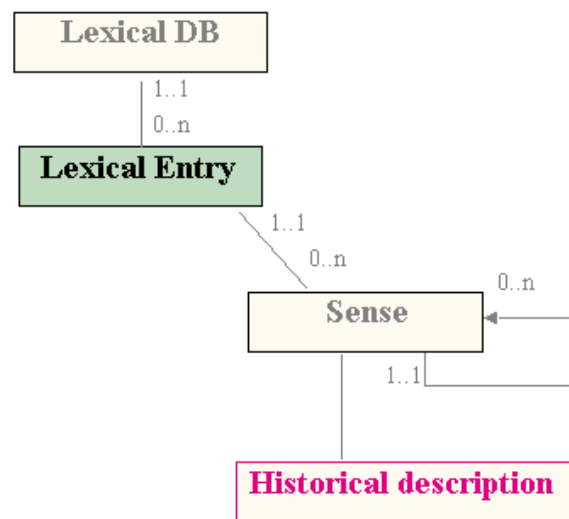


Figure 4

## Implementation in the framework of the TEI

The final paper will show precisely how the two types of structures described above can be implemented using the latest version of the specification platform of the TEI (ODD — One Document Does it all; Burnard & Rahtz). In particular, we will show that, on the one hand, we can extend the scope of the existing <etym> element from the P4 guidelines, and, on the other hand, it is necessary to introduce a new element dedicated to the representation of historical notes, which mimics the behavior of related entries (sub-structure with a strong structural analogy to a full entry), combined with dating and bibliographical descriptors. Depending on the feedback we will receive from the lexicographic community, these extensions could be incorporated into the next version (P5) of the TEI guidelines.

## Bibliography

Baldinger, K. "L'étymologie d'hier et d'aujourd'hui." *Etymologie*. Ed. R. Schmitt. Darmstadt, 1977.

Burnard, L., and S. Rahtz. "Relaxing with Son of ODD, or What the TEI did Next." Paper delivered at the Extreme Markup Languages conference, Montréal (Canada), 2-6 August 2004. 2004.

Hausmann, F.J., O. Reichmann, H.E. Wiegand, and L. Zgusta, eds. *Wörterbücher. Ein internationales Handbuch zur Lexikographie*. Berlin / New York: Walter de Gruyter, 1990.

Ide, N., and L. Romary. "International Standard for a Linguistic Annotation Framework." *International Journal on Natural Language Engineering* (Forthcoming).

Romary, L., S. Salmon-Alt, and G. Francopoulo. "Standards going concrete: from LMF to Morphalou." Coling Workshop on Enhancing and Using Electronic Dictionaries, Geneva, 29 août 2004. 2004.

## Spanish Morphosyntactic Disambiguator

**Octavio Santana Suárez** (*osatana@dis.ulpgc.es*)

*Departamento de Informática y Sistemas.*

*Universidad de Las Palmas de Gran Canaria*

**José Rafael Pérez Aguiar** (*jperez@dis.ulpgc.es*)

*Departamento de Informática y Sistemas.*

*Universidad de Las Palmas de Gran Canaria.*

**Luis Javier Losada García** (*losada@dis.ulpgc.es*)

*Departamento de Informática y Sistemas.*

*Universidad de Las Palmas de Gran Canaria*

**Francisco Javier Carreras Riudavets**

(*fcarreras@dis.ulpgc.es*)

*Departamento de Informática y Sistemas.*

*Universidad de Las Palmas de Gran Canaria*

## 1. Introduction

The written expression of an idea is not achieved only through the simple combination of the different components of the grammar based on a given syntax. Other factors take part in the process, such as semantics and context. But it is obvious that a first approach requires at least a correct syntactic analysis, and for this it is necessary, from the computer-science point of view, to obtain results similar to those obtainable by human knowledge. In this work, a first approach is achieved by the identification and then disambiguation of the elements that are part of a sentence.

Traditionally, syntactic analysis requires a specialized knowledge of the language, all the more so in the case of Spanish, due to its wide range of variations which turn the syntactic analysis into a task only for experts. From the educational point of view, syntactic analysis is very useful to help learn to distinguish the different symbols implied: on the one hand, the correct combination of the elements by means of the application of grammar rules, and on the other hand, the incorporation of less tangible, although necessary aspects, like semantics and context. People usually perform an intuitive use that hides the true difficulty of the problem.

This system is intended to provide a close view of the Spanish grammar to researchers, enhancing their performance and

reliability. This is a first step that will allow, with the addition of new features, to keep improving until reaching 100% accuracy. Any automated processing of a text entails inevitably the syntactic analysis of its sentences, following the morphosyntactic disambiguation of the elements that compose it, allowing for different possible applications: a) to provide a precise synonym for a given word, b) to analyze its literary style, c) to reveal its semantics, d) to extract information or summarize its contents, e) to make trustworthy translations to other languages, f) to answer to concrete questions on its content, etc.

## 2. Methodology

In this work, the number of erroneous syntactic representation trees, obtained by the application of the rules of the Spanish grammar by means of a set of structural disambiguation rules, is notably reduced. In spite of the remarkable amount of necessary combinations, this system does not limit itself to subgroups of the grammar like most of the other proposals, but instead it uses a system of rules which covers all the possible combinations of the Spanish grammar. In addition to being the starting point for an automated syntactic analysis system, it complements the local functional disambiguator developed by the Group of Data Structures and Computational Linguistics of the University of Las Palmas de Gran Canaria ( <http://www.gedlc.ulpgc.es/investigacion/desambigua/desambigua.htm> ). As an indicator of its performance, the accuracy of the disambiguation is raised from 87% to 96%.

A solution is provided to the problem of the appearance of structural ambiguities that are generated during the process of construction of syntactic representation trees. The syntactic structures are combined to each other to allow for the syntactic representation trees. Many of these combinations generate erroneous trees. Direct conflicts between rules have been identified as one of the main causes of the problem. The characteristics of the different syntactic structures and how they must be considered at the time of accepting or not the construction of a representation symbol have been studied for the development of methods of structural disambiguation.

In view of the great number of possible combinations of the grammar elements (more evident in verb-phrase constructions which allow any number of elements and almost in any combination), the adequate representation mechanisms have been defined so that all the possibilities are covered, not leaving valid options unrepresented. When allowing any combination of possible elements in the verb-phrase, some combinations appear, which should not be allowed, and would be rejected in the structural disambiguation processes. In this way, all the

possible combinations are represented, from a structural point of view, and those not allowed are rejected.

Groups of semantic identification oriented to the recognition of syntactic structures are catalogued. The processes of structural disambiguation include some rules that introduce semantic information. The generated lists have been obtained from the tables of the ideological dictionaries that can be related to certain syntactic structures.

## 3. Knowledge base

The grammar used is based mainly on the description made by Gili Gaya. To achieve maximum system completeness and include all the syntactic structures that can appear we followed Gutiérrez Araus. The examples cited by Gómez Torrego (2002a, 2002b), were useful to test the system and contributed mainly to illustrate the aspects relative to the compound sentences that remained to be refined.

For this work, the tagger developed by GEDLC was used ( <http://www.gedlc.ulpgc.es/investigacion/scogeme02/lematiza.htm> ) which gathers the main lexicographical repertoires of the Spanish language<sup>1</sup>, and admits 151103 canonical forms and something more than 4900000 inflected and derived forms (without adding the inherent extension to the prefixes and the enclitic pronouns that have also been contemplated).

## 4. Related works

There are other authors that approach this problem for the Spanish language from diverse points of view. In the same way as our work, which can be used for free at discretion through the Internet ( <http://www.gedlc.ulpgc.es/investigacion/desambigua/morfosintactico.htm> ), we have only been able to find one other operative tool of this kind on the network: the parser from the Center of Language and Computing of the University of Barcelona. Given the high complexity of the problem, they have chosen to write down exclusively those elements that are explicitly present in the sentence, which had led them to a simplified treatment of some syntactic aspects like coordination and some subordinated types that they leave unsolved. Also, they abandon the concept of sentence understood like noun-phrase and verb-phrase, opting for a list of components instead.

Although the computer methodologies applied are different, they try to reach the same objectives. Our work is based on the real and complete study of: a) a Spanish grammar that includes all the possibilities available in the written language, b) the direct structural ambiguities that cause the appearance of multiple syntactic representation trees, c) the symbols that



cannot cover all the sentence, d) the complex verbal form, e) other situations where ambiguities can be solved based on linguistic knowledge about words, grammar categories and objects involved, and f) the considerations for the generation of the predicate symbol. Nevertheless, other methodologies apply statistical criteria for the resolution of ambiguities, with the consequent loss of reliability for unfrequent cases. The richness of our language and, particularly, the writers' freedom in the construction of syntactic structures makes us reconsider the probabilistic methods as the only solution to this complex problem.

## 5 Conclusions

This work is not limited to subsets of the grammar, but is based instead on a system of rules for the Spanish grammar in spite of the remarkable quantity of necessary combinations.

It provides a solution to the problem of the appearance of functional ambiguities. First a disambiguation process is applied, based on local syntactic structures that reach an accuracy of 87%; and second, another disambiguation process is applied, based on trees of syntactic representation that improve the average accuracy level up to 96%.

The importance of this work lies on the fact that it fosters the development of future applications, because:

1. It accelerates the process of syntactic analysis when pruning incorrect structures.
2. It improves the precision in the results of advanced word searches.
3. It allows the discarding of non valid options in information extraction.
4. It detects grammatical errors in the written constructions.

- 
1. Alvar Ezquerro; Casares; García Márquez & Hernández; Diccionario General de la Lengua Española Vox; Gran Diccionario de la Lengua Española; Gran Diccionario de Sinónimos y Antónimos; Moline; Real Academia Española.

## Bibliography

Bosque, I., V. Demonte, and F. Lázaro Carreter. *Gramática descriptiva de la lengua española*. Madrid: Espasa, 1999.

Casares, J. *Diccionario ideológico de la lengua española*. Barcelona: Gustavo Gili, 1994.

*Diccionario General de la Lengua Española Vox, Edición en CD-ROM*. Barcelona: Bibliograf, S.A., 1997.

Ezquerro, Alvar. *M. Diccionario de voces de uso actual*. Madrid: Arco-Libros, 1994.

García Márquez, Gabriel, and Humberto Hernández. *Clave. Diccionario de Uso del Español Actual, Edición en CD-ROM*. Madrid: Ediciones SM, 1997.

Gili Gaya, S. *Curso Superior de Sintaxis Española (Higher Course on Spanish Syntax)*. Barcelona: Bibliograf S.A., 1998.

Gómez Torrego, L. *Análisis sintáctico: Teoría y práctica*. Madrid: Ediciones SM, 2002a.

Gómez Torrego, L. *Gramática didáctica del español*. Madrid: Ediciones SM, 2002b.

*Gran Diccionario de la Lengua Española*. Barcelona: Larousse Planeta, S.A., 1996.

*Gran Diccionario de Sinónimos y Antónimos*. Madrid: Espasa-Calpe, 1991.

Gutiérrez Araus, M.L. *Estructuras sintácticas del español actual (Syntactic Structures of Current Spanish)*. Madrid: Sociedad General Española de Librería, S.A, 1978.

Moliner, M. *Diccionario de Uso del Español, Edición en CD-ROM*. Madrid: Gredos, 1996.

Quesada, J.F. *Un modelo robusto y eficiente para el análisis sintáctico de lenguajes naturales mediante árboles múltiples virtuales*. Sevilla: Centro Informático Científico de Andalucía (CICA), 1996.

Real Academia Española. *Esbozo de una nueva gramática de la lengua española*. Madrid: Espasa-Calpe, 1989.

Real Academia Española. *Diccionario de la Lengua Española, Edición electrónica*. Madrid: Espasa-Calpe, 1995.

Santana, O., J. Pérez, Z. Hernández, F. Carreras, and G. Rodríguez. "FLAVER: Flexionador y lematizador automático de formas verbales." *Lingüística Española Actual* XIX, 2 (1997): 229-282.

Santana, O., J. Pérez, F. Carreras, J. Duque, Z. Hernández, and G. Rodríguez. "FLANOM: Flexionador y lematizador automático de formas nominales." *Lingüística Española Actual* XXI, 2 (1999): 253 - 297.

Santana, O., J. Pérez, L. Losada, and F. Carreras. "Hacia la desambiguación funcional automática en Español." *Procesamiento del Lenguaje Natural* 28 (2002): 1-22.

# Una Herramienta de Recuperación Morfológica Aplicada a *Microsoft Word*

---

**Octavio Santana Suárez** (*osantana@dis.ulpgc.es*)

*Universidad de Las Palmas de Gran Canaria*

**Zenón Hernández Figueroa**

(*zhernandez@dis.ulpgc.es*)

*Universidad de Las Palmas de Gran Canaria*

**Gustavo Rodríguez Rodríguez**

(*grodriguez@dis.ulpgc.es*)

*Universidad de Las Palmas de Gran Canaria*

**Luis Losada García** (*llosada@dis.ulpgc.es*)

*Universidad de Las Palmas de Gran Canaria*

---

## 1. Introducción

Uno de los aspectos de la investigación en lingüística es el estudio del uso de la lengua en documentos escritos; se trata de identificar y analizar la aparición de determinadas construcciones, lo que, en gran medida, puede entenderse como una clase particular de lo que en informática se conoce como recuperación de información. En el ámbito de la recuperación de información se ha tenido desde siempre conciencia de la insuficiencia de las búsquedas exacta y parcial de las palabras de un texto, y también de la necesidad de incorporar información lingüística para una recuperación más completa. Las ya antiguas búsquedas con truncamiento parten de la hipótesis de que las distintas formas de una palabra se componen de una raíz fija acompañada de un sufijo o un prefijo variables; tal hipótesis suele ser bastante acertada para lenguas poco flexivas, pero resulta muy pobre con lenguas muy flexivas y con altas tasas de irregularidad. Las búsquedas con máscara, por similitud o en base a expresiones regulares no incorporan la adecuada información sobre la naturaleza morfológica de las palabras.

## 2. Antecedentes

El Grupo de Estructuras de Datos y Lingüística Computacional (GEDLC, <<http://www.gedlc.ulpgc.es>>) del Departamento de Informática y Sistemas de

la Universidad de las Palmas de Gran Canaria lleva algún tiempo desarrollando trabajos en morfología computacional, sintaxis automatizada, análisis de textos y lexicografía que incluyen lematizadores y flexionadores del español, así como el estudio de relaciones morfológicas entre las palabras.

El bagaje de conocimientos acumulado y la experiencia en el desarrollo de herramientas en el campo se ponen en este trabajo al servicio del desarrollo de sistemas de localización de fenómenos morfológicos del español dentro de un texto. Se ha realizado una aplicación de búsqueda lingüística aplicada a un procesador de textos popular —*Microsoft Word*.

El hecho de que el diálogo "Buscar y reemplazar" de *MS-Word XP* ofrezca una opción llamada «Todas las formas de la palabra» que según la ayuda de la aplicación sirve para «Buscar o reemplazar sustantivos, adjetivos o tiempos verbales» demuestra el interés de este tipo de búsquedas en el contexto de un procesador de textos. Pero la propia ayuda de la aplicación hace dudar del alcance de tales búsquedas al poner ejemplos como: «reemplace 'manzana' por 'naranja' y, al mismo tiempo, reemplazará 'manzanas' por 'naranjas'» o «reemplace 'peor' por 'mejor' y, al mismo tiempo, reemplazará 'el peor' por 'el mejor'»; ambos casos corresponden a simples sustituciones de cadenas de caracteres que no requieren ningún conocimiento lingüístico especial y que, de hecho, se pueden realizar sin seleccionar la opción «Todas las formas de la palabra»; más prometedor parece el ejemplo de los verbos: «reemplace 'dormir' por 'salir' y, al mismo tiempo, reemplazará 'dormido' por 'salido'», pero el *GEDLC* no ha logrado verlo funcionar.

## 3. La herramienta desarrollada

Se ha desarrollado una herramienta de búsqueda textual para *MS-Word* que incorpora los aspectos flexivos, derivativos y prefijales entre otros mecanismos de formación de palabras del español, lo que aporta una gran potencia de búsqueda.

A la hora de diseñar una aplicación que permita especificar patrones de búsqueda que tengan en cuenta aspectos flexivos y derivativos de la lengua hay que observar una cuestión fundamental: la gran cantidad de detalles que son susceptibles de configuración —la flexión verbal admite 116 configuraciones diferentes.

### 3.1 Organización

La aplicación se ha diseñado para presentar distintos niveles de detalle. El nivel básico muestra: una caja de entrada de texto, en la que el usuario debe introducir la palabra a buscar, un botón para iniciar la búsqueda, otro para usar la palabra como parte de una coocurrencia, y un par de flechas que dan acceso a mayores detalles.



Figure 1

El usuario sólo tiene que escribir una palabra y pulsar el botón Buscar. El patrón de búsqueda que se aplicará será el que esté configurado —por defecto corresponde a "cualquier palabra del texto que tenga una forma canónica que coincida con alguna de las formas canónicas de la palabra de búsqueda y que, para esa forma canónica, tenga la misma flexión".

En el siguiente nivel de detalle se pueden elegir los grados de derivación y de flexión; se usan tres escalas independientes: una para la derivación y otras dos para la flexión de las formas verbales y de las no verbales.

Si el usuario requiere una recuperación más precisa accederá al último nivel de detalle de la flexión —las relaciones morfológicas continúan en un nivel paralelo—, donde se podrá modificar el patrón de búsqueda, ampliando o recortando elecciones de flexión. Existe la posibilidad de añadir o quitar del patrón de búsqueda prefijos y, en el caso de los verbos, pronombres enclíticos.

Cabe tener en cuenta las formas canónicas que correspondan a la palabra de búsqueda o ignorarlas: por ejemplo, buscar palabras que sean "primera persona del singular del presente de indicativo de un verbo introducido" o, ignorando la forma canónica, "primera persona del singular del presente de indicativo de cualquier verbo".

Eligiendo una forma canónica se accede a la interfaz de configuración de las relaciones morfológicas en donde se puede indicar qué formas relacionadas se desea incluir en la búsqueda.

Si el usuario escribe un asterisco en lugar de una palabra, se abre la posibilidad de configurar un patrón de búsqueda por características morfogramaticales, sin determinación léxica; por ejemplo, localizar todas las palabras que sean "sustantivos femeninos plurales" o "formas verbales del presente de indicativo", independientemente de cualquier forma canónica.

Además de la búsqueda de palabras individuales, es posible la localización de coocurrencias, tanto con determinación léxica, como por características morfogramaticales —lo que permite afinar la búsqueda hasta el punto de poder situar fenómenos lingüísticos específicos, tales como: perífrasis verbales, regímenes preposicionales y colocaciones léxicas.

## 4. Conclusiones

Se ha elegido *MS-Word* por ser, seguramente, el procesador de textos más extendido bajo el entorno *MS-Windows* y disponer de interfaz COM (Component Object Model) que facilita la interoperabilidad con otras aplicaciones. La concepción de la herramienta en sí es tal que podría interactuar con otras aplicaciones que ofrezcan interfaces COM.

De hecho, el objetivo principal consistió en cómo configurar una interfaz que aprovechara los motores de lematización desarrollados por el *GEDLC* para realizar búsquedas que incorporen conocimiento lingüístico de forma potente, usable y efectiva. La decisión de que la herramienta desarrollada se aplicase a un procesador de textos pretendió evitar las distracciones derivadas de problemas particulares de otros ámbitos, tales como los de la navegación, si la herramienta se aplicaba a realizar búsquedas en la red, por ejemplo. El siguiente paso será adaptar la interfaz desarrollada para aplicarla a entornos más complejos que un procesador de textos, tales como: el análisis de corpus, el estudio del uso de la lengua en Internet, herramientas de apoyo a la enseñanza, etc. Es un proceso abordable dada la experiencia que también posee el *GEDLC* en ese campo, como se refleja en trabajos previamente publicados sobre analizadores de páginas Web. Análogamente, la herramienta desarrollada podría aplicarse a la recuperación de información en bases de datos textuales.

## Bibliografía

- Figuerola, Carlos G., Raquel Gómez Díaz, Angel F. Zazo Rodríguez, and José Luis Alonso Berrocal. "Spanish monolingual track: the impact of stemming on retrieval." *Evaluation of Cross-Language Information Retrieval Systems. Second Workshop of the Cross-Language Evaluation Forum, CLEF 2001. Darmstadt, Germany, September 2001; Revised papers, volume LNCS 2406 of Lecture Notes in Computer Science* (2002): 253-261.
- Figuerola, Carlos G., Raquel Gómez, Angel F. Zazo Rodríguez, and José Luis Alonso Berrocal. "Stemming in Spanish: A first approach to its impact on information retrieval." *Results of the Cross-Language System Evaluation Campaign CLEF 2001, Darmstadt, Germany*. Ed. Carol Peters. September 2001. 197-202.
- Santana, O., F. Carreras, J. Pérez, and G. Rodríguez. "Relaciones morfológicas prefijales del español." *Procesamiento de Lenguaje Natural* 32 (2004): 9-36.

Santana, O., F. Carreras, J. Pérez, and G. Rodríguez. "Relaciones morfológicas sufijales en español." *Procesamiento de Lenguaje Natural* 30 (Marzo, 2003): 1-73.

Santana, O., J. Pérez, Z. Hernández, F. Carreras, and G. Rodríguez. "FLAVER: Flexionador y lematizador automático de formas verbales." *Lingüística Española Actual* XIX.2 (1997): 229-282.

Santana, O., J. Pérez, F. Carreras, Z. Hernández, J. Duque, and G. Rodríguez. "FLANOM: Flexionador y lematizador automático de formas nominales." *Lingüística Española Actual* XXI.2 (1999): 253-297.

---

## A Digital Environment for Neolatin Studies

---

**Ross Scaife** (*scaife@gmail.com*)

*University of Kentucky*

**Andrew Gollan** (*tharoth@gmail.com*)

*University of Kentucky*

**William du Cassé** (*villelmus@gmail.com*)

*University of Kentucky*

**Jennifer Nelson** (*jnels2@uky.edu*)

*University of Kentucky*

---

**T**he 2005 annual convention of the ACH/ALLC to be hosted at the University of Victoria will have the intersection of multilingualism and humanities computing as a prominent theme. It is therefore the perfect venue for presentation of our work.

The graduate Institute for Latin Studies at the University of Kentucky, begun in 2000, strongly emphasizes the active use of the Latin language via pervasive oral immersion and rigorous assignments in written composition. The faculty conducts a sequence of graduate classes solely in Latin, and students meet spontaneously in various extracurricular venues for further practice. Equally against the grain, the Institute does not restrict its focus to Republican and Imperial Latin of the ancient world as Classical Studies departments typically do, but rather encourages an appreciation of Latin from *all* periods, from antiquity through early modern times.

This presentation will explore the many challenges and opportunities posed for humanities computing by the practices and emphases of the UK Institute for Latin Studies. These challenges fall into the general categories of availability of materials, interface design for a reading environment, and general computational infrastructure for the study of Latin.

Regarding materials, we are responding to a general lack of adequate resources for the active use of Latin as well as the diachronic study of the language. We have produced TEI-conformant XML editions of numerous Neolatin texts, including the *Moriae Encomium* of Erasmus, the *Orationes* of Muretus, the *Argenis* and *Icon Animorum* of Iohannes Barclaius, the *Eudemia* of Ianus Nicius, and the *Psyche Cretica* of Ioannes Ludovicus Praschius. We have begun to produce the world's only unified archive (in any medium) of all known Latin

colloquia, dialogues written expressly for the purpose of teaching people how to speak Latin. We are now completing work on electronic editions of colloquia by Erasmus, Corderius, Duncanus, and Vives, and have already identified many more examples from this genre for inclusion in our archive. We intend to have a constant stream of students writing (Latin) commentaries on the texts that we publish. As an example, last year a graduate student in our program made a selection from Erasmus' *Colloquia*, equipping the texts with introductions, notes and questions to help teachers guide students through texts that may be unknown to the teachers themselves. This kind of work will be important in advancing the utility of our text archive.

Another current digitization project that is related to our goal of keeping Latin learners' minds in Latin is our production of a TEI-XML edition of an all-Latin grammar, *Grammatica classicae latinitatis* by J. Llobera and E. Alvarez (Barcelona 1919). Also, with support from the UK Center for Computational Sciences, this year we are developing a separate grant proposal that would fund a multiyear project to digitize important all-Latin lexica by Forcellini and Du Cange.

We realize that most teachers currently lack experience in conducting their classrooms in spoken Latin, so we will also create our own web-based interactive dialogues (in audiovisual formats) of progressive sophistication. Such exercises in spoken Latin will facilitate students' understanding and will foster an immediate familiarization with the language.

All of the digitized materials we create will be freely available online and will carry appropriate Creative Commons licenses, permitting students and scholars worldwide to use them without cost or restrictions. Our repository will link with others via the Classical Text Services Protocol now being developed under the auspices of the Center for Hellenic Studies, broadening the selection of high quality marked up classical texts available worldwide, as well as allowing our other tools to operate, at least partially, on texts outside our project.

What we anticipate as an ultimate synthesis of all these activities is a digital reading and learning environment, built on the *Apache Cocoon* platform, that allows our own resources, and others flowing in from elsewhere, to be tied together and used selectively to meet the needs of a vertical spectrum of Neolatinists. For students we will associate well spoken versions of the source texts, allow them to click through to morphological and hence lexical data, connect them with grammatical and historical notes, show them a variety of translations, and perhaps even give feedback on their own pronunciations. More advanced scholars may be able to see photographs of manuscripts, automatically collate manuscript transcriptions according to their own stemmatological theories, or call up related commentaries and other scholarship.

Many of the individuals interested in later Latin are not professional scholars, but the Internet allows talented amateurs back into the academic market. How can we best manage the signal to noise ratio? Moreover, those professional scholars of Neolatin who do exist are scattered worldwide. We need to provide tools that not only give everyone access to rarer texts (and ancillary resources), but also harness the expertise and enthusiasms of the reading public to improve them. The work done at the University of Kentucky on the *Suda On Line* project showed how a web-based community of scholars could tackle a project too large for any one individual and gradually make real progress. Accordingly we plan on extending the results of that experiment with a framework for lexicography that enables the continuous improvement of our fundamental reference works by a distributed set of users.

## Bibliography

[No source references provided. Eds.]

## Letters and Lacunae: Editing an Electronic Scholarly Edition of Correspondence

---

**Susan Schreibman** (*sschreib@umd.edu*)

*University of Maryland*

**Gretchen Gueguen** (*ggueguen@wam.umd.edu*)

*University of Maryland*

**Amit Kumar** (*amitku@uiuc.edu*)

*University of Illinois at Urbana Champaign*

**Ann Saddlemyer** (*sadlemy@uvic.ca*)

*University of Victoria*

---

Encoding editions of documentary texts, particularly editions of correspondence, within the *Text Encoding Initiative (TEI) Guidelines* raises special challenges not encountered when editing previously published works. The challenges fall into three broad categories: 1) difficulties in capturing bibliographic meta-information describing the physical object and its transmission history; 2) challenges in developing a controlled vocabulary suitable to the informal nature of texts which were never intended for publication; and 3) difficulties in encoding both physical characteristics of the documentary texts, as well as their intellectual content, i.e. adopting a principle of encoding the text either as a physical artifact or as a conceptual work. These challenges, particularly as they relate to encoding letters, will be explored by through an edition currently being edited entitled *Thomas MacGreevy and George Yeats: A Friendship in Letters*.

During the next two years members of *The Thomas MacGreevy Archive* team will be creating for online publication an edition of the correspondence between George Yeats (1893-1968), wife of the Irish poet W.B. Yeats, and Thomas MacGreevy (1893-1967), Irish poet, art and literary critic, and Director of the National Gallery of Ireland (1950-63). It is a collection spanning 41 years, comprising 148 letters. The letters are fascinating documentary records which provide a window not only into the personal lives of the authors, but into the artistic and political circles in which they moved, providing a unique insight into the new Irish Free State and the cultural climate of Europe during the first half of the twentieth century. The letters are being encoded using Extensible Markup Language (XML) according to newly released *P5 TEI Guidelines* to take advantage of the TEI's new chapter on Manuscript Description.

Although the *TEI Guidelines* were not developed specifically to encode previously published texts, many of the rules built into the syntax of the Document Type Definitions (DTDs) favor this document type. To cite but one example, the content model of `tei.divbot` does not allow for a paragraph `<p>` element after the closer element `<closer>`. While the need for additional paragraphs after closing material in published texts may be uncommon, letters frequently have a closing salutation, followed by a postscript. Moreover, it has proved difficult within the TEI header to detail the type of descriptive information that editors, scholars, and bibliographers require when engaging with handwritten documents.

Individual projects (such as *DALF: Digital Archive of Letters in Flanders Project*) and subject- area consortiums (such as *The Model Editions Partnership*) have developed their own extensions to the *TEI Guidelines* to accommodate the needs of electronic editions of correspondence. After a brief survey of the strategies employed by these and other editions, we will discuss how TEI's new chapter on manuscript description alleviates some of the problems previous projects solved with local solutions. The chapter on Manuscript Description builds on the work of two separate initiatives which have been recently combined: *MASTER* project (1999-2001), an EU-funded project headed by Peter Robinson, and the work of the *TEI Medieval Manuscripts Description Work Group* (1998-2000), headed by Consuelo Dutschke and Ambrogio Piazzoni. The new elements available in this tagset provide for detailed description of primary texts including transmission, physical description, the relationship between parts of the manuscript (for example, when a poem is enclosed with a letter), dimensions, location, manuscript identification, provenance, and history of ownership.

Another area to be discussed is the difficulties in developing an ontology or controlled vocabulary for a correspondence. The ontology, the backbone for the search page, is more difficult to develop for a collection of letters than other document types. Subject headings, such as the *Library of Congress Subject Headings (LCSH)*, which are used to describe entire collections or self-contained bodies of information, are not suitable for this project which describes each letter individually. The problem with using schemes such as *LCSH* is twofold: one, the letters cover many subjects and follow no formal organization pattern, making it difficult to make a faceted indexing schema like *LCSH* worthwhile; secondly, the subject headings were meant to be used in the cataloging of cohesive works or collections, and were not designed to be brief entries in the index for a specific work or collection.

The indexing done for this edition more closely resembles back-of-the-book style indexing in terms of its description of the details of the text. Standard controlled vocabularies that might be used in this type of indexing, like the *Getty Art and Architecture Thesaurus*, on the other hand, are too specific and

terms do not sufficiently summarize or categorize the topics discussed. Capturing, representing, and, indeed, interpreting a multitude of topics present in any given letter — from general subjects to more intimate personal details — is of paramount importance. If ontology is defined as a "formal, explicit specification of a shared conceptualization" (Fensel 11), the burden of interpreting by a third party what a "shared conceptualization" of a text written for an intended audience of one is immense. Indeed, as the correspondence itself often indicates, meaning is often misconstrued by the intended recipient. Given these difficulties, other types of structured data, such as annotation and abstracts, may be used to mitigate issues of keywords conveying different meanings when taken out of textual context.

Another challenge when editing documentary texts for electronic publication is choosing a philosophy by which to encode. This is particularly true in the case of editing modern correspondence. Editors have had to traditionally decide whether the purpose of the encoding is to capture the physical appearance of the page (regardless of the text's logical sequence), or whether it is to record the textual/ontological flow (regardless of the text's physical appearance). In traditional print publications, editions (except for facsimiles) reflect a logical sequencing of the text. For example, text which appears in the margins is placed where the editor feels it belongs logically, even when the writing crosses page boundaries (such as finishing a letter in the margins of the first page when the author ran out of room on the last).

This edition is exploring methods of encoding both the physical appearance of the page, as well as the letter's logic. This is particularly challenging when encoding, for example, marginalia. To represent the marginalia within the logical sequence of the text, the editor must decide where it is to be anchored within the textual flow. To represent it in a physical representation, the editor must provide coordinates that will anchor the text vertically and horizontally in relation to the main body of the work. While some of this positioning is absolute, for example, anchoring text at the top of the page, other positioning is relative, for example, anchoring marginalia relative to the paragraph it appears next to. While the encoding must take into account, in some measure, the technologies available to us today, XSLT, CSS, and JavaScript, for example, at the same time it must also be encoded with a view to future presentations, independent of current technologies.

These are a sampling of issues that will be discussed.

## Bibliography

Chestnutt, R. David. "The e Model Editions Partnership: 'Smart Text' and Beyond." *DLib Magazine* (July/August 1997). <<http://www.dlib.org/dlib/july97/07chestnutt.html>>

*DALF: Digital Archive of Letters in Flanders Project*. Centrum voor Teksteditie en Bronnenstudie (KANTL). Accessed 2005-03-21. <<http://www.kantl.be/ctb/project/dalf/>>

DeRose, Steven J., David G. Durand, Elli Mylonas, and Allen H. Renear. "What is Text, Really?" *Journal of Computing in Higher Education* 2.1 (Winter) (1990): 3-26.

Farrow, John. "All in the Mind: Concept Analysis in Indexing." *The Indexer* 19.4 (1995): 243-247.

Fensel, Dieter et al. *Spinning the Semantic Web*. Cambridge, Massachusetts: MIT Press, 2005.

Matthews, Douglas. "Indexing Published Letters." *The Indexer* 22.3 (2001): 135-141.

Renear, Allen H., Elli Mylonas, and David G. Durand. "Refining Our Notion of What Text Really Is: The Problem of Overlapping Hierarchies." *Research in Humanities Computing 4: Selected Papers from the ALLC/ACH Conference, Christ Church Oxford, April 1992*. Ed. Susan Hockey and Nancy Idle. Oxford: Oxford University Press, 1996. 263-280.

Schreibman, Susan. *The Thomas MacGreevy Archive*. Accessed 2005-03-21. <<http://macgreevy.org>>

*TEI Guidelines P4*. Accessed 2005-03-15. <<http://www.tei-c.org/Guidelines2/index.html>>

*TEI Guidelines P5, Manuscript Description Chapter*. Accessed 2005-03-15. <<http://www.tei-c.org/Activities/MS/FASC-ms.pdf>>

## **The *Blackwell Companion to Digital Humanities*: a Roundtable Discussion**

---

**Susan Schreibman** ([sschreib@umd.edu](mailto:sschreib@umd.edu))

*University of Maryland*

**Ray Siemens** ([siemens@uvic.ca](mailto:siemens@uvic.ca))

*University of Victoria*

**John Unsworth** ([unsworth@uiuc.edu](mailto:unsworth@uiuc.edu))

*University of Illinois*

**Willard McCarty** ([willard@mccarty.me.uk](mailto:willard@mccarty.me.uk))

*Kings College, London*

**Martha Nell Smith** ([mnsmith@umd.edu](mailto:mnsmith@umd.edu))

*University of Maryland*

**Geoffrey Rockwell** ([georock@mcmaster.ca](mailto:georock@mcmaster.ca))

*McMaster University*

**Abby Smith** ([asmith@clir.org](mailto:asmith@clir.org))

*Council on Library and Information Resources*

**Claire Warwick**

*University College London*

**Perry Willett**

*University of Michigan*

---

This session will reflect on the recently published *Blackwell Companion to Digital Humanities* by six of its contributors and its three editors. This collection marks a turning point in the field of digital humanities: for the first time, a wide range of theorists and practitioners, those who have been active in the field for decades, and those recently involved, disciplinary experts, computer scientists, and library and information studies specialists, have been brought together to consider digital humanities as a discipline in its own right, as well as to reflect on how it relates to areas of traditional humanities scholarship.

The participants for this panel discussion reflect the broad range of themes and disciplinary areas of interest that the 38 chapters of the *Companion* address. Rockwell's chapter "New Media" (co-authored with Andrew Mactavish) is part of the History section, which considers the field from disciplinary perspectives. Perry Willett's chapter, "Perspectives and

Communities", and Willard McCarty's chapter entitled "Modelling", represent the second section, 'Principles', which includes chapters on databases, text encoding, and communities. Martha Nell Smith, writing about "Electronic Scholarly Editing", and Claire Warwick, writing on "Print Scholarship and Digital Resources", are part of the *Companion's* third section, entitled 'Applications', which covers a wide range of cross-disciplinary perspectives on how computer-mediation has changed our approach from fields as diverse as authorship studies, robotic poetics, and speculative computing. Abby Smith's chapter on "Preservation", is from the *Companion's* last section entitled 'Production, Dissemination, Archiving', which covers a broad range of practical issues (including project design, conversion of primary sources, text tools, and preservation).

The panel will open with a discussion of the collection's origins in the research carried out over the past half a century on textually-focused computing in the humanities. It will, however, quickly move on to how broadly the field now defines itself, which is evident from even the most cursory glance at the *Companion's* table of contents. The field remains deeply interested in text. But as advances in technology have made it first possible, then trivial to capture, manipulate, and process other media, the field has redefined itself to embrace the full range of multimedia. Especially over the last decade with the advent of the World-Wide Web, digital humanities has broadened its reach. At the same time, it has remained in touch with the goals that have animated it from the outset: using information technology to illuminate the human record, and bringing an understanding of the human record to bear on the development and use of information technology. In it is in these areas that the chapters by Willett and McCarty are especially relevant.

The first eleven chapters of the *Companion* address the field from disciplinary perspectives. Although the breadth of fields covered is wide, what is revealed is how computing has cut across disciplines to provide not only tools, but also methodological focal points. What the editors discovered when doing final editing of the volume is that there exists a common focus across disciplines on preserving physical artifacts whether these have been left to us by chance (ruin, and other debris of human activity), or that which has been near-impossible to capture in its intended form (music, performance, and event). Yet many disciplines have gone beyond simply wishing to preserve these artifacts, to re-representing and manipulating them so that their hidden properties and traits can be revealed. Moreover, digital humanities now also concerns itself with the creation of new artifacts which are born digital and require rigorous study and understanding in their own right.

What was also revealed in editing the volume was the widespread notion that there is a clear and direct relationship



between the interpretive strategies that humanists employ and the tools that facilitate exploration of original artifacts based on those interpretive strategies. More simply put, those working in the digital humanities have long held the view that application is as important as theory. This point is clearly demonstrated in the chapters by Martha Nell Smith and Clare Warwick. Thus exemplary tasks traditionally associated with humanities computing hold the digital representation of archival materials on par with analysis or critical inquiry, as well as theories of analysis or critical inquiry originating in the study of those materials. The field also places great importance on the means of disseminating the results of these activities as well as the realization that strategies for preserving digital objects must be built into the design process at the very earliest stages of project design, as evidenced by Abby Smith's contribution.

The panel will close with a discussion of how the *Companion* serves as a historical record of the field, capturing a sense of the digital humanities as they have evolved over the past half-century, and as they exist at the moment. Yet, if one looks at the issues that lie at the heart of nearly all contributions to this volume, one will see that the taken as a whole, this collection reflects a relatively clear view of the future of the digital humanities, which the panel will also consider. Lastly, the panel will address how digital humanities is addressing many of the most basic research paradigms and methods in the disciplines, to focus our attention on important questions to be asked and answered, in addition to important new ways of asking and answering that are enabled by our interaction with the computer.

Unsworth, John. Inaugural E-humanities Lecture at the National Endowment for the Humanities, April 3, 2001. Accessed 2005-03-21. <<http://www.iath.virginia.edu/~jmu2m/KR/>>

## Bibliography

Hockey, Susan. *Electronic Texts in the Humanities*. Oxford: Oxford University Press, 2000.

McCarty, Willard. *What is Humanities Computing? Toward a Definition of the Field*. Accessed 2005-03-09. <<http://www.kcl.ac.uk/humanities/cch/wlm/essays/what/>>

Schreibman, Susan. "Computer-mediated Discourse: Reception Theory and Versioning." *Computers and the Humanities* 36.3 (2002): 283-293.

Schreibman, Susan, Ray Siemens, and John Unsworth. *A Companion to Digital Humanities*. Oxford: Blackwell's, 2004.

Siemens, R.G. "'A New Computer-Assisted Literary Criticism?' An introduction to A New Computer-Assisted Literary Criticism?" [*A special issue of*] *Computers and the Humanities* 36.3 (2002): 259-267.

# Representation of Meaning: a Graphical and Interactive Approach

---

*Gary Shawver* (*gary.shawver@nyu.edu*)

NYU

*Oliver Kennedy* (*xthemage@nyu.edu*)

NYU

---

You shall know a word by the company it keeps!

(Firth 179)

We assume in the design of this software a view of language that sees the meaning of a word as inextricably linked to its contexts, a view that is in some ways foreshadowed in Augustine's advice to readers of Scripture to resolve semantic ambiguities by examining the context of a passage (3.2.2). To some extent, our software's function and design is also informed by a Saussurean structuralist semantics that holds that a word's meaning is "determined by the [horizontal] paradigmatic and [vertical] syntagmatic relations" (Lyons 268) between that word and others in a system of language. This is illustrated in Saussure's example of the game of chess in which the value of individual pieces is not the result of any intrinsic qualities but rather of their relation to other pieces (Saussure 108, 87 ff.). Of course, the method of deducing what a word named (its reference) for Augustine becomes a mode of meaning for Wittgenstein, and while Saussure's focus is upon a word's place within a language system (its sense), that of Wittgenstein is upon a word's place within instances of discourse.

J. R. Firth was perhaps the first linguist to give Wittgenstein's statement serious consideration when he proposed that one could define a word by looking at its linguistic context, or "the mere word accompaniment, the other word-material in which [a word is] most commonly or most characteristically embedded" (Firth 180). Seeking to differentiate this from *context of situation*, Firth called it *collocation* (195). In a sense, this study of collocation, the linguistic material or *linguistic context* within which a word commonly occurs, involves the mapping of pathways leading to and from each word (Palmer 76). Firth's idea of collocation, first presented in 1951, introduced a new type of word meaning, but would have to await the introduction of computers and electronic texts before it could be used to study texts of any significant size (Berry-Rogghe 103). His ideas have most famously borne fruit in the *COBUILD* dictionaries and grammar and in the field of corpus linguistics. The collection and classification of collocates require the computer's unique ability to perform repetitive tasks quickly and efficiently. Software like *TACT* brought this ability to desktop PC users and thus solved the problem of collection, but introduced a new set of challenges, one of which was how to represent the data gathered by such tools. A multitude of familiar paths lead off from these words in every direction. (Wittgenstein 143e, 144e)

During the initial phases of memory encoding, it is as if one is preserving the shape of a letter by walking a particular route on a lawn. The pattern is dynamic, and is only evident in the way one moves. After a period of time, however, the grass will wear through, creating a dirt pathway. At that point, one can stop walking; the information is preserved structurally (Kosslyn & Koenig, 351)

## Introduction

This paper will consist a discussion of some of the theoretical assumptions underlying the design of a piece of software and a demonstration of its function. This software, the *Fixed Phrase Tool* is part of the first release of *TAPoR* (*text analysis portal for research*). I first saw the desirability of such software while doing my doctoral dissertation in Toronto, a computer-aided text analysis of the semantics of *story* and *tale* in Chaucer. It was clear that the lists generated by traditional texts analysis tools such as *TACT* were both tedious to examine and presented information in a serial fashion that was not suitable to data that were, in some sense, relational. It was also clear that attempts to reformat the output of such 'listware' into more relational forms is exceedingly laborious. Of course, this practical need presupposes certain assumptions about meaning, which follow this rough outline.

## Theory

But if both meanings, or all of them. . . remain ambiguous after the faith has been consulted, then it is necessary to examine the context of the preceding and following parts of the ambiguous place, so that we may determine which of the meanings among those which suggest themselves it would allow to be consistent.

(Augustine, 2.2)

The value of the chess pieces depends upon their position upon the chess board, just as in the language each term has its value through its contrasts with all the other terms.

(Saussure 88)

For a large class of cases —though not for all— in which we employ the word 'meaning' it can be defined thus: the meaning of a word is its use in the language.

(Wittgenstein 20e)

---

The significance of a collocate lies repetition and syntagmatic relation. Therefore, the *Fixed Phrase Tool (FPT)* focuses on collecting phrasal repetends, a subclass of collocation. As Ian Lancashire has noted, phrasal repetends fall into two categories: "repeating fixed phrases" in which word order is unchanging and "repeating collocations, in which words co-occur, although in varying order, sometimes with words intervening" (100). Repeating fixed phrases (*fixed phrases* hereafter) may most closely represent the network of *familiar paths* that constitute a person's language. From a neuroscientific perspective, as Lancashire has noted, they are the result of such paths (188 ff). He also attempted to represent these neurological structures in what he called "phrasal repetend graphs" (217).

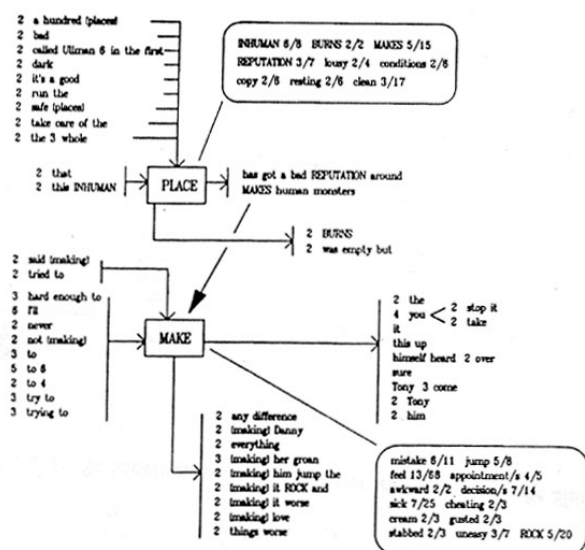


Figure 1

We have adopted this model fully in the *FPT*. Although it outputs the standard list of fixed phrases, the tool can also automatically format such lists into a form very much resembling that above, though, at present showing only fixed phrases. The neurological and philosophical roots of the path model of meaning also imply a degree of interactivity. This is implemented in the Java version of the *FPT*, where each node word in the graph can itself become a node in its own graph.

## Bibliography

Augustine of Hippo. *On Christian Doctrine*. Trans. D.W. Robertson. New York: Macmillan, 1958.

Berry-Rogghe, Godelieve L.M. "The Computation of Collocations and Their Relevance in Lexical Studies." *The Computer and Literary Studies*. Ed. A.J. Aitken, R.W. Bailey

and N. Hamilton-Smith. Edinburgh: Edinburgh University Press, 1973. 103-112.

Firth, J.R. "Modes of Meaning." *Papers in Linguistics, 1934-1951*. London: Oxford University Press, 1957. 190-215.

Firth, J.R. "A Synopsis of Linguistic Theory, 1930-55." *Selected Papers of J.R. Firth 1952-59*. Ed. F.R. Palmer. London, Eng: Longmans, 1968. 168-205.

Kosslyn, Stephen M., and Olivier Koenig. *Wet Mind: The New Cognitive Neuroscience*. New York: The Free Press, 1995.

Lancashire, Ian. "Uttering and Editing: Computational Text Analysis and Cognitive Studies in Authorship." *Texte: Revue de Critique et de Théorie Littéraire: Texte et Informatique* 13.14 (1993): 173-218.

Lancashire, Ian. "Phrasal Repetends and 'The Manciple's Prologue and Tale'." *Computer-Based Chaucer Studies*. Ed. Ian Lancashire. Toronto: University of Toronto, 1993. 99-122.

Lyons, John. *Semantics*. 2 vols. Cambridge, UK: Cambridge University Press, 1993.

Lyons, John. *Linguistic Semantics: An Introduction*. Cambridge, UK: Cambridge University Press, 1995.

Lyons, John. "Firth's Theory of 'Meaning.'" *In Memory of J. R. Firth*. Ed. C.E. Bazell. London: Longmans, 1966. 288-302.

Palmer, F.R. *Semantics*. 2nd ed. Cambridge, UK: Cambridge University Press, 1981.

Saussure, Ferdinand de. Ed. Charles Bally and Albert Sechehaye. *Course in General Linguistics*. Trans. Roy Harris. Chicago: Open Court, 1996.

Wittgenstein, Ludwig. *Philosophical Investigations*. Trans. G.E.M. Anscombe. 2nd ed. Oxford: Basil Blackwell, 1967.

## Keyword Extraction in Information Retrieval

---

**Harold Short** (*harold.short@kcl.ac.uk*)

*King's College London*

**Marilyn Deegan** (*marilyn.deegan@kcl.ac.uk*)

*King's College London*

**Laszlo Hunyadi** (*hunyadi@ling.arts.klte.hu*)

*University of Debrecen*

**Paul Baker** (*p.baker@lancaster.ac.uk*)

*Lancaster University*

**Dawn Archer** (*d.archer@lancaster.ac.uk*)

*University of Central Lancashire*

**Tony McEnery** (*a.mcenery@lancaster.ac.uk*)

*Lancaster University*

---

**S**ession chair: Harold Short.

### Keyword extraction: an Overview

#### Laszlo Hunyadi

With the ever increasing amount of information made available in all areas of life, including economy, science, education and culture, there is an imperative need to retrieve, elaborate and present this information in the most optimal way. Since the bulk of information is textual, information retrieval is mainly concerned with texts. This talk will present an outline of the approaches, principles and techniques used in textual data retrieval based on keyword extraction. There will be an analysis of the capabilities of various approaches showing their appropriate uses.

### Keyword extraction in the project Forced Migration Online

#### Marilyn Deegan

*Forced Migration Online* ( <http://www.forcedmigration.org/> ) provides instant access to a wide variety of online resources dealing with the situation of forced migrants worldwide. Being one of the most comprehensive textual resources dealing with humanitarian issues across a large

number of countries, it faces the challenges of multilingualism, multiculturalism and the essential requirement to be up-to-date and informative. That is why the organisation and retrieval of textual information as well as operability have high priority in the design and functioning of the system. This talk presents the essentials of this online resource, including its pioneering beginnings and view of future development.

### Querying Keywords: Questions of difference, frequency and sense

#### Paul Baker

This paper examines issues to do with interpreting keyword lists (Scott 1999), such as over-attending to lexical differences whilst ignoring differences in word usage and/or similarities between texts. Using a variety of techniques (e.g. analysis of key clusters or annotated data), I show how researchers can use keyword analyses to obtain a more accurate picture of the distinctive features of their texts or corpora.

### Love - a familiar or a devil? An exploration of key domains in Shakespeare's Comedies and Tragedies

#### Dawn Archer, Jonathan Culpeper, Paul Rayson

Love is a common theme in Shakespeare's works. In this paper, we show how the *UCREL Semantic Annotation Scheme* (henceforth *USAS*), a software program for automatic dictionary-based content analysis, can help us to explore the semantic field of 'love' within a selection of Shakespeare's plays. Specifically, we will explore 3 love-tragedies (*Othello*, *Antony and Cleopatra*, and *Romeo and Juliet*) and 3 love-comedies (*A Midsummer Night's Dream*, *The Two Gentlemen of Verona* and *As You Like It*) to determine differences in their (re)presentation of 'love'. We will also discuss how the semantic field of 'love' co-occurs with different domains in the plays, and assess the implications this has on our understanding of 'love' as a concept. This research builds on (i) Jonathan Culpeper's work on keywords in Shakespeare, using *Wordsmith* (Culpeper 2002), (ii) Paul Rayson's comparisons of key word and key domain analysis (Rayson 2003), and (iii) Dawn Archer and Paul Rayson's work on the identification of key domains in refugee literature, using *USAS* (Archer and Rayson forthcoming).

### Key words and the analysis of discourses in historical contexts

#### Tony McEnery

This paper examines the use of keywords to approach the discourse of moral panic evident in the writings of the Society

for the Reformation of Manners in late seventeenth/early eighteenth century England. The keyword approach, I will argue allows one to populate a model of moral panic discourse, while simultaneously showing how, in that historical context, links were forged between concepts which, while unlinked then, have become naturalised as being linked in modern English. By showing how keywords relate to discourse, and ultimately to a process whereby meanings and objects become linked, the paper will argue that keywords are important tools for the historical linguist in studying the shifting patterns of word association in language.

## Bibliography

Archer, D., and P. Rayson. "Using the UCREL automated semantic analysis system to investigate differing concerns in refugee literature." *The Keyword Project: Unlocking Content Through Computational Linguistics*. Ed. M. Deegan, L. Hunyadi and H. Short. Office for Humanities Communication Publications, Forthcoming.

Culpeper, J. "Computers, language and characterisation: An analysis of six characters in Romeo and Juliet." *Papers from the ASLA symposium, Conversation in life and literature*. Ed. Ulla Melander Marttala, Carin Ostman and Merja Kyto. Uppsala: Association Suedoise de Linguistique Appliquee, 2002. 11-30.

Rayson, P. "Matrix: A statistical method and software tool for linguistic analysis through corpus comparison." Ph.D. thesis, Lancaster University.

## Theory and Practise in Literary Textual Analysis Tools

**Ray Siemens** (*siemens@uvic.ca*)

*University of Victoria*

**Geoffrey Rockwell** (*georock@mcmaster.ca*)

*McMaster University*

**Susan Schreibman** (*sschreib@umd.edu*)

*University of Maryland*

**Matthew Jockers** (*mjockers@stanford.edu*)

*Stanford University*

## Panel Description

**T**hrough discussion of several exemplary literary textual analysis tools, participants on this panel explore elements of the literary studies community's reaction to textual analysis computer tool development -- and, particularly, how theorists perceive the development of tools as an activity that supports, tests, models, and expands upon their work. Panel contributors challenge the oft-perceived disparity between the 'lower' criticism (enumerative, bibliographic, re-presentative, &c.) in which most computing tools that we use have their origins and the 'higher' criticism often associated with thematically-oriented literary critical theory.

**Geoffrey Rockwell**, McMaster U (presenter)

**Matt Jockers**, Stanford U (presenter)

**Susan Schreibman**, U Maryland (presenter)

**Ray Siemens**, U Victoria (chair and respondent)

## Interrupting the Machine to Think About It

**Geoffrey Rockwell**

"A machine may be defined as a *system of interruptions* or breaks (*coupures*). . . Every machine, in the first place, is related to a continual material flow (*hylè*) that it cuts into." (Deleuze and Guattari 36)

Text analysis tools (and for that matter any form of analysis) perform two types of operations. They interrupt the flow of continuous analog information in order to break it down into samples that can be quantified and then they synthesize new eruptions out of the samples. Even the representation of a text in digital form is a matter of machined sampling and

quantitative representation whether you chose to represent a printed page as pixels or characters.

This interrupting and breaking down is a process that constrains what computer-based tools can do and that is the first point of this paper. The sampling and quantization also makes it possible to develop synthetic processes that create new hybrid artefacts like text visualizations or sonoric representations, the second point of this paper.

Finally, the breaking down (and not the transparent functioning) is the (error) message of the textual machine. We know the machine when it fails, when it is in error, and when it delivers monstrous results. To stand back and look at a machine, as opposed to looking through it, is to think through ambitious failure.

Such a thinking through a computer is pragmatic theorizing in a tradition of thinking while tinkering - a thinking often provoked by what is at hand. What is proposed is a theory of computer assisted text analysis that addresses the way such ruptures stress interpretation. Development happens in rupture, both the programming development that scripts computers and the performance of thinking (about machines and texts) called developing a theory.

In the meantime, *The Bug* that mocks us and interrupts our demonstrations is also what provokes reflection and adaptation. We wouldn't want it any other way, except at the moment of machined interruption, for which reason a demonstration of TAPoRware text analysis tools will interrupt this paper.

## Bibliography

Deleuze, Gilles, and Félix Guattari. *Anti-Oedipus: Capitalism and Schizophrenia*. Trans. Robert Hurley. Minneapolis: University of Minnesota Press, 1983.

Ullman, Ellen. *The Bug*. New York: Nan. A. Talese, 2003.

Yan, Lian, and Geoffrey Rockwell. *TAPoRware*. Accessed 2005-03-22. <<http://taporware.mcmaster.ca/>>

## Visualizing the Hypothetical, Encoding the Argument

### Susan Schreibman

The *Versioning Machine* (VM) <<http://www.mith2.umd.edu/products/ver-mach>> was launched at ACH/ALLC 2002 as a tool to display multiple witnesses of deeply encoded text. It was designed as a presentation tool so that editors could engage with the challenging work of textual editing, rather than becoming experts in other technologies, such as XSLT, JavaScript and CSS, all components of the *Versioning Machine*. The application allows encoders who

utilize the *Text Encoding Initiative's* Parallel Segmentation method of encoding to view their documents through a browser-based interface which parses the text into its constituent documents (at present the VM works best with *Internet Explorer* 6.0 and higher, but it also works with *Firefox* for PC and Mac). The *Versioning Machine* also provides several features for the end user to engage with texts, including highlighting a structural unit (paragraphs, lines, or divs) across the witness set, synchronized scrolling, and the ability to display a robust typology of notes.

The *TEI's* Critical Apparatus tagset (as outlined in Chapter 19 of the *TEI's Guidelines*) provides a method for capturing variants across a witness set. This highly structured encoding brings together in one document n number of witnesses which an editor considers the same work. The encoding enabled by parallel segmentation provides a typology for indicating what structural units of text, or parts of structural units, belong to each witness. In this way, content which appears in more than one version of the work is encoded once, with attribute values indicating which witness or witnesses it belongs to. It is an extremely efficient way of encoding in that the editor is saved the repetitious work of encoding the content which persists over multiple witnesses, as one would do if each witness were encoded as a separate document.

The apparatus element or <app> acts as a container element binding together the various readings, which are encoded within a reading <rdg> element. Attribute values indicate which witness or witnesses a particular structural unit (a paragraph or line, for example), or subunit, belongs to (See figure 1.).

```
<lg n="1">
<l n="1">
<app>
<rdg wit="a1 a2 a3 a4 pub">The sun burns
out,</rdg>
</app> </l>
<l n="2">
<app>
<rdg wit="a1">The world withers,</rdg>
<rdg wit="a3 a4">The world
withers,<milestone unit="stanza"/></rdg>
<rdg wit="a2 pub">The world
withers<milestone unit="stanza"/></rdg>
</app> </l>
```

Figure 1. A fragment of parallel segmentation encoding

When parsed in the *Versioning Machine*, the aforementioned fragment, the title of the text, along with the first few lines, is rendered as follows for the first three versions:



Figure 2: The title of 'Autumn' rendered in the Versioning Machine

In Lessard and Levison's 1998 article "Introduction: quo vadimus", they argue that computational humanities research has not achieved a level of acceptance because of the differences in "opposing intellectual paradigms, the scientific and the humanistic". The scientific, they argue, is based on formulation of hypotheses, collection of data and controlled testing and replication. The humanistic paradigm, they argue is based on argument from example, "where the goal is to bring the interlocutor to agreement by coming to see the materials at hand in the same light" (263).

While the *Versioning Machine* was designed as a visualization tool, it is no less importantly an environment within which editors realize a theory of the text, bringing readers to an understanding of the work as embodied in its multiple witnesses. It can thus be seen within Lessard and Levison humanistic paradigm, as a tool for presenting a reading of the work through its editing and encoding, itself a primary theoretical event (McGann 75). Moreover, this primary event can be illuminated and explicated through more traditional scholarly apparatus, such as annotation, adding an additional layer of textual analysis.

Thus the *Versioning Machine* provides a venue not only to realize contemporary editorial theory, but to challenge it. It meets the requirement that Stéfan Sinclair outlines in his 2003 article "Computer-Assisted Reading; Reconceiving Text Analysis" in that it is a tool which is relevant to literary critics' current approaches to textual criticism (178). The *Versioning Machine* is an active editing environment: it has been used by encoders editing texts as different as Renaissance plays and Dadaist poetry. The *Versioning Machine* is a tool which takes as its premise that the goal of much contemporary editing is not to create a definitive edition, but rather a "hypothesis" of the text (Kane-Donaldson as quoted in McGann 77), which can be read alongside an unedited edition of the text (that is, a reproduction of an image of the text in documentary form; McGann 77, Siemens). As such, it makes visible encoding as criticism, providing an environment to challenge our approaches to complex texts in terms of theories of encoding, as well as contemporary editorial theory.

## Bibliography

- Lessard, G., and M. Levinson. "Introduction: quo vadimus?" *Computers and the Humanities* 31.4 (1998): 261-269.
- McGann, Jerome. *Radiant Textuality: literature after the World Wide Web*. New York: Palgrave, 2001.
- Schreibman, Susan, Amit Kumar, and Jarom McDonald. "The Versioning Machine." *Literary and Linguistic Computing* 18.1 (2003): 101-107.
- Siemens, Ray. "'Unediting and Non-Editions' The Theory (and Politics) of Editing." *Anglia* 119.3 (2001): 423-455.
- Sinclair, Stéfan. "Computer-Assisted Reading; Reconceiving Text Analysis." *Literary and Linguistic Computing* 18.2 (2003): 175-184.
- Sperberg-McQueen, C.M., and L. Burnard, eds. *TEI P4: Guidelines for Electronic Text Encoding and Interchange*. Text Encoding Initiative Consortium, 2002. Accessed 2004-10-09. <<http://www.tei-c.org/P4X/>>
- Vetter, Lara, and Jarom McDonald. "Witnessing Dickinson's Witnesses." *Literary and Linguistic Computing* 18.2 (2003): 151-165.

## Electronic Text Analysis and a New Methodology for Canonical Research

### Matt Jockers

Using a combination of 'typical' text analysis tools (concordance and collocation) and other custom tools developed by the author, this paper demonstrates that conventional 'higher' criticism with its fashionable and thematically-oriented theoretical approaches fails as a means of assessing and generalizing about canons and genres of literature. Drawing on a case-study of the canon of Irish-American prose, the paper employs a quantitative and, indeed, scientific methodology to offer a radical reinterpretation of the canon.

In support of this research the author collected, coded, and categorized a database collection of prose literature including over 750 individual works written by some 280 different authors. The collection spans a period of 300 years and nearly being comprehensive in terms of its scope and coverage of the prose canon and genre of Irish-American ethnic literature. In addition to the usual metadata associated with electronic archives, each work in the collection is tagged with metadata related to the nature of the work: metadata includes geographic setting (East or West of the Mississippi), regional setting (Northeast, Southwest, Mountain, Pacific, and etc), information about whether the work is set in an urban or rural environment as well as data specific to the author of each text. Using his own *Corpus Analysis Tools Suite (CATools)*, a set of analytic

tools developed using php and mysql for doing both semantic and quantitative text-analysis of materials specifically housed within a relational database structure, the author has mined the material in order to reveal latent chronological, semantic, and geographic trends within the overall canon since its beginning in the late 18th century to the present.

The results of this work not only challenge the best available scholarship on the subject of Irish-American literature but further challenge the efficacy of contemporary and fashionable theoretical approaches to literature that are based on the 'close-readings' of texts. In making the case for a re-evaluation of the Irish-American canon, the paper challenges the basic and fundamental methodology of traditional literary study, and demonstrates in clear and indisputable terms that a quantitative and, indeed, scientific analysis of the literary data is not only valuable to the study of a genre or a canon of literature but essential if we are to ever go beyond the mere 'readings' and interpretations of texts.

## Response

Ray Siemens

---

# The *Virtual Lightbox* for Museums and Archives: A Distributed Solution for Structured Data Reuse Across Multiple Visual Resources

---

*Amy Smith* ([a.c.smith@reading.ac.uk](mailto:a.c.smith@reading.ac.uk))

*University of Reading*

*Brian Fuchs* ([fuchs@mpiwg-berlin.mpg.de](mailto:fuchs@mpiwg-berlin.mpg.de))

*Max Planck Institute for the History of Science, Berlin*

*Leif Isaksen* ([li103@soton.ac.uk](mailto:li103@soton.ac.uk))

*University of Southampton*

---

**T**he *Virtual Lightbox for Museums and Archives (VLMA)* is a tool for collecting and reusing distributed visual archives via RDF syndication and P2P technology. It aims to assist students and scholars in locating and exporting learning object/metadata groups that may then be reused and shared among users and user-groups.

The *VLMA* is a response to the challenge of contextualization. It grew out of the experience of digitizing a small university collection —the *Ure Museum of Classical Archaeology* (<http://www.reading.ac.uk/Ure/>) — and integrating its contents into a humanities portal —the *ECHO* project (<http://echo.mpiwg-berlin.mpg.de/>) — with diverse content. Research and teaching required access to external resources, in order to provide context and points of comparison, while integration into a large portal with diverse objects demonstrated the need for a means to assemble groups of objects from diverse sources in a manner independent of their online presentation. It also became apparent that the need to allow students to collect objects from diverse online sources for reuse in other contexts, e.g. presentations, as opposed to integration into a collection's website, could not easily be met in the current situation of online collections. Recent technological innovations, such as metadata unification (e.g. *Dublin Core*), distributed metadata (e.g. *Open Archives Initiative*), and meta-metadata (CIDOC's *Conceptual Reference Model*), had indeed done much to ease the difficulties involved in content integration; but in all of these cases, content integration has remained almost exclusively the province of the data provider, who is responsible for repackaging harvested



material. Thus, even when suitable metadata is available, collection and reuse of distributed content requires Herculean efforts on the part of the individual user. And because of the heterogeneous nature of the material collected, the individual user's efforts typically cannot themselves be rewarded with reuse in any significant fashion.

The *VLMA* seeks to complement other methods of content integration with a 'point-of-reuse' approach. Through this approach collections with intrinsically heterogeneous metadata sets are syndicated via RDF and then 'collected'—browsed, stored, viewed, and reused—at the peer/client level. The idea is to add to the current integration options available to collections publishers/users a 'distributed' variant, in which each peer determines its own strategy for metadata integration and content reuse. The problem of how to integrate collection metadata then becomes a question of reuse and syndication at the level of the individual user, rather than the provider, with content federation strategies flourishing or withering depending upon the current needs of the user community.

Syndication of content can take several forms—a lecture, a student paper, or actual resyndication across the network in the form of a new collection. The latter possibility is particularly exciting, as it provides an easy method for bringing added value to published content (for example, an online scholarly article that discusses related artworks in diverse collections might provide a unified set of illustrations rather than, as is necessary at present, laboriously providing links to the diverse websites on which these objects might be illustrated) as well as a simple way of creating thematically related collections with distributed content.

The *VLMA* method for content syndication is designed to be simple. A content producer seeds the network by syndicating already published content using a syndication tool which writes RDF to a lightbox namespace. The basic units of lightbox namespace are services, collection objects, and images, each of which is represented by an RDF fragment. In the simplest instantiation of a service, a consumer browses online objects in a collection, which s/he then captures to the lightbox. The lightbox then displays the images and metadata sets associated with this object, and "syndicates" them as a local collection, which appears in the service hierarchy alongside other collection browsers that have been discovered on the network. The consumer then has several reuse options, such as annotation, publication, export and local storage, which allow syndication with added value.

The *VLMA* is an open-source tool, written in Java under a GPL, and funded in its first phase (through March 2005) by *JISC*. A functioning prototype applet is available from the project's website (<http://www.reading.ac.uk/Ure/VLMA.htm>); <http://www.sourceforge.org/VLMA/>). A crucial feature of the current implementation is an RDF store

web service, which has been implemented with *Sesame* (<http://www.openrdf.org>), an open-source RDF database, as a backend. The use of such a service allows not only reduction in applet size but also significant latency reduction in RDF harvesting and querying, as visual collections clients typically access the same RDF material. The project has also agreed to pursue parallel development with Virtual Lightbox (<http://mith2.umd.edu/products/lightbox/>), developed at the Maryland Institute for Technology in the Humanities, and has designed its code in a modular fashion so as to be easily able to incorporate developments at MITH.

## Bibliography

- ECHO (European Cultural Heritage Online)*. European Commission. Accessed 2005-03-21. <http://echo.mpiwg-berlin.mpg.de/>
- Sesame*. openRDF.org. Accessed 2005-03-21. <http://www.openrdf.org/>
- Ure Museum of Classical Archeology*. University of Reading. Accessed 2005-03-21. <http://www.rdg.ac.uk/Ure/index.php>
- Virtual Lightbox*. Maryland Institute for Technology in the Humanities. Accessed 2005-03-21. <http://mith2.umd.edu/products/lightbox/>
- Virtual Lightbox for Museums and Archives*. Joint Information Systems Committee & Max Planck Institute for the History of Science. Accessed 2005-03-21. <http://www.reading.ac.uk/Ure/VLMA/>

## ***Callimachus*: A Virtual Archivist for Electronic Markup Projects**

---

**Jeff Smith** ([jeffs@smithicus.com](mailto:jeffs@smithicus.com))

University of Saskatchewan

**Joel Deshaye** ([joel.deshaye@usask.ca](mailto:joel.deshaye@usask.ca))

University of Saskatchewan

**Peter Stoicheff** ([peter.stoicheff@usask.ca](mailto:peter.stoicheff@usask.ca))

University of Saskatchewan

---

**T**he field of electronic text editing in the Humanities has been somewhat polarized along boundaries of preferred technology. Some feel that relational databases provide a more robust and powerful representation scheme while others cleave to the expressive power and transferability of XML.

The *Callimachus* project at the University of Saskatchewan was conceived as a way to have our cake and eat it too - merging the robust scalability of formal database technologies with the expressive power and Humanist-friendly accessibility of HTML and XML schema.

First applied to the hypertext edition of Faulkner's *The Sound and the Fury*, *Callimachus* delivers an easy-to-use schema development and implementation environment that eliminates the need for client-side editing, document management and revision control systems.

We propose to present our experiences with *The Sound and the Fury*, sharing the lessons learned and some of the surprising and unanticipated scholarly results that were achievable only with a system that allowed us to change our minds repeatedly and reconceive our schema as we moved along.

*Callimachus* was designed to allow the possibility of using a true database application to markup Faulkner's 1929 novel on a token (or word) level. Using free software tools such as *MySQL*, we built a Web interface with the database, thereby circumventing the need for client-side software, and enabling more than one editor to alter the database simultaneously. By storing each word of the novel in a separate record, we avoided the problems caused by imposing a strict hierarchy (like TEI) on a literary text. The new approach enables us to layer and overlap tags without fear of corrupting a strictly structured markup; we can also use conventional data-mining algorithms to reveal unforeseen relationships between tagged elements in the text.

These relationships are crucial in understanding any literary text that builds meaning through association and structure. Faulkner's *The Sound and the Fury* is a prime example. With this custom database and interface, we can discover when and how a concept appears in the novel. We can discover which characters dwell on what concepts and to what extent. We can discover how many words (how much narrative space) characters use when talking or thinking about specific topics. We can display relationships with charts and graphs computed with any combination of variables. And, to share our data, we can easily script the software to transform our database in conformance with the TEI or any other schema.

Before the invention of *Callimachus*, the first version of the hypertext edition of Faulkner's novel used HTML and JavaScript to visualize the complicated and apparently disordered narrative in the book's first chapter. Faulkner, telling this part of the story through the mind of an 'idiot', normally provides only obscure clues to mark the mnemonic flashing from one event to another. The narrative does not follow the chronological sequence of events in the novel. However, using HTML and JavaScript to tag each event, we built an interface that links events in the narrative sequence with events in a chronologically correct version of the text. For the first time, readers could reorient themselves in the chronology by clicking a button, leaving behind the much more confusing original narrative.

The hypertext edition helped us clarify our understanding of the novel and yielded some surprising results. We knew that the 'idiot', narrator, named Benjy, would relive an event (such as his grandmother's death), would trigger a sequence of flashbacks, and would often repeatedly return to that initial event. Benjy's memory of his grandmother's death is interrupted 17 times by other flashbacks. When we isolate this event from the interruptions, we notice that it is transmitted chronologically. Hidden in the chaos of so many relived events are small, coherent, chronological narratives.

This archive ( [www.usask.ca/english/faulkner](http://www.usask.ca/english/faulkner) ) was recently called "one of the best applications of the true potential of hypertext to date" (Neyt 140) . However, there were only a few surprises yielded through the approach of this early edition. In general, we knew what to expect; the manual markup in HTML and JavaScript made our innovative display possible, but without search functions or sophisticated computing, we had to draw graphs manually and show proportions based on manual word counts. At the *ACH / ALLC* conference in 2001, we saw the promise of the TEI and, very slowly, began rethinking our approach. The *Callimachus* archive structure is the result.

The point of *Callimachus* (named after the Greek poet and grammarian who was the chief librarian at Alexandria) is to free Humanities researchers from the burden of having to

specify their destination before starting their journey. And we do so in a way that allows each participating project to take advantage of analysis tools that might have been originally created for a different text.

In addition to providing a universally accessible web-based editing infrastructure, *Callimachus* offers powerful analysis tools: on-the-fly visualization to produce graphs of the relationships inherent in the text; data mining to help identify textual relationships that were not immediately apparent; and translation to allow the user to transform the text into arbitrary formats (such as HTML, TEI or other XML schema) for exchange with other parties.

*Callimachus* is designed to grow and adapt with the user, but without invalidating his or her previous work. The user does not have to learn another markup language or data formalization in order to begin exploring the text with state-of-the-art analysis tools. We prefer to leave the construction of hierarchical representations schemes until after we've learned what those relationships are, rather than presupposing what we are going to find in order to begin.

## Bibliography

Neyt, Vincent. "Review of Stoicheff, Muri, Deshayé, et al. (eds.): *The Sound and the Fury: A Hypertext Edition*." *Literary and Linguistic Computing* 19.1 (2004): 137-143.

---

## Integrating a Massive Digital Video Archive into Humanities Teaching and Research

---

*Lisa Spiro* ([lspiro@rice.edu](mailto:lspiro@rice.edu))

*Rice University*

*Diane Butler* ([diane@rice.edu](mailto:diane@rice.edu))

*Rice University*

*Chris Pound* ([pound@rice.edu](mailto:pound@rice.edu))

*Rice University*

---

**A**lthough the use of digital video in humanities teaching and research is growing, there are few studies that examine how it is used and what its impact is. One of the largest collections of digital video for use in education is the *Shoah Visual History Archive*, which was established by Steven Spielberg following the filming of *Schindler's List* to create a visual record of Holocaust survivors for use in education. The archive collects over 52,000 testimonies by survivors and witnesses of the Holocaust in 32 languages, yielding more than 110,000 hours of video. Over 38,000 of these testimonies have been digitized, and over 34,000 have been indexed using a set of more than 30,000 descriptive terms.

During the 2003-2004 academic years, Rice University, Yale University, and the University of Southern California jointly participated in a project funded by the Mellon Foundation to explore possible uses of the archive in research and teaching. Although the project has ended, all three universities continue their support of the archive. The team from Rice University (<http://shoah.rice.edu>) will report on applications of the archive in humanities teaching and research between 2003 and 2005 and cover three basic topics: how the archive was used in research projects in the humanities, how instructors adapted the archive for specific pedagogical purposes, and how the nature and content of the archive affected student engagement with their course material.

## Humanities research

**B**ecause it collects so much data and captures the first-person accounts of Holocaust survivors in a dynamic form, the Visual History Archive is a rich resource for research. Students have used the archive for research projects in topics ranging from rhetoric to violence and trauma. For instance,

three graduate students in anthropology presented their observations on working with the archive as part of a panel at the American Anthropological Association conference in December, 2004. Research studies include "On forms of the Other and of the chronotope in memories of Rus Czerwona" (Potoczniak), "Cultural logics of memorialization reflected in the Survivors of the Shoah archive" (Baum), and "History and memory: Turkish and Jewish accounts of communal life before and after WWII on the island of Rhodes" (Erkan). In addition, a medical ethicist is using the archive to examine the role of doctors and nurses in resistance activities.

Despite the archive's potential for research, the process of identifying and viewing relevant testimonies proved difficult for some, since the keyword structure is elaborate, the number of testimonies potentially overwhelming, the technology occasionally unreliable, and the length of each testimony (anywhere from an hour to 18 hours, with 2.5 hours being the average) daunting. Student and faculty experiences with the archive, as well as a usability study undertaken by a graduate student in psychology, yielded significant recommendations for improvements to the web interface and point to ways that video archives can be better integrated into education.

## Humanities pedagogy

No new courses were created in conjunction with the project. Rather, we set out to integrate the archive into existing courses and thereby assess its pedagogical implications in a broad context, investigating how humanities instructors would use a vast archive of digital video. Courses in religious studies, comparative literature, film studies, rhetoric, women and gender studies, and even classics employed the archive.

Participating faculty were interviewed to elicit their pedagogical vision. Most perceived that the archive offered an opportunity for students to work with primary sources and develop multimedia projects. But in each case, the instructors imagined specific reasons and practices for engaging with the archive. Our report includes several vignettes demonstrating divergent rationalities in humanities pedagogy, giving multiple meanings to a common digital resource. Among the assignments given to students were creating documentary videos based upon the Shoah materials, making presentations comparing pre- and post-War Jewish life as revealed in Shoah testimonies, and writing an essay analyzing two testimonies in relation to Samantha Power's recent book on genocide, *A Problem from Hell*.

## Emotions and intellectual engagement

The faces and voices in the Shoah archive are captivating. The survivors speak for as long as they want on any experience they care to recount, and every aspect of their testimony is preserved. The stories are emotionally charged, and the possibility that any detail of the testimony could acquire a sudden significance leads most viewers to attend to the matter carefully.

Our project team developed a survey to assess the emotional impact of the testimony on students, finding that students who responded emotionally to the video also reported a higher level of intellectual engagement with the course material. An important secondary finding in our assessment of student engagement was that students also responded well to unexpected particularities, atypical experiences, and pre-/post-Holocaust contexts represented in the archive, justifying the archive's attempt at comprehensiveness and its storage in a digital format to enable essentially random access. For many students, video proved to be more compelling than written sources, since they could see the facial expressions of the survivors and hear the tone of their voices.

## Bibliography

- Baum, Eric. "Cultural logics of memorialization reflected in the Survivors of the Shoah archive." *Paper delivered at the 2004 American Anthropological Association Annual Meetings, Atlanta, Georgia*. 19 December 2004.
- Potoczniak, Anthony. "On forms of the Other and of the chronotope in memories of Rus Czerwona." *Paper delivered at the 2004 American Anthropological Association Annual Meetings, Atlanta, Georgia*. 19 December 2004.
- Saka, Erkan. "History and memory: Turkish and Jewish accounts of communal life before and after WWII on the island of Rhodes." *Paper delivered at the 2004 American Anthropological Association Annual Meetings, Atlanta, Georgia*. 19 December 2004.
- Survivors of the Shoah Visual History Foundation*. Accessed 2005-04-06. <<http://www.vhf.org/>>

## *Early Modern Literary Studies:* Preparing for the Long Run

*Matthew Steggle* ([M.Steggles@shu.ac.uk](mailto:M.Steggles@shu.ac.uk))  
Sheffield Hallam University

*Early Modern Literary Studies (EMLS)*<sup>1</sup> is a peer-reviewed online journal publishing articles on all aspects of early modern literature. No registration or subscription is required, and it is available for free to anyone anywhere in the world with access to a web browser. All articles which are submitted to it undergo double-blind peer review, but those which are successful are usually published within less than a year of submission: a process much faster than comparable print journals. Since its foundation in 1994 *EMLS* has published over a hundred and fifty scholarly articles, and over two hundred and fifty reviews of books, films, plays, and multimedia products. In a typical week, its servers appear to record around 6,000 different readers in perhaps eighty different countries.

But although it is a veteran in terms of the internet, it is still a newcomer in a relatively slow-moving field where many of its rival print journals have pedigrees going back over a century. This paper reviews the progress of the journal since its foundation in 1995, and asks how a project like this one should be preparing for a long-term future.

The most obvious forms of this problem relate to questions of formatting and archiving. This paper will describe *EMLS*'s involvement with different forms of archiving system including the National Library of Canada and the Stanford University *LOCKSS* project<sup>2</sup>, a project to create multiple caches of the journal's contents at research libraries around the world. The paper will also review *EMLS*'s policies around file formats, principally proprietary formats, HTML, and XML, and problems around revision policies.

But equally important to the journal's long-term future are the systems for determining its current success or otherwise. In particular, methods for determining the success of a commercial website typically include raw number of hits recorded; revenue generated through subscriptions and advertizing; and sales resulting directly or indirectly from the site. Methods for determining the success of an academic journal typically include print run; citation of articles in it elsewhere; and peer recognition among leading experts in the field. By which of these sets of standards should an academic website seek to measure its success? Or must a new set of standards be developed to describe this activity? This paper details the results

of research into *EMLS*'s readership statistics (raw statistics online at <http://www.shu.ac.uk/emls/stats/>), addressing the question of what can and can't be deduced from them, before moving on to a consideration of other forms of esteem factor in terms of their effect in an institutional context.

If scholarly electronic publishing is to have a long-term future, it needs to be able to sustain publications over a scale of decades rather than merely years. This paper will conclude with recommendations for how to achieve longevity in an electronic publication.

1. <http://purl.oclc.org/emls/emlshome.html>
2. <http://lockss.stanford.edu/>

## Bibliography

Dyck, Paul, R.G. Siemens, Jennifer Lewin, and Joanne Woolway Grenfell. "The Janus-Face of Early Modern Literary Studies: Negotiating the Boundaries of Interactivity in an Electronic Journal for the Humanities." *Early Modern Literary Studies* 5.3 / *Special Issue* 4 4 (2000): 1-20. <http://purl.oclc.org/emls/05-3/dslwemls.html>

Heimpel, Rod. "Legitimizing Electronic Scholarly Publication: A Discursive Proposa." *Computing in the Humanities Working Papers* A.15 (October, 2000): . <http://www.chass.utoronto.ca/epc/chwp/heimpel2/>

Keller, Michael A., Vicky Reich, and Andrew Herkovic. "What is a library anymore anyway?" *First Monday* 8.5 (2003): . [http://www.firstmonday.org/issues/issue8\\_5/keller/index.html](http://www.firstmonday.org/issues/issue8_5/keller/index.html)

McCarty, Willard. "Changing Shape: The On-line Journal as a Scholarly Resource." (Panel abstract) DRH 97 Proceedings. St. Anne's College, Oxford. September 14-17, 1997.

Reich, Vicky, and David S. H. Rosenthal. "LOCKSS: A Permanent Web Publishing and Access System." *D-Lib Magazine* 7.6 (June 2001): . <http://www.dlib.org/dlib/june01/reich/06reich.html>

# Profiling Stylistic Variations in Dickens and Smollett through Correspondence Analysis of Low Frequency Words

---

*Tomoji Tabata (tabata@lang.osaka-u.ac.jp)*  
*Osaka University*

---

The aim of this paper is to present the result of a corpus-driven, quantitative analysis of the style of Dickens in comparison with the style of Smollett. The particular problem discussed is the differing distribution of *-ly* adverbs in the texts written by the two authors. By applying a multivariate stylo-statistics model, this study illustrates how sharply the two authors differ in their uses of adverbs as well as how texts are differentiated according to genre and chronology within authorial groups.

On the relationship between linguistic registers and adverbs, Biber et al. (1999, 541) present interesting findings from a large-scale corpus:

It is interesting to note that, overall, fiction ... uses many different descriptive *-ly* adverbs, although few of these are notably common (occurring over 50 times per million words). Rather, fiction shows great diversity in its use of *-ly* adverbs. In describing fictional events and the actions of fictional characters, writers often use adverbs with specific descriptive meanings.

In fact, *-ly* adverbs found in Dickens are quite diverse. In the 23 texts used in this study, the number of types amount to 1,728; Smollett employed 634 types. Among those, a few types are highly frequent, such as *really* and *certainly*, occurring more than one thousand times. Conversely, a large number of adverbs occur only once. Such *hapax legomena* include a few types which sound very much Dickensian, such as *evil-adverbiously*, *patientissamentally*, *Shakespeareanly*. Although the number of tokens of *-ly* adverbs account for only a little more than 1% of total word-tokens in the texts, the findings by Biber et al. suggest that *-ly* adverbs deserve special attention in stylistic study of fiction.

This study deals with a corpus of texts comprising Dickens' and Smollett's major works. Dickens' set includes fifteen 'serial fictions', six 'sketches', one 'miscellany', and one 'history'. Smollett's contains six 'fictions' and one 'sketch'. The total word-tokens in the corpus amount to 5.8 million, with the Dickens component containing 4.7 million tokens and the

Smollett component totalling 1.1 million word-tokens. The present project was initiated as a study based on a comprehensive collection, not a sample corpus, of texts by the targeted authors. Therefore, the imbalance in the number of texts as well as tokens is inevitable. However, due attention will be paid in the choice of variables to minimize a potential effect of the differences in the population of the two sets. All the texts in the corpus have been annotated with the POS tags, using Eric Brill's *Rule-Based Tagger* (also known as the *Brill Tagger*). Manual post-editing has been conducted to eliminate a number of ill-assigned tags.

In an early successful attempt at a computational description of literary style, Milic compared the style of Jonathan Swift with the writings of his contemporaries, with special reference to the relative frequencies of word-classes in the texts and to grammatical features such as seriation and connection. Cluett (1971 & 1976) adopted a similar approach to conduct a diachronic study of prose style across 4 centuries: from the 16th to the 20th centuries. Brainerd's works (1979 & 1980) are ambitious attempts to apply discriminant analysis to the question of genre and chronology in Shakespeare plays. Takefuta's approach to text typology, or register variation, is among the first to successfully employ factor/cluster analysis to the lexical differences between registers. His pioneering work, however, is not widely acknowledged because it was written in Japanese. Since Burrows (1987) and Biber (1988), it has become popular practice to employ multivariate techniques in quantitative studies of texts. Biber carried out *factor analysis (FA)* on 67 linguistic features to identify co-occurring linguistic features that account for dimensions of register variation. A series of research projects based on Biber's *Multi-Feature/Multi-Dimensional* approach have been successful in elucidating many interesting aspects of linguistic variation, such as language acquisition, ESP, diachronic change of prose style, and differences between conversational styles in British and American English, to give a handful of examples (Biber & Finegan; Conrad & Biber eds.).

The Biber model is one of the most sophisticated approach by far. Yet it is not without its critics. Nakamura (1995) raises a major objection. He argues that Biber's variables are "quite arbitrarily selected with no definite criterion and mixed levels" (1995, 77-86). Further, Sigley (1997) notes that almost half of Biber's 67 linguistic features are too rare in texts of 2,000 words.

Burrows (1987), on the other hand, applied a *Principal Component Analysis (PCA)* to the thirty most common words in the language of Jane Austen. The method demonstrates that differing frequency patterns in these very common words show significant differentiations among Austen's characters, and that the statistical analysis of literary style may lead not only to a deeper understanding of the novel itself but may also contribute to our deeper appreciation of it. In this use of a PCA, the

frequencies of common words are used as variables. The Burrows method seems to have higher replicability and feasibility; since it focuses on common words, most of the variables are frequent enough to produce stable statistical results. In addition, it does not require a multi-layered tagging scheme optimised for Biber's MF/MD approach.

A particular strength of the Burrows methodology is in testing cases of disputed authorship and national differences in the English first-person retrospective narrative, known as 'history'. Among the most successful applications are Burrows (1989, 1992 & 1996), Craig (1999a, b, & c). The Burrows approach or similar methodology has been applied to Bible stylometry. Some scholars like Linmans, Merriam, and Mealand use *Correspondence Analysis (CA)* instead of PCA. In the context of text typology, Nakamura (1993) applied CA to the frequency distribution of personal pronouns to visualize association between personal pronouns and 15 text categories in the LOB corpus.

My earlier work (Tabata) also used CA to analyse the distribution patterns of Part-of-Speech in Dickens's 23 texts and identified a contrast between serial fiction and sketches. The present study is different from the Burrows model in that it extends the range of variables to include low-frequency words, or rare words, by applying CA in the analysis of -ly adverbs. CA is one of the techniques for data-reduction alongside PCA and FA. Unlike PCA and FA, however, CA does not require intervening steps of calculating correlation matrix or covariance matrix, and can therefore process the data directly to obtain solution. CA allows examination of the complex interrelationships between row cases (i.e., texts), interrelationships between column variables (i.e., adverbs), 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 like Figures 1 to 4, relative distance between variable entries indicates affinity, similarity, association, or otherwise between them. One advantage CA has over PCA and FA is that PCA and FA cannot be computed on a rectangular matrix where the number of columns exceeds the number of rows, a concern of the present study. Yet CA can handle such types of a data table with, for example, the row cases consisting of thirty texts and the column variables consisting of hundreds of adverbs.

Figures 1 & 2 Correspondence Analysis of —ly adverbs in Dickens & Smollett: based on the commonest 1,278 types that appear in two or more texts

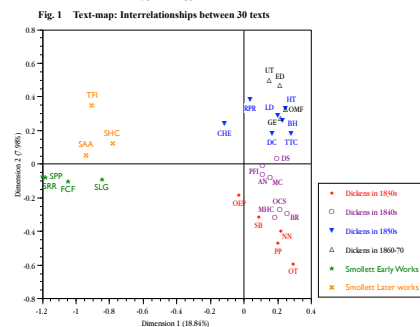


Figure 1: Correspondence Analysis of —ly adverbs in Dickens & Smollett based on 1,278 types that appear in two or more texts: Text-map showing interrelationships between 30 texts

Fig. 2 Word-map: Interrelationships between 1,278 —ly adverbs which appear in two or more texts

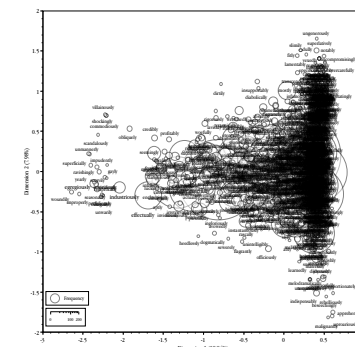


Figure 2: CA: Word-map showing interrelationships between 1,278 types of —ly adverbs

Figures 3 & 4 Correspondence Analysis of —ly adverbs in Dickens & Smollett: based on the commonest 99 types that appear in both Dickens and Smollett

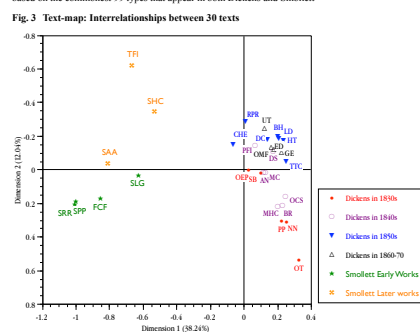


Figure 3: Correspondence Analysis of —ly adverbs in Dickens & Smollett based on the most common 99 types: Text-map showing interrelationships between 30 texts





Milic, L. T. *A Quantitative Approach to the Style of Jonathan Swift*. The Hague: Mouton, 1967.

Nakamura, J. "Text Typology and Corpus: A Critical Review of Biber's Methodology." *English Corpus Studies* 2 (1995): 75-90.

Nakamura, J. "Statistical Methods and Large Corpora: A New Tool for Describing Text Types." *Text and Technology: In Honour of John Sinclair*. Ed. M. Baker, G. Francis and E. Tognini-Bonelli. Amsterdam: John Benjamins, 1993. 293-312.

Sigley, R. "Text Categories and Where You Can Stick Them: A Crude Formality Index." *International Journal of Corpus Linguistics* 2.2 (1997): 199-237.

Tabata, T. "Investigating Stylistic Variation in Dickens through Correspondence Analysis of Word-Class Distribution." *English Corpus Linguistics in Japan*. Ed. T. Saito et al. Amsterdam: Rodopi, 2002. 165-182.

Takefuta, Y. コンピューターの見た現代英語: ボキャブラリーの科学 ('*The Computer Analysis of the Contemporary English Language: a quantitative study of vocabulary*'). Tokyo: Educa, 1981.

---

## Disciplined: Using Curriculum Studies to Define 'Humanities Computing'

---

*Melissa Terras* ([m.terras@ucl.ac.uk](mailto:m.terras@ucl.ac.uk))

*School of Library, Archive and Information Studies, University College London*

---

**H**umanities Computing remains an emergent discipline. Although it has been a full half-century since the work of Roberto Busa, and the activities described as 'Humanities Computing' continue to expand in number, sophistication, and scholarly importance, the field is continually changing, developing, and defining itself. Most introspection regarding the role, meaning, and focus of 'Humanities Computing' as a discipline has come from a practical and pragmatic perspective from scholars and educators within the field itself. This paper aims to appropriate techniques from Education and Curriculum Studies to provide an alternative, externalised, viewpoint of the history and focus of Humanities Computing, by analysing the discipline through its teaching programs and message they deliver, either consciously or unconsciously, about the focus and scope of the discipline.

It is now over thirty years since the Association of Literary and Linguistic Computing was founded (in 1973), and almost twenty years since the first issue of *Literary and Linguistic Computing* was issued in 1986.<sup>1</sup> Attempts have been made to discuss how Humanities Computing as a field should be defined (McCarty 1998, Unsworth), how it relates to other disciplines (McCarty 1999), and how it should support the activities with which it is associated, both on an institutional (McCarty et al. 1997) and scholarly level (Orlandi). Histories of and companions to the discipline have begun to emerge (Fraser 1996, Schreibman et al. 2004), from both research, scholarly, and institutional perspectives (Warwick). Attention to these issues is intrinsic to such a multi-disciplinary field; the emergence of related discussions was a major reason for the creation of the Humanist discussion list in 1987,<sup>2</sup> and associated issues continue to appear on its postings.

Humanities Computing may now be pursued as part of undergraduate degree courses, and there are now various graduate programmes in the humanities, in both European and North American Universities, in which a computing component has a significant role. Humanities Computing can therefore be seen to exist as an independent academic subject, with undergraduate and graduate students, specialist faculty and

research staff, and coherent systems of communication, publication, peer review, and funding criteria, as well as reflective historical and academic analysis which has been undertaken by practitioners in the field.

From the viewpoint of Educational Studies, teaching can be seen to be at the heart of a discipline (Maskell and Robinson 2001), and the curriculum, or "content of a particular subject or area of study" (Kelly 3) shapes the field. Moreover, the curriculum can be seen to define the field in the way the publication record cannot: it is the 'hidden' history of the subject, the core skill set practitioners have chosen to pass on to younger scholars, and describes the purposes of the transmission of knowledge content, and an exploration of the effects that exposure to knowledge is likely to have (Kelly). An awareness of the complexity of the relationship between content, application, and intention of the curriculum is necessary to create coherent teaching practices and develop a homogeneous definition of an academic discipline (Knight). Additionally, the writing styles and practices stressed and encouraged within a curriculum can in themselves define the field (Monroe). The curriculum can be taken, therefore, to define what a particular discipline represents (Becher et al.).

Discussion of the curriculum of Humanities Computing is not novel. Indeed, there was an entire conference devoted to The Humanities Computing Curriculum: The Computing Curriculum in the Arts and Humanities (2001), at Malaspina University College, Nanaimo, British Columbia, Canada.<sup>3</sup> Most papers necessarily described the practical aspects of setting up Humanities Computing programs and courses, and defining an overview of the contents of courses. For example, Gilfillan and Musick outlined the practicalities involved in promoting the use of computing in humanities based teaching and research at the University of Oregon, and Hockey examined the role of computing in the humanities curriculum at both postgraduate and undergraduate levels. Additionally, there was also a seminar series which was undertaken to define and generate a syllabus for a graduate course in knowledge representation for humanists at the University of Virginia, which resulted in a comprehensive syllabus for a Master's Degree in Digital Humanities (Drucker et al.). The accompanying report and proposed syllabus serves as a reference to those who may undertake the teaching of similar masters programs in the future. More generally, the Aco\*hum project produced a study on how Computing was, is, and could be used in Humanities subjects (de Smedt). These studies all serve to illustrate how important defining the curriculum is to Humanities Computing, and how, as a nascent subject, much is still being done to define the teaching program, and the field: although their focus is mostly (and necessarily) a practical approach to how teaching programs can be implemented and integrated into academic departments and scholarly frameworks.

However, from an Educational and Curriculum studies perspective, the term 'curriculum' applies not only to the content of a particular subject are of study, but refers to the total programme of an educational institution: being

the overall rationale for any educational programme, including those more subtle features of curriculum change and development and especially those underlying elements [explanation and justification] — which are the most crucial element in Curriculum studies.

(Kelly 3)

This paper will report on an analysis of the Humanities Computing curriculum from an Educational, and Curriculum, Studies holistic perspective, to be carried out between November 2004 and May 2005 in the department of Education and Professional Development, at University College London.

A study of Humanities Computing in this manner will be useful and interesting for a number of reasons. Firstly, as a nascent discipline, much of the documentation regarding the development of the curriculum is still available, and many practitioners and educators, who have seen the curriculum develop, are still working in the community, making access to such material relatively easy. Secondly, from the point of view of Curriculum Studies, it is quite rare to be able to study a field which is at this point of breakthrough into becoming an accepted academic discipline: compare this to more established academic subjects, where early development of the curriculum are all but lost. Thirdly, by focussing on the whole curriculum, an alternative viewpoint can be propagated as to what Humanities Computing is, and what it does. By rationalising the Humanities Computing curriculum in terms of Educational Theory, it may be possible to provide an alternative overview and definition of the discipline which is not merely limited to describing the content of courses, or programme syllabuses, but embraces the curriculum as a totality of purpose and content, including its formal, informal, planned, received, and hidden agendas.

This research will be carried out by undertaking an analysis of all available material gathered from the established teaching programs in the field (from both the UK and the USA), discussion lists, and educators, combined with surveys and interviews with leading practitioners (both teachers and researchers) and computational and content analysis of published, survey and interview material. Findings will be related to current theory and practice in Curriculum Studies. In doing so, it will be a sustained professional enquiry into the teaching and learning process of Humanities Computing, adopting the standard techniques from Curriculum Studies to analyse and understand the disciplinarity of the subject (Rowland 1993, Rowland 2000).

Questions asked in this study will include: can an analysis of the curriculum aid in defining Humanities Computing? How does the curriculum currently on offer differ from the research

agenda, as demonstrated through conference and publication records? Is there a common curriculum in existence between individual institutions and programs? How can the definition and rationalisation of the curriculum of this nascent discipline aid it in becoming entrenched in more traditional academic disciplines? How does the intention of Humanities Computing as a teaching discipline differ from the reality? What hidden implications and definitions are propagated about Humanities Computing through its curriculum? How does the role of computing in the discipline detract from the centralized control of the teacher necessary for steering the curriculum? What strategies for curriculum change, control, assessment, evaluation, appraisal, and accountability have been implemented in the Humanities Computing community? How do the writing styles promoted by Humanities Computing, through its curriculum, define and shape the field? If a common curriculum cannot be defined, does Humanities Computing as a subject really exist?

As this research will be carried out throughout late 2004 / early 2005, it is impossible to summarise its findings here: however, the fact that this is a novel and alternative approach to answering the perennial question "What is Humanities Computing?", this research should yield useful insights. As Kelly notes:

A study of curriculum, while not offering us spurious answers to questions of values, will... draw our attention to important questions that need to be asked about policies and practices and help us achieve the kind of clarity which will enable us to see underlying ideologies more clearly.

(19)

Viewing Humanities Computing from another perspective may aid us in defining and steering the direction of the discipline, whilst propagating a useful and alternative definition of the subject.

- 
1. The Association of Literary and Linguistic Computing published their journal twice yearly from 1980 to 1985, when this was merged with *ALLC bulletin* to become *Literary and Linguistic Computing* (1986).
  2. <http://www.princeton.edu/~mccarty/humanist/>
  3. The Humanities Computing Curriculum: The Computing Curriculum in the Arts and Humanities, Malaspina University College, Nanaimo, British Columbia, Canada. November 9-10, 2001. <http://web.mala.bc.ca/siemensr/HCCurriculum/>.

## Bibliography

- Becher, T., and P.R. Trowler. *Higher Education: A Critical Business*. Buckingham: Open University Press/SRHE, 2001.
- de Smedt, K., H. Gardiner, E. Ore, T. Orlandi, H. Short, J. Souillot, and J. Vaughan, eds. *Computing in Humanities Education, a European Perspective*. Bergen: University of Bergen, 1999. Accessed 2005-03-03. <http://helmer.aksis.uib.no/AcoHum/book/>
- Drucker, J., J. Unsworth, and A. Laue. *Final Report for Digital Humanities Curriculum Seminar*. Media Studies Program, College of Arts and Science, University of Virginia, 2002. Accessed 2005-03-03. <http://www.iath.virginia.edu/hcs/dhcs/>
- Fraser, M. *A Hypertextual History of Humanities Computing*. Oxford University, 1996. Accessed 2005-03-03. <http://users.ox.ac.uk/~ctitext2/history/>
- Fung, Glenn, and Olvi L. Mangasarian. "What is Humanities Computing and What is Not?" Talk delivered in the Distinguished Speakers Series of the Maryland Institute for Technology in the Humanities at the University of Maryland, College Park MD. 5 October 2000.
- Gilfillan, D., and J. Musick. "Wiring the Humanities at the University of Oregon: Experiences from Year 3." Paper delivered at The Humanities Computing Curriculum: The Computing Curriculum in the Arts and Humanities, November 9-10, 2001, Malaspina University College, Nanaimo, British Columbia, Canada. November 9-10, 2001. <http://web.mala.bc.ca/siemensr/HCCurriculum/>
- Hockey, D. "Towards a Curriculum for Humanities Computing: Theoretical Goals and Practical Outcomes." Paper delivered at The Humanities Computing Curriculum: The Computing Curriculum in the Arts and Humanities, November 9-10, 2001, Malaspina University College, Nanaimo, British Columbia, Canada. November 9-10, 2001. <http://web.mala.bc.ca/siemensr/HCCurriculum/>
- Kelly, A.V. *The Curriculum: Theory and Practice*. 4th ed. London: Paul Chapman, 1999.
- Knight, P.T. "Complexity and Curriculum: A process approach to curriculum-making." *Teaching in Higher Education* 6.3 (2001): 369-381.
- Maskell, D., and I. Robinson. *The New Idea of a University*. London: Haven Books, 2001.
- McCarty, Willard. *What is humanities computing? Toward a definition of the field*. Accessed 2005-03-21. <http://www.kcl.ac.uk/humanities/cch/wlm/essays/what/>

McCarty, Willard. "Humanities computing as interdiscipline. Is Humanities Computing an Academic Discipline?" Paper delivered at IATH, University of Virginia. 5 November 1999.

McCarty, Willard, L. Burnard, M. Deegan, J. Anderson, and H. Short. "Root, trunk, and branch: institutional and infrastructural models for humanities computing in the U.K." Panel session at the Joint International Conference of the Association for Computers and the Humanities and the Association for Literary & Linguistic Computing, Queen's University, Kingston, Ontario, Canada. 3-7 June 1997.

McCarty, Willard, and H. Short. *A Roadmap for humanities computing*. 2002. Accessed 2005-03-21. <<http://www.kcl.ac.uk/humanities/cch/allc/reports/map/>>

Monroe, J. "Introduction: The shapes of fields." In *Writing and Revising the Disciplines*. Ed. J. Monroe. Ithaca: Cornell University Press, 2002. 1-12.

Orlandi, T. *The Scholarly Environment of Humanities Computing: A Reaction to Willard McCarty's talk on The computational transformation of the humanities*. Accessed 2005-03-21. <<http://rmcisadu.let.uniroma1.it/~orlandi/mccarty1.html>>

Rowland, S. *The Enquiring University Teacher*. : Milton Keynes: Society for Research into Higher Education and Open University Press, 2000.

Rowland, S. "An interpretative approach to teaching and learning." *The Enquiring Tutor: Exploring the Process of Professional Learning*. Ed. S. Rowland. Learning: Falmer Press, 1993. 16-33.

Schreibman, S., R. Siemens, and J. Unsworth, eds. *A Companion to Digital Humanities*. : Blackwell Publishing, 2004.

Warwick, Claire. "No Such Thing as Humanities Computing? An analytical history of digital resource creation and computing in the humanities." Paper presented at ALLC/ACH 2004, Gothenburg. 2004.

---

## Finalizing the Multiple-Text Electronic *King Lear* for Use in the Classroom

---

*Stephanie F. Thomas*  
([activerreading@btinternet.com](mailto:activerreading@btinternet.com))  
*Sheffield Hallam University*

---

### Summary

As the teaching of Renaissance texts becomes more and more technologically enabled, it is even more significant that these technological enhancements are developed appropriately. Working with both lecturers and students, the Active Reading project has developed a number of different interfaces and tools for analyzing variants in multiple-text editions. The quarto and folio texts of *King Lear* are imposing in length alone, and for students to aptly demonstrate their understanding of the texts, it is important to create an appropriate learning environment. The most interesting element of the work appears to be how these interfaces or tools are being used actively in the classroom. By studying students' interactions with the online texts and recording their feedback, I have been able to form my own conclusions about the most useful ways of presenting a multiple-text electronic edition and adequately incorporating its textual variants. This paper will present the findings of these studies.

### Introduction

The Active Reading project is involved in developing an electronic scholarly edition of a Renaissance text illustrating the textual variants between published editions of that work. Two quartos and the folio text of *King Lear* have been selected for development in this way, allowing the editorial processes to be unravelled, and the Active Reading process to be encouraged through interactive involvement. In examining several paper-based editions of a work for textual variants, readers may become disoriented between the editions, and find comparisons difficult to make. In developing an electronic edition that combines all the versions of a text, it is possible to form an interactive resource for comparison of variants, and indeed for composing new editions of a text and taking on the role of editor.

Initially, a prototype combining all the editions of a short twenty-one-line poem was developed. This was encoded in XML, and XSL and JavaScript were employed in producing the interaction methods. Pilot studies were undertaken, looking at the ways in which readers interact with the electronic edition and how they compare variants. The results of the studies enabled the development of a considerably longer text, that of *King Lear*. Initially texts were encoded with a scheme developed specifically for the project, but TEI (Text Encoding Initiative) standards have since been adopted to allow for more simplified sharing and greater dissemination of the material.

Empirical studies into the use of the tool and its effect on the process of Active Reading have enabled refinements in development. These studies examine the ways in which readers actively compare variants of a text - through recording interactive involvement, and by observing the editorial decisions they make. Two sets of user groups were established so that the edition could be observed in use. The first set of users were several groups of undergraduate students on the BA English Studies degree, taking the course "Introduction to Poetry". These students were looking at the edition within the confines of what would be their usual seminar session on electronic resources in the computing labs. The second set of users were postgraduate students either from the MA English Studies (Renaissance Literature) degree, or research students completing their work on Renaissance themes. This second group of students were observed using the edition in the same computer labs, but not within any formal seminar session. The results from these studies provide useful feedback directly from the student target users, who will be using the edition again within their future coursework. Overall the project is largely experimental, exploring the ways in which the material could most effectively be displayed, and looking at the ways in which readers interact with the texts and the variants.

## Conclusions

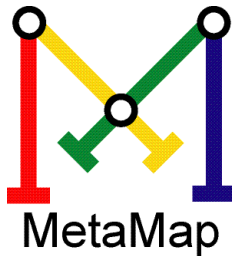
**T**his edition and the related research presents novel ways of comparing textual variants in editions of a Shakespearean text, and offers support for actively reading and understanding these texts. The edition has been used in teaching and as a learning tool, but is also intended to act as a template for the creation of future electronic editions. In designing and developing a new edition it has been helpful to be able to employ methods from the sphere of HCI (Human-Computer Interaction), and to understand the pedagogical requirements of the application, enhancing the experience for the reader.

In developing an electronic edition of this kind it is possible for the reader to compose new editions of a text, effectively taking on the role of editor. This paper aims to demonstrate the issues associated with developing the *King Lear* scholarly

edition. The findings presented will illustrate the advantage of using interactive technologies and text encoding tools to: a) provide a facility to examine textual variants independent of the text; b) allow active involvement in students' understanding of the text, and; c) offer greater insight into students' requirements when set the task of editing a multiple-edition text.

## The *MetaMap*, an Online Tool for Learning about Metadata

*James Turner* ([james.turner@umontreal.ca](mailto:james.turner@umontreal.ca))  
*Université de Montréal*



The *MetaMap* (<http://mapageweb.umontreal.ca/turner/>) is a graphic study aid and reference tool to help people understand how approximately 200 metadata initiatives worldwide are related to information management in digital libraries. It represents these in the form of a subway map to help the user navigate metaspace, and leans heavily on the conventions of the *London Underground Map* (London), noted for its clarity in helping users sort out complex reality. Each metadata standard, set, or initiative (MSSI) is represented as a station on a line. Each line has a theme, and these include processes of information management (Creation, Organisation, Dissemination, Preservation), institutions with expertise in information management (Libraries, Archives, Museums), and types of digital documentation (Text, Still Images, Moving Images, Sound). In addition, organisations deeply involved in Web activity and metadata norms, such as the *World Wide Web Consortium*, the *IETF*, and the *IEEE* are included on a separate line. When the user passes the mouse over a station name, the expanded name of the acronym appears. A mouse click on the name opens a window with other useful information such as the purpose of the MSSI, its sponsor, and links to the official and other useful Web sites. The *MetaMap* is available in English and French, and there is some information online in Spanish and Portuguese while these versions are being built. We are also negotiating with potential partners to build versions in several other languages.

The arrival of the World Wide Web and the new networked environment for information has already radically changed the approach to information management. This new information environment has no analogue in history and over the past ten years or so, it has been necessary to re-think the techniques and methods used for organising information. Various types of

metadata are being developed in response to various information management needs. Descriptive metadata is used for identification, discovery and access, and can also help in evaluating resources. Recordkeeping metadata helps to order, to validate, and to archive an organisation's resources and, with the arrival of electronic information is also considered "a tool that can help ensure the meaning, manageability and longevity of records and the information they contain" (New South Wales). In addition, preservation metadata plays the specific role of contributing to longterm conservation of digital resources. But whatever the particular reasons for which metadata is used, all types of metadata have in common the physical and intellectual management of resources to ensure access to them both now and in the long term. Basic readings which are helpful to those who wish to get a grasp on the concept of metadata include articles by Hodge and Hillman. The latter article is on use of the *Dublin Core* (DCMI). Without common rules and principles for metadata construction, the metadata and the corresponding resources would remain underused or would not be used at all (Soft Experience). This need for uniformity explains why the adoption of metadata standards quickly became so necessary.

The metadata standards and sets constructed to date already constitute a long list. As this list continues to grow, it becomes more and more difficult to keep track of this information, which forms the basis of the *Semantic Web*. In addition to standards and sets, a number of initiatives have been undertaken, often with a number of collaborators, in order to serve as testbeds for metadata sets and standards and to demonstrate the effectiveness of additional techniques for organising networked information. A number of major players assume the responsibility for several initiatives (e.g. *IETF* 2004, *IEEE* 2003, *W3C* 2004, *DCMI* 2004).

The idea of developing a tool such as the *MetaMap* arose from the idea that it would be useful to gather in a single place information about the many MSSIs that have come into existence over the last several years. Since the focus of attention in the information management community is the Web, and since the Web is the chief source for information about metadata for managing networked information, it was thought to be particularly helpful to produce a Web-based tool, in addition to a poster (in colour, French and English recto-verso, 90 x 60 cm) which is available free of charge. The *MetaMap* is sponsored by the *Groupe départemental de recherche en information visuelle* (GRIV 2003) at the Université de Montréal. Work on the *MetaMap* is funded by *CoRIMedia* (<http://www.corimedia.org>), a research consortium based at the Université de Sherbrooke.

## Bibliography

- dublincore.org*. Accessed 2004-11-21. <<http://www.dublincore.org/>>
- Hillman, Diane. *Using Dublin Core*. . Accessed 2001-04-12. <<http://www.dublincore.org/documents/2001/04/12/usageguide/>>
- Hodge, Gail. *Metadata made simpler*. . Accessed 2005-03-25. <[http://www.niso.org/news/Metadata\\_simple\\_r.pdf](http://www.niso.org/news/Metadata_simple_r.pdf)>
- ieee.org*. Accessed 2005-03-21. <<http://www.ieee.org/portal/index.jsp>>
- ietf.org*. Accessed 2005-03-14. <<http://www.ietf.org/>>
- London Underground: tube maps*. . Accessed 2005-03-21. <<http://tube.tfl.gov.uk/content/tubemap/default.asp>>
- mapageweb.umontreal.ca*. Accessed 2004-11-21. <<http://mapageweb.umontreal.ca/turner/francais/g Riv.html>>
- New South Wales Recordkeeping Metadata Standard*. . Accessed 2004-03-02. <<http://www.records.nsw.gov.au/publicsector/rk/rib/rib18-en.pdf>>
- Peccatte, Patrick. *Métadonnées: une initiation Dublin Core, IPTC, EXIF, RDF, XMP, etc.*. Soft Experience, 2004. <<http://peccatte.karefil.com/Software/Metadata.htm>>
- The MetaMap*. . Accessed 2005-01-08. <<http://mapageweb.umontreal.ca/turner/>>
- W3C: The World Wide Web Consortium*. Accessed 2005-03-21. <<http://www.w3c.org>>

## Visual Knowledge: Textual Iconography of the *Quixote*, a Hypertextual Archive

**Eduardo Urbina** ([e-urbina@tamu.edu](mailto:e-urbina@tamu.edu))

Texas A&M University

**Richard Furuta** ([furuta@cs.tamu.edu](mailto:furuta@cs.tamu.edu))

Texas A&M University

**Steven Escar Smith** ([ssmith@lib-gw.tamu.edu](mailto:ssmith@lib-gw.tamu.edu))

Texas A&M University

At present, there is no catalogue, in print or online, covering in a comprehensive manner the textual iconography of the *Quixote*. Attempts were made in 1879 and 1895 to offer a representative sampling but the coverage is very limited in both cases, amounting in the former to 101 illustrations from 60 selected editions, and to 23 plates from a single edition in the latter.<sup>1</sup>

Two key obstacles have prevented the publication of a comprehensive collection or archive based on the textual iconography of the *Quixote*: 1) the rarity of and difficult access to the materials, and 2) the technical and financial difficulties in compiling and disseminating such an archive in print format. The advent of hypertext, digital libraries, and the Internet, among other technological factors of the information technology revolution, make the impossible dream of visualizing the *Quixote* a realizable goal. Nevertheless, there remain still considerable obstacles and challenges if the result is to be both effective and valuable as an educational tool and a research resource in the humanities.

In this context, the *Cervantes Project (CP)* has under way the creation of a fully accessible, searchable and documented electronic database and digital archive of all the illustrations that form the textual iconography of the *Quixote*, as permitted by copyright limitations, along with the necessary interfaces and visualization tools to allow for the kind of access and study until now unavailable.<sup>2</sup> We further envision the archive as a research depository to complement the textual and bibliographical electronic resources already present in the *CP*, as well as a unique digital variorum image collection able to extend the value of our *Electronic variorum edition of the Quixote*, initiated in 1999. The archive will allow worldwide electronic access to these unique and rare textual and graphic resources by scholars, students, and users interested in

Cervantes' work and the influence of his masterpiece through 400 years from several perspectives: textual, artistic, critical, bibliographical, and historical.

In the last few years, critical interest on the illustrations of the *Quixote* has been renewed as demonstrated by the publication of three major monographs by J. Hartau, R. Paulson, and R. Schmidt.<sup>3</sup> Of equal significance is the recent exhibition at the Museo del Prado in Madrid entitled *Images of Don Quixote*, as well as the richly illustrated and documented catalogue of the exhibition prepared under the direction of Patrick Lenaghan, Curator of prints at the Hispanic Society of America in New York.<sup>4</sup> These studies and events place the illustrations in new and diverse cultural, aesthetic and historical contexts, demonstrating their key critical value and role in the reception and interpretation of the novel, and make evident the urgent need to provide a more complete and accessible resource to the rich artistic tradition of the textual iconography of the *Quixote* in order to better understand its significant contribution to the editorial history and critical reception of Cervantes' novel still largely unknown to readers and unexamined by critics.

The main rare book collection supporting our project is the *Cervantes Project (CP)* Collection at the Cushing Memorial Library and Archives of Texas A&M University. In recent years, the *Cervantes Project (CP)* and the Cushing Memorial Library have acquired a large number of significant illustrated editions for the purpose of creating a special collection of illustrated editions of the *Quixote*. At present (November 2004) the *Quixote* textual iconography collection includes 352 editions, published since 1620. The collection comprises over 650 volumes and is concentrated in 18th and 19th century English, French, and Spanish illustrated editions. We estimate the digital archive of the collection will eventually include upwards of 8,000 images and a fully searchable database complemented by innovative visualization tools.<sup>5</sup>

An important component of our initial work is the specification of a comprehensive taxonomy of the episodes, adventures, themes and characters in the *Quixote*. The taxonomy, representing the logical narrative structure of the work, will provide the addressing mechanism by which illustrations, texts, and other components can be associated with one another automatically. Through manipulation of the structure of the taxonomic elements and through specification of the desired interrelationships, hypotheses about the work can be posed and examined through coordinated inspection of text, illustration, commentary, and bibliography.

Specifically for the *Cervantes Project*, an XML schema, compatible with the TEI DTD, is being created representing the complex and highly significant interrelationships of episodes and adventures traceable throughout the entire text of *Don Quixote* as identified and tagged in our narrative taxonomy. In the first phase of the textual iconography project, one base text

of the *Quixote* is being fully encoded in TEI XML. Given the bilingual nature of our site and the international scope of its users, we next plan to encode J. Ormsby's English translation, already available in our digital library of electronic texts. Since this mark-up will include elements created by project staff to represent the various episodes, adventures, themes and motifs present in the narrative, these texts will provide an even richer searching opportunity for Cervantes scholars and will allow the subsequent incorporation of other key critical/textual editions.

The *Cervantes Iconography* project represents a close collaboration among four administrative groups on the Texas A&M University campus-Cervantes scholars based in the Hispanic Studies Department, professional staff members from the Texas A&M University Libraries, Information Science researchers from the Center for the Study of Digital Libraries (CSDL), and digital imaging specialists from the Texas A&M University Digital Library (TAMUDL), as diagrammed in Figure 1. Three separate representations of the collection are maintained at present-an edition master list (maintained in Hispanic Studies), a production collection (maintained in the A&M Libraries), and a research collection (maintained in the CSDL). The edition master list (Figure 2) serves as a master index and is imported by the two collections. As editions are digitized, the collected images are distributed simultaneously to both collections. The production collection maintains a set of Dublin-core-based metadata records, which are kept for each edition (Figure 3) and for each image of interest within each image (Figure 4). These metadata fields initially are populated from the master list and are augmented with additional information as the edition is catalogued, imaged, and associated with the comprehensive taxonomy. The research collection currently provides a Web-based "proofing" interface (illustrated in Figure 5 and accessible from our Web pages at <http://www.cSDL.tamu.edu/cervantes>) for the purpose of allowing quick browsing and searching of the growing collection.



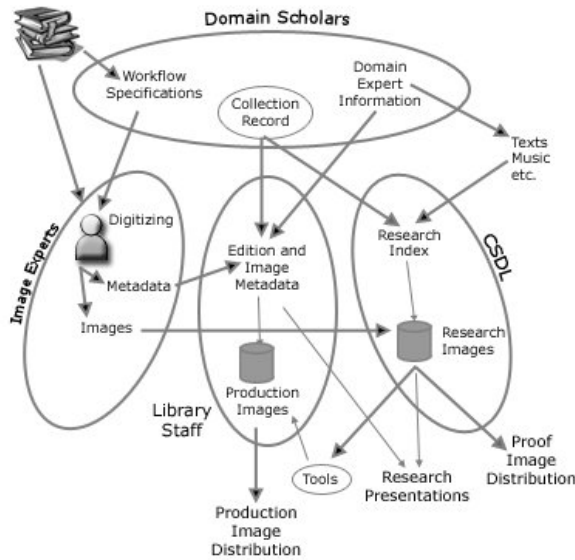


Figure 1

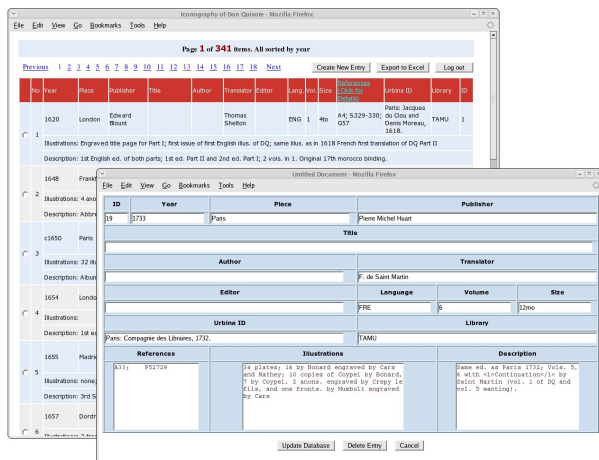


Figure 2

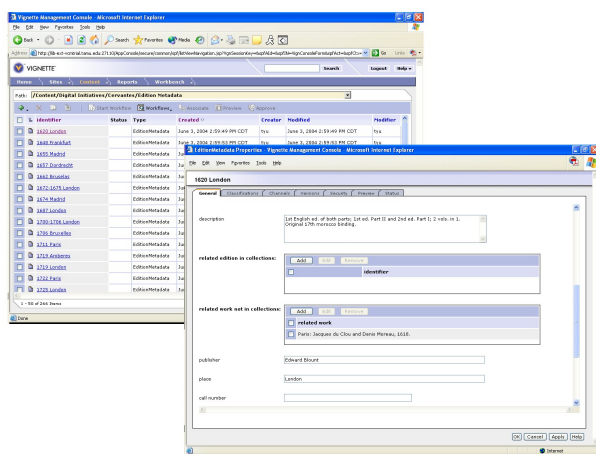


Figure 3

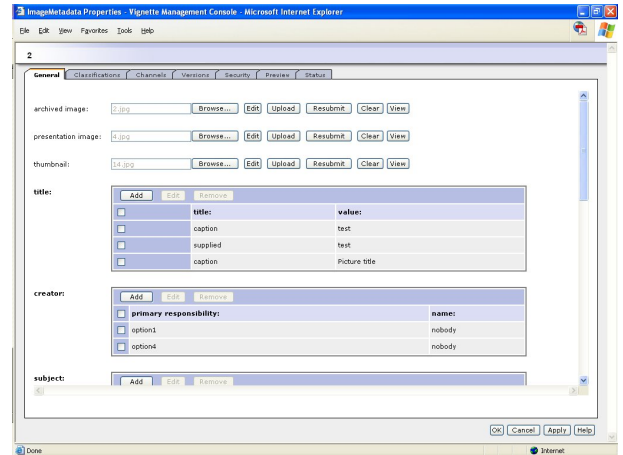


Figure 4

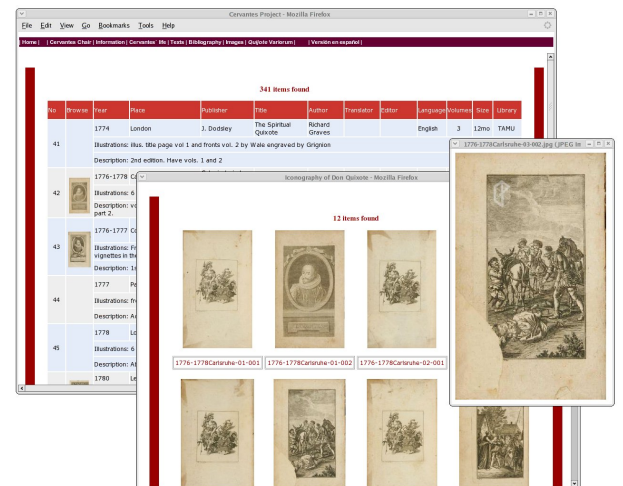


Figure 5

At the time of writing, the project has been ongoing for about a year and a half. The first year's activities focused on defining the metadata fields and imaging standards. The TAMUDL's imaging work began with test scans in Fall 2003, and currently is reaching production levels. To date, 32 editions have been imaged locally and another 15 obtained from other sources (the Library of Congress and the Spanish National Library), with a total of 1659 images currently available.

The added value of the illustrations in the *Quixote* textual iconography digital archive derive in particular from their innovative treatment and relationship with the collection of electronic text available already at the CP and in the linkages allowing connectivity between images, metadata, and bibliography entries. In addition, the archive provides interactivity between digital images and electronic texts, from different entry points. For example, one can browse single images, images with same metadata in a given field (content), sequential images from same edition, or all images related to a particular chapter or adventure by same artist or in the same

edition. We also plan to develop a tool to compare, juxtapose and collage related images from several editions, artists, etc., as part of our research to create new approaches and techniques to display images for analysis, beyond browsing and searching; in that connection we will expand archival description methods, and advance ways in which to integrate texts and images with metadata, as previously done for the images included in the electronic variorum edition of the *Quixote*.

The digital images, electronic databases, hypertextual archive, and visualizations tools to be created will be fully accessible free of charge through customized interfaces at the Internet portal of the *Cervantes Project*, as well as, in Spanish, through the portals of the National Library of Spain and the University of Castilla-La Mancha. The wide interdisciplinary interest in the *Quixote* throughout the centuries, its canonical and seminal status in the creation of the novel as a genre, its traditional inclusion in world literature courses, and its iconic status in Hispanic culture, are all factors insuring that the potential audience for the reference materials to be made available by our proposed project will be large, constant, and varied. It will include scholars in literary and book history interested in evaluating the reception and development of the written and visual text, students of the novel and of illustrations researching the role and function of iconography in narrative, and curious readers interested in seeing and appreciating for the first time a rich artistic tradition.<sup>6</sup>

- 
1. *Iconografía de Don Quijote; reproducción heliográfica y tipográfica de 101 láminas elegidas entre las 60 ediciones, diversamente ilustradas, que se han publicado durante 257 años...destinadas a la primera edición de Don Quijote* (Barcelona: P. Riera, 1879). *Catálogo de la exposición celebrada en la Biblioteca Nacional en el tercer centenario de la publicación del Quijote* (Madrid, 1905); *Exposición cervantina en la Biblioteca Nacional para conmemorar el CCCXXX aniversario de la muerte de Miguel de Cervantes Saavedra* (Madrid, 1946); Juan Givanel Mas, *Catálogo de la exposición de iconografía cervantina* (Barcelona, 1944).
  2. The *Cervantes Project* (CP) is an ongoing long-term project and research initiative dedicated to the development of a comprehensive digital archive based on the works of Miguel de Cervantes (1547-1616), the cornerstone of Hispanic letters and one of the world's most influential authors. In partnership with the Center for the Study of Digital Libraries and the Cushing Memorial Library and Archives, a division of the Texas A&M University Libraries, our goal is to create an online repository of textual, documentary, bibliographic, and visual electronic resources to serve the needs of students and scholars interested in Cervantes' life, times and work, and focused in particular on the study of *Don Quixote de la Mancha* ( <<http://www.csdl.tamu.edu/cervantes>> ).
  3. Johannes Hartau, *Don Quijote in der Kunst: Wandlungen einer Symbolfigur* (Berlin: Mann, 1987); Ronald Paulson, *Don Quixote*

*in England: the Aesthetics of Laughter* (Baltimore: The Johns Hopkins UP, 1998); Rachel Schmidt, *Critical Images: The Canonization of Don Quixote through Illustrated Editions of the Eighteenth Century* (Montreal: McGill-Queen's UP, 1999).

4. Patrick Lenaghan, together with Javier Blas y José Manuel Matilla, *Imágenes del Quijote: Modelos de representación en las ediciones de los siglos XVII a XIX* (Madrid: Hispanic Society of America-Museo Nacional del Prado-Real Academia de Bellas Artes de San Fernando, Calcografía Nacional, 2003). See also the catalogue prepared for another related exhibition, *El Quijote ilustrado: Modelos de representación en las ediciones españolas del siglo XVIII y comienzos del XIX* (Madrid: Ministerio de Educación, Cultura y Deporte-Real Academia de Bellas Artes de San Fernando, 2003), which visited Texas A&M University in March-April 2004 during the celebration of Spain Week.
5. Carlos Monroy et al. "Texts, Images, Knowledge: Visualizing Cervantes and Picasso". *Proceedings Visual Knowledges Conference*. John Frow, ed. University of Edinburgh: Institute for Advanced Studies in the Humanities, 2003. <<http://webdb.ucsd.edu/~malts/other/VKC/dsp-all-papers.cfm>> .
6. We provide a more complete acknowledgement of the many participants in this project at <<http://www.csdl.tamu.edu/cervantes/V2/CPI/images/iconography-acks.html>>

## Bibliography

*Catálogo de la exposición celebrada en la Biblioteca Nacional en el tercer centenario de la publicación del Quijote*. Madrid: Biblioteca Nacional, 1905.

*Cervantes Project*. <<http://www.csdl.tamu.edu/cervantes>>

*csdl.tamu.edu*. Accessed June 6, 2002 3:08:05 PM. <<http://www.csdl.tamu.edu/cervantes/V2/CPI/images/iconography-acks.html>>

*El Quijote ilustrado: Modelos de representación en las ediciones españolas del siglo XVIII y comienzos del XIX*. Madrid: Ministerio de Educación, Cultura y Deporte-Real Academia de Bellas Artes de San Fernando, 2003.

*Exposición cervantina en la Biblioteca Nacional para conmemorar el CCCXXX aniversario de la muerte de Miguel de Cervantes Saavedra*. Madrid: Biblioteca Nacional, 1946.

Hartau, Johannes. *Don Quijote in der Kunst: Wandlungen einer Symbolfigur*. Berlin: Mann, 1987.

Lenaghan, Patrick, Javier Blas, and José Manuel Matilla. *Imágenes del Quijote: Modelos de representación en las ediciones de los siglos XVII a XIX*. Madrid: Calcografía Nacional, 2003.

Mas, Juan Givanel. *Catálogo de la exposición de Iconografía cervantina*. Barcelona: Biblioteca Central, 1944.

Monroy, Carlos, et al. "Texts, Images, Knowledge: Visualizing Cervantes and Picasso." *Proceedings of the Visual Knowledge Conference*. Ed. John Frow. University of Edinburgh Institute for Advanced Studies in the Humanities, 2003. Accessed 2003. <<http://webdb.ucs.ed.ac.uk/malts/other/VKC/dsp-all-papers.cfm>>

Paulson, Ronald. *Don Quixote in England: the Aesthetics of Laughter*. Baltimore: The Johns Hopkins UP, 1998.

Riera, P. *Iconografía de Don Quixote; reproducción heliográfica y foto-tipográfica de 101 láminas elegidas entre las 60 ediciones, diversamente ilustradas, que se han publicado durante 257 años...destinadas a la primera edición de Don Quijote*. Barcelona, 1879.

Schmidt, Rachel. *Critical Images: The Canonization of Don Quixote through Illustrated Editions of the Eighteenth Century*. Montreal: McGill-Queen's UP, 1999.

---

## Mining the Differences between Penninc and Vostaert

---

**Karina van Dalen-Oskam**

([karina.van.dalen@niwi.knaw.nl](mailto:karina.van.dalen@niwi.knaw.nl))

Dept. Dutch Linguistics and Literary Studies

**Joris van Zundert**

([joris.van.zundert@niwi.knaw.nl](mailto:joris.van.zundert@niwi.knaw.nl))

Dept. Dutch Linguistics and Literary Studies

---

**T**he Middle Dutch *Roman van Walewein* (Romance of Gauvain, ca. 1260) was written by two authors, Penninc and Vostaert. Only one manuscript containing the complete text, explicitly dated as copied in the year 1350, is left to us. Some fragments of another, probably somewhat younger manuscript contain about 400 lines. The text in the complete manuscript consists of 11,202 lines of rhyming verse. The manuscript was written by two clerks. The first seems to have written the lines 1-5.781 and the second the lines 5,782-11,202.

The second author, Vostaert, explicitly claims to have added about 3,300 lines to Penninc's text. Because scholars of Middle Dutch literature came up with other amounts, we decided to try out modern authorship attribution techniques to find out whether these would point to a specific line in the text where the text before and the text after contrasts most. We used a lexical richness measure, Udney Yule's Characteristic K, and Burrows's Delta, measuring the differences of frequencies of the most frequent words in different parts of the text. We split the text into largely overlapping parts of 2000 lines, moving through the text in order to search for an exact line in the text where the contrast before and after would be the most significant. For measuring Burrows's Delta this meant that for the sake of our focus on one text (or two, in a way), we considered the text as a group of texts' and every part' of 2000 lines as a separate text, to be compared with the other 'texts'.

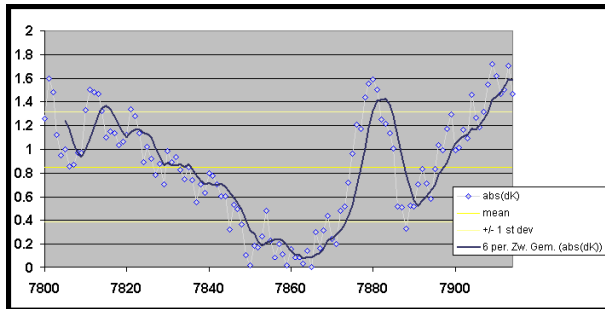


Figure 1: Lexical Richness according to Yule's K.

At the conference in Gothenburg in 2004 we were able to show that both measures yielded the lines 7,881-2 as the point of the most contrast. In Fig. 1 we present the results of Yule's K for that part of the text and in Fig. 2 the results of our creative use of Burrows's Delta can be found. It is very intriguing that both measurements point to the same place in the text. This suggests that line 7,882 could very well be the place where Vostaert took over from Penninc.

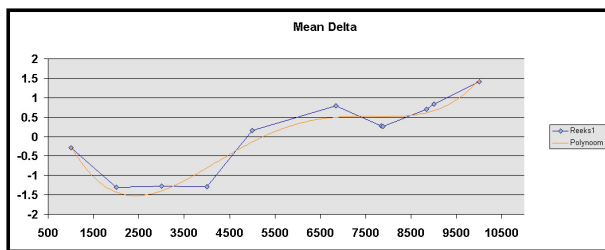


Figure 2: Differences in frequencies of the 150 most frequent words according to Burrows's Delta

We continue our research by concentrating on a quantitative analysis of the differences between the two parts of the text. What are in fact the lexical differences between the text parts before and after line 7,881-2? To find out, we made a list of lemmata (headwords, comprising all spelling variants or inflections etc. of a word) that occur significantly more in the lines before and in the lines after. The top of this list looks as follows:

		stdev	>0.05242999
		mean	0.0166
	<b>Penninc</b>		<i>z-score</i>
<i>be, his</i>	zijn	0.8413	15.7293
<i>I</i>	ik	0.8042	15.0217
<i>me</i>	mij	0.6790	12.6328
<i>you</i>	gij	0.5059	9.3325
<i>my, mine</i>	mijn	0.4223	7.7364
<i>may</i>	mogen	0.3158	5.7060
<i>it</i>	het	0.2957	5.3222

<i>stand</i>	staan	0.2665	4.7663
<i>we</i>	wij	0.2514	4.4775
<i>lord</i>	heer	0.2328	4.1224
<i>that</i>	dat	0.2195	3.8692
<i>yonder</i>	gene	0.2137	3.7587
<i>your</i>	uw	0.2131	3.7465
<i>you</i>	u	0.2095	3.6793
<i>say</i>	zeggen	0.2022	3.5387
<i>god</i>	god	0.1903	3.3124
<i>live</i>	leven	0.1774	3.0663
<i>come</i>	komen	0.1702	2.9290
<i>need</i>	moeten	0.1653	2.8359
<i>gate</i>	poort	0.1650	2.8300
<i>see</i>	zien	0.1599	2.7316
<i>squire</i>	knaap	0.1524	2.5898
<i>then</i>	doe	0.1485	2.5157
<i>give</i>	geven	0.1485	2.5150
<i>well, rather</i>	wel	0.1479	2.5043
<i>over</i>	over	0.1474	2.4931
<i>king</i>	koning	0.1454	2.4555
<i>thus</i>	dus	0.1396	2.3445
<i>stay</i>	blijven	0.1392	2.3375
<i>inside</i>	binnen	0.1267	2.0992
<i>not</i>	ne	0.1229	2.0275
<i>at</i>	aan	0.1147	1.8707
<i>shall</i>	zullen	0.1038	1.6623
<i>you</i>	jij	0.1034	1.6550
<i>loyal</i>	trouw	0.1011	1.6111
<i>go</i>	gaan	0.1009	1.6075
<i>serpent</i>	serpent	0.0958	1.5093
<i>allow</i>	laten	0.0954	1.5030
<i>desire</i>	begeren	0.0915	1.4280
<i>day</i>	dag	0.0878	1.3569
<i>where</i>	waar	0.0821	1.2481
<i>all</i>	al	0.0807	1.2211

		stdev	0.03920838
		mean	0.0167
	<b>Vostaert</b>		<i>z-score</i>
<i>the, this</i>	die	0.6234	15.4755

<i>he</i>	hij	0.4112	10.0614
<i>to</i>	te	0.3670	8.9353
<i>knight</i>	ridder	0.3659	8.9071
<i>large</i>	groot	0.3406	8.2613
<i>duke</i>	hertog	0.3051	7.3573
<i>very, pain</i>	zeer	0.2951	7.1002
<i>they, she</i>	zij	0.2886	6.9355
<i>Walewein</i>	walewein	0.2823	6.7757
<i>there</i>	daar	0.2748	6.5846
<i>so, thus</i>	zo	0.2260	5.3397
<i>of</i>	van	0.2242	5.2924
<i>Isabele</i>	isabele	0.1844	4.2767
<i>maiden</i>	jonkvrouw	0.1813	4.1977
<i>hit, slay</i>	slaan	0.1607	3.6728
<i>in</i>	in	0.1382	3.0998
<i>horse</i>	hors	0.1349	3.0160
<i>how</i>	hoe	0.1348	3.0117
<i>self</i>	zelf	0.1334	2.9774
<i>other</i>	ander	0.1330	2.9662
<i>fox</i>	vos	0.1228	2.7068
<i>no</i>	geen	0.1196	2.6245
<i>to</i>	toe	0.1171	2.5612
<i>man</i>	man	0.1131	2.4601
<i>many</i>	menig	0.1074	2.3153
<i>black</i>	zwart	0.1023	2.1845
<i>also</i>	ook	0.0985	2.0859
<i>begin</i>	beginnen	0.0980	2.0739
<i>because</i>	want	0.0969	2.0465
<i>brave</i>	stout	0.0961	2.0252
<i>speak</i>	spreken	0.0957	2.0155
<i>to</i>	tot	0.0942	1.9779
<i>helmet</i>	helm	0.0925	1.9352
<i>(some)one</i>	men	0.0918	1.9169
<i>sweet</i>	lief	0.0912	1.9009
<i>on</i>	op	0.0910	1.8953
<i>blood</i>	bloed	0.0884	1.8290
<i>and</i>	en	0.0873	1.8027
<i>walk</i>	lopen	0.0852	1.7485
<i>merciful</i>	goedertieren	0.0820	1.6672

<i>hour</i>	stonde	0.0812	1.6466
<i>do</i>	doen	0.0804	1.6262

[etc.]

Summarizing, Penninc makes significantly more use of the first and second person of the personal pronoun, in contrast to a significantly higher use of the third person by Vostaert. Penninc also applies a lot more modal verbs. But why? Are there several reasons for these differences, or can all be explained by only one or two 'special effects' of the individual authors?

The first hypothesis we will explore is that a difference in the amount of *dialogue* between the two parts of the text may give rise to several of the differences we have found. The paper will investigate whether this is the case. We will present an analysis of the vocabulary of both authors differentiating between dialogue, narrator's text, and *erlebte Rede*' (narrated monologue). We will also list other possibly differentiating elements and test whether these play a part in the contrast we discovered by using Yule's K and Burrows's Delta. This qualitative phase in the research is meant to yield an overview of elements contributing to the (quantitative) contrast on the one hand, and to lead us to a list of key elements in the lexicon of the two authors on the other. The list of actual differences will be the input for a new quantitative and qualitative literary analysis of the character and voice of Penninc and Vostaert. Furthermore, we will look forward to the next purely quantitative step we hope to take, in which the results of the above can help us to establish a formula for authorship distinction in the genre of Middle Dutch Arthurian Romance, and help us, so to speak, to leap from the mining to the modelling of the differences.

## Bibliography

- Burrows, J. "'Delta': a Measure of Stylistic Difference and a Guide to Likely Authorship." *Literary and Linguistic Computing* 17 (2002): 267-287.
- Burrows, J. "Questions of Authorship: Attribution and Beyond." *Computers and the Humanities* 37 (2003): 5-32.
- Es, G.A. van, ed. *De jeeste van Walewein en het schaakbord van Penninc en Pieter Vostaert*. 2 vols. : Zwolle, 1957.
- Holmes, D.I. "Authorship Attribution." *Computers and the Humanities* 28 (1994): 87-106.
- Johnson, D.F., and G.H.M. Claassens, eds. *Dutch Romances I: Roman van Walewein*. Trans. D.F. Johnson and G.H.M. Claassens. Cambridge: Cambridge, 2000.

Love, Harold. *Attributing Authorship: An Introduction*.  
Cambridge: Cambridge, 2002.

## The *e-Laborate* Project and the Usability of Another Textual Paradigm

---

**Joris van Zundert**

([joris.van.zundert@niwi.knaw.nl](mailto:joris.van.zundert@niwi.knaw.nl))

Dept. Dutch Linguistics and Literary Studies

**Karina van Dalen-Oskam**

([karina.van.dalen@niwi.knaw.nl](mailto:karina.van.dalen@niwi.knaw.nl))

Dept. Dutch Linguistics and Literary Studies

---

In 2003 we embarked upon the project *e-Laborate: a digital platform for partnerships in the humanities and social sciences*. The web application (at <http://www.e-laborate.nl> /> ) resulting from this project is intended as a virtual workplace for researchers in the humanities and social sciences. The *e-Laborate* collaboratory contains text collections, collections of statistical data and basic content management tools for sharing and working on text material and datasets. The project allows individual researchers as well as research groups to explore the potential of the collaboratory and to generate feedback. The tools enable users to expand the collection of material continuously and to improve its quality. In our paper we will present *e-Laborate* as an on line research collaboratory and as a web enabled tool for editing and analysing textual content. We will also show how *e-Laborate* provided a research environment in which we can explore the usefulness and usability of a specific text paradigm.

The text material we used in our project issued from the historical cultural journal "Vaderlandsche Letteroefeningen". The title means "National Literary Exercises" and in academic writing is usually shorthanded as "VLO". Published between 1761 and 1876, the "VLO" is of great importance for every research discipline concerned with the study of culture in the Netherlands during that period. There has long been a widely held desire to see a complete set of the journal available in digital form. However, because of its huge size and the enormous costs that digitisation would entail this has not been possible before now. The approach we have chosen differs fundamentally from the way in which textual material has usually been digitised and published in the past. The "VLO" component of the *e-Laborate* project uses a bottom up collaborative approach, drawing upon the assistance of researchers, to produce a continuous developing and evolving digital version of the publication. Using this approach NIWI

will now be able to publish facsimiles (scans) of the first 50.000 pages of "VLO" editions by the first quarter of 2005.

We will describe the development process used in building *e-Laborate*. The *eXtreme Programming* protocol (XP) was closely followed. Researchers' demands concerning the texts were closely monitored during the project and used to drive the development of the electronic tools for joint working on text and textual material. Every two weeks new elements were delivered, tested and approved of or commented on. Also critique and additional wishes were communicated with the developers. In this way we made sure that the tools would really be what researchers collaboratively working on text wanted and needed. The participating researchers are enthusiastic about this development approach and about the tools delivered up till now. Formal evaluation and retrospection showed especially appreciation for the pragmatically forward looking vision of the project (i.e. building the collaborative brick by brick, feature by feature).

The paper will provide a functional and architectural overview of *e-Laborate* as a collaborative tool for supporting the production of digital editions. At the core of *e-Laborate* is the *transcription object*. The transcription object is a container object holding the scanned image of a page from an original publication and a transcription field. Each transcription object's authorisation may be tailored by its creator / owner. Depending on the user's authorisation the transcription field of a transcription object is depicted either as a text edit box or as rendered text. Arbitrary additional metadata may be added. In the case of the "VLO" a standard id field is added to hold the number of the year, volume and page the scan shows. Standard content management utilities available within the *e-Laborate* platform allow for the arbitrary placing and grouping of individual transcription objects into a page or folder hierarchy. Any transcription object is automatically indexed so an authorised user or editor can search through the text base and present the search results in a comprehensive way. A fuzzy matching algorithm amends search input as well as the indexed material for spelling variants. In the future tools to further process or statistically analyse those results may be added. The addition of modules, tools, or components to *e-Laborate* is easily facilitated by its plain plugin architecture and open source nature. Current additions under development are the inclusion of an open source OCR engine to facilitate text recognition on demand for uploaded scans.

Current work in the project is focused on the development of a flexible annotation tool. This tool will empower researchers to create annotations to every part of a scan or the transcription text of a transcription object, simply by pointing to and highlighting the image part or text range they desire to annotate. Researchers will also have the possibility to react to annotations by annotating the annotation (*ad infinitum*). A researcher may

choose to categorize his or her annotation using a standard or personalised typology of annotations. Standard annotation typologies that will be provided concern a.o. basic formatting (italic, bold, capitalization etc.), ranges of interpretation (word, part of the text etc.) and information type (back ground historical information, biographical etc.). Any annotation may be categorised in multiple typologies.

The annotation tool will be as much WYSIWYG as possible. This means that a researcher wanting to add annotations will not be bothered by laborious tagging and will need no prior knowledge of any particular mark up language. This is a design choice fundamental to our view of text and textual research. We think that it's not a researcher's concern to produce or validate XML or any other marked up form of text. Knowing about mark up is not fundamental to the task of a text researcher, but making inferences about the meaning, structure and form of a text and putting such inferences into annotations is. Therefore tools for the production and enrichment of digital editions should focus on that research related task and not on mark up particulars. As a consequence the digital editing tools of *e-Laborate* will take care of the creation of valid mark up 'in the background', providing ample information about the name of the user who created the annotation, the date and time of creation, the part of the text or scan the annotation belongs to, and of course, the annotation text itself and any additional metadata provided by the user.

Elementary for our project are the leading principals behind the design choices described in the preceding paragraph. That is, the design choice not to define yet another mark up solution, but to concentrate on the researcher's interactions with the textual material, leaving the description of these interactions in the form of XML to the application. We will show that these principals define another textual paradigm, meaning *another* textual paradigm than the text paradigm implicitly emanating from the concepts of *TEI*.

At present a powerful surge of *TEI*-driven edition projects, seems to have propagated *TEI* into a de facto standard. Although undeniably useful as a means for marking up texts for editorial use, the apparent all round applicability and efficiency of *TEI* needs to be contested. We will argue that *TEI* in it's form of explicit mark up is not a very efficient means of editorial mark up. We will also argue that *TEI* is far from efficient nor very useful when computer supported textual analysis is the focus of research. We will show that the use of *TEI* forces an a priori, top down view of text onto a researcher trying to model a text using *TEI*-tagging. *TEI*'s particular use of XML and its DTD implicitly present a vision of a text being a flat hierarchy of meaningful text elements. To a researcher wanting to express and analyse overlapping interpretations, associative relations, layered narratives (to name but a few common textual constructs *TEI* has difficulty expressing) *TEI* does not provide effective

or efficient solutions. We will argue that such a researcher would be better off considering the use of lightly embedded mark up solutions and layered cross tagged mark up as provided for example by the *JITT* and *LMNL* models. Although problematic in themselves, these models do address the non linear, non hierarchical nature of texts more adequately than *TEI*. We will also argue how these models can be combined to provide an intuitive way of structuring and annotating text, resulting in a dynamic layered model of text that can be represented by proper XML. We will show how within the context of *e-Laborate* a graphical user interface enables structuring and annotating texts according to this dynamic model of text representation. We are convinced that this interface enables a researcher to interact with a text on a research and interpretative level rather than a mark up level. We will also show that in such a dynamic research environment it is still possible to provide backward compatibility with *TEI* mark up using transformational languages.

*SURF*. Accessed 2004-11-20. <<http://www.surf.nl/en/home/index.php>>

*TEI and TEI-Consortium*. Accessed 2004-11-20. <<http://www.tei-c.org/>>

van Dijk, S. "Introduction." *'I have heard about you'. Women's writing crossing borders*. Ed. S. van Dijk, P. Broomans, J.F. van der Meulen and W.R.D. van Oostrum. Hilversum: Verloren, 2004. [Information about the "VLO".]

*Women Writers*. Accessed 2004-11-20. <<http://www.roquade.nl/womenwriters/>>

*XML and the World Wide Web Consortium*. Accessed 2004-11-20. <<http://www.w3c.org/XML>>

*Xpast*. Accessed 2004-11-20. <[http://www.e-laborate.nl/nl/new\\_2/toon](http://www.e-laborate.nl/nl/new_2/toon)>

## Bibliography

Agosti, M., I. Ferro, I. Frommholz, and U. Thiel. "Annotations in Digital Libraries and Collaboratories." *Proceedings of the 8th European Conference, EDCL 2004. Bath, UK, September 12-17, 2004*. Ed. R. Heery and L. Lyon. Berlin: EDCL, 2004. 244- 255.

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

*DARE, Digital Academic Repositories*. Accessed 2004-11-20. <<http://www.surf.nl/en/themas/index2.php?oid=7>>

*e-Laborate*. Accessed 2004-11-20. <<http://www.e-laborate.nl/>>

*JITT*. Accessed 2004-11-20. <<http://www.sbl-site2.org/Extreme2002/>> and <[http://www.idealliance.org/papers/xml02/dx\\_xml02/index/titl e/e93017c13fc3874332dee40367.html](http://www.idealliance.org/papers/xml02/dx_xml02/index/titl e/e93017c13fc3874332dee40367.html)>

*LMNL*. Accessed 2004-11-20. <<http://lmnl.net/>>

McGann, J.P. "Dialogue and interpretation at the interface of man and machine, reflections on textuality and a proposal for an experiment in machine reading." *Computers and the Humanities* 36 (2002): 95-107.

*NHDA*. Accessed 2004-11-20. <<http://www.niwi.knaw.nl/en/geschiedenis/collecties/>>

*NIWI-KNAW*. Accessed 2004-11-20. <<http://www.niwi.knaw.nl>>



---

## The Abraham Lincoln Historical Digitization Project, the World Wide Web, and the Public Humanities

---

*Drew VandeCreek (drew@niu.edu)*  
*Northern Illinois University*

---

The *Abraham Lincoln Historical Digitization Project (Lincoln Project)* at Northern Illinois University presents primary source materials shedding light upon Abraham Lincoln's life and context in *antebellum* Illinois (1831-1861) on its *Lincoln/Net* World Wide Web site ( <<http://lincoln.lib.niu.edu>> ). Begun in 1998 as a model digital library project, it has gathered texts and images (over twenty-five million words of text, and over 2500 images) from significant historical collections in the state of Illinois in a single set of searchable databases, and thus dramatically expanded the ability of students, teachers, scholars, and the public to use them.

But the *Lincoln Project* has grown to become more than a large digital library. It also features original interpretive materials, written by leading scholars, which help the site's users to think about the historical context in which Lincoln lived, and in which he and his contemporaries produced the historical materials on display. In recent years the project has also produced a documentary film examining Lincoln's role in the Black Hawk War of 1832 and a nationwide reading program for public libraries (in collaboration with the *American Library Association*). These projects have added new resources to the web site, including over one hundred streaming video clips featuring segments of the project film and leading scholars discussing major themes in this period's history.

While scholarly researchers make wide use of the *Lincoln/Net* site, the *Lincoln Project* also represents an attempt to use the opportunities presented by digital technology and the World Wide Web to expand the scope and effectiveness of the public humanities.

These opportunities have emerged at several levels. First, a web site like *Lincoln/Net* provides its users with an opportunity to explore primary source materials within an integrated learning environment including interpretive materials. Many digital library users who are not professional scholars often wonder "What should I search for?" These secondary resources

help site users to learn about the fundamental questions that scholars ask about this period in American history. They also help site users to begin to fashion research questions with which they may explore the databases. This ability to pursue original explorations in primary source materials can contribute in new, rare ways to the public's appreciation for historical contingency and change.

The World Wide Web and digital technology also provide a new opportunity to support and enrich more traditional humanities programming. Traditional public programs in the humanities provide isolated opportunities to attend lectures, films, symposia, or other events. But often attendees are left with nothing more than the few notes they may have scribbled during the event. The World Wide Web provides an opportunity to furnish program attendees with an opportunity to follow up their lecture experience with further explorations in pertinent primary source materials and additional scholars' interpretations of them. A web site also provides more prosaic services: the opportunity to read the text of the original lecture at one's leisure, or the ability to view a film again via streaming digital video.

While many members of the *ACH* and *ALLC* have devoted their efforts to devising new scholarly applications for digital technology and the World Wide Web, I would like to argue that they represent an unprecedented opportunity for scholars to open a new dialogue with members of the public who may harbor an interest in humanities subjects, but find little opportunity to nourish it. I have built the *Abraham Lincoln Historical Digitization Project* on the hypothesis that a significant segment of the public harbors an interest not only in exploring primary source materials, but also in imbibing and considering scholars' research questions and conclusions. The *Lincoln/Net* World Wide Web site thus presents an opportunity for these scholars to reach out to a new audience, not by persuading the public to read their monographs or textbooks, but by boiling their research conclusions down into readily accessible interpretations of the primary source materials that are, thanks to new technology, readily at hand.

This public discourse can serve the humanities well. It can return them to the civic role that they often enjoyed before the era of ever-narrower scholarly specialization. Increased appreciation for, and comprehension of, scholarly work by even a limited segment of the public can also prove valuable in a political climate that has often witnessed taxpayer assaults upon university and college budgets.

I hope that this poster presentation and/or paper will lead conference attendees to consider using new technology to do more than refine research techniques, and reach out to a public audience.

# Databases and Prosopographies: *The Prosopography of Anglo-Saxon England (PASE) a Case Study*

---

*Hafed Walda* ([hafed.walda@kcl.ac.uk](mailto:hafed.walda@kcl.ac.uk))

King's College London

*Alex Burghart* ([alex.burghart@kcl.ac.uk](mailto:alex.burghart@kcl.ac.uk))

King's College London

---

## Summary

### Dr Hafed Walda

#### Definition

David Pelteret, of the *Prosopography of Anglo-Saxon England*, wrote that: "in essence prosopography can be interpreted as the study of identifiable persons and their connections with others for the purpose of enabling the modern student to discern patterns of relationships." (Pelteret 13).

The *Prosopography of Anglo-Saxon England Project (PASE)* is based in the department of History, the Centre for Computing in the Humanities at King's College London and in the Department of Anglo-Saxon, Norse, and Celtic at the University of Cambridge. The aim of the *PASE Project* is to record everything that is known about all Anglo-Saxon individuals mentioned in any source written between 597 and 1042. This will create a comprehensive register of the recorded inhabitants of the period. *PASE* will be accessible in the form of a freely available, searchable on-line database.

The past two decades have witnessed enormous growth in the number and importance of Prosopographies such as *PASE*. Following this period of rapid growth will the academic community find a common technological ground for all these Prosopographies? This paper explores the issues surrounding the search for this common ground.

#### Historical overview

The father of prosopographies is the *Corpus Inscriptionem Latinarum (CIL)* edited by Christian Matthias in 1858. After

the publication of the *CIL*, Mommsen worked on the original *Prosopographicum Inscriptionem Romanorum (PIR)* until 1877. In that he made use of his considerable experience with *CIL*, hence his first work mimicked the *Inscriptions* but was supplemented by literary sources and papyrology. After a long delay Professor A.H.M. Jones continued Mommsen's work with the help of his two pupils John Morris and John Martindale in the 1950s. The continuation of Mommsen's work became an international affair. The huge task was divided between the French (under the direction of H.-I. Marrou) and the British (under A.H.M. Jones).

A.H.M. Jones died before the first volume of the *Prosopography of the Later Roman Empire (PLRE)*, covering the years AD 260-395, was published in 1971. He did however manage to read through and edit the final draft. The British Academy ensured the survival of the project by providing financial aid from 1970. John Morris and John Martindale continued to work on the project, volume two (covering the years AD 395-527) being published just after the death of John Morris in 1980. John Martindale was left to edit the third volume, eventually published in 1992 in two volumes covering AD 527-64, before he retired in 2000.

The French *Prosopographie Chrétienne* was divided on a regional as well as chronological basis. Marrou and Mandouze produced the volumes for Africa (AD 303-533) in 1982. Another two volumes covering Italy (AD 313-604) were published in 1999, under the direction of Charles and Luce Pietri.

#### Data and development issues

With the development of scholarship through the increased availability of written sources, the publication of new editions, and through the publication of *Inscriptions*, coins and seals, the research materials have also increased. New methods using searchable databases were developed to deal with the sheer quantity of material available in various formats.

In the past two decades, computer based methods of recording and manipulating data have offered historians in a variety of fields new opportunities of data manipulation that go beyond what was formerly feasible for scholars using traditional research methods geared for paper publication. This was hailed as an ideal way of converting data into information by processing and presenting them for human interpretation.

The database approach to the development of Prosopographies has been found attractive to scholars in Anglo-Saxon studies (*PASE*), *Clergy of the Church of England (CCE)* and *Prosopography of the Byzantine World (PBW)*. All these Prosopographies are based in King's College London and developed in collaboration with the Centre for Computing in the Humanities (CCH). The role of CCH involves a whole range

of activities, including data analysis, system design, the application of computing tools, and technical advice and long-term support.

These Prosopographies have benefited hugely from the technical and academic knowledge that has been accumulated at King's College London.

In this paper I will draw a comparison between the various methods of collecting and displaying prosopographical data in different formats from the earlier book-based to electronic editions, and will analyze the advantages of each method. The comparison will involve using actual historical records.

I will be looking closely at the advantages, costs and risks of using the Relational Database model to drive the data and the use of web browsers as interfaces to display the information. The Prosopography of Anglo-Saxon England database will be used as an example of a database driven prosopography.

## The Prosopography of Anglo-Saxon England, 597-1042

### Alex Burghart

This section will demonstrate the research possibilities made available by *PASE*. It will give a live presentation of how data from sources has been entered, collated and reconciled.

A wide range of source types survive from Anglo-Saxon England, these include chronicles, letters, biographies, and legal texts. Sources such as Bede's *Ecclesiastical History* and the *Anglo-Saxon Chronicle* have generated a prodigious amount of secondary literature and probably an uncountable number of editions in a wide variety of vernacular languages. Yet these editions could not, on their own, answer such basic questions as, how many Anglo-Saxons held a certain office; or establish the links between people and overall groupings with systematic and accessible structure.

Perhaps the most substantial advance that *PASE* has made on standard prosopography is that it records not only data concerning individuals (their status, what they owned, to whom they were related, what they ate for breakfast &c.) but also information about how individuals were connected with each other. This has been embodied in the database by the creation of *EVENT* in which is recorded all meetings / relations between one or more people. This is a significant step in understanding Anglo-Saxon history because for the first time historians will be able to search the whole corpus of Anglo-Saxon sources for associations between people. This lies at the heart of what prosopography should be.

The major source of associations in Anglo-Saxon history is that of the charters. The term 'Anglo-Saxon charter' covers a multitude of documents ranging in kind from the royal diplomas issued in the names of Anglo-Saxon kings between the last quarter of the seventh century and the Norman Conquest, which are generally in Latin, to the wills of prominent churchmen, laymen, and women, which are generally in the vernacular. A large proportion of the surviving corpus of about 1500 charters is made up of records of grants of land or privileges by a king to a religious house, or to a lay beneficiary. The corpus also includes records of settlements of disputes over land or privileges, leases of episcopal property, and records of bequests of land and other property. Its importance for *PASE* lies in what they tell us about individuals. Most charters include invaluable information about ownership and status, but, as legal documents, they also frequently include lists of people who gave their agreement to the settlement described in the charter. These names give us an insight into the workings of royal and local courts and communities which we would have otherwise been denied. In collating these names and relating them to other source material, *PASE* grants the researcher the ability to reveal something of the lives which hide behind them.

Anglo-Saxon charters are the best represented corpus of medieval documents on the internet. The completion of *PASE* will confirm and bolster this position.

## Bibliography

Bradley, John, and Harold Short. "Texts into databases: The Evolving Field of New-style Prosopography." Paper delivered at the ACH/ALLC Conference, University of Georgia, Athens Georgia. Summer 2003.

Cameron, Averill, ed. *Fifty Years Of Prosopography: The Later Roman Empire, Byzantium and Beyond*. Proceedings of the British Academy, number 118. New York: Oxford University Press, for the British Academy, 2003.

*Clergy of the Church of England*. King's College London. Accessed 2005-03-21. <<http://maple.cc.kcl.ac.uk:8080/cce/rochester/index.html>> (Not yet launched.)

Jones, A.H.M., J. R. Martindale, and J. Morris. *The Prosopography of the Later Roman Empire*. Cambridge: Cambridge University Press, 1971.

Klebs, E., P. von Rohden, and H. Dessau. *Prosopographia Imperii Romani*. 3 vols. Berlin: Walter de Gruyter, 1956. A new second edition was published in 1998.

Martindale, John Robert, ed. *Prosopography of the Byzantine Empire I (641-867): The CD of the First Period*. King's College London, 2002.

Peltert, David. "[Title not provided]." *History and Computing* 12.1 (2002): 13.

*Prosopography Centre*. Modern History Research Unit, University of Oxford. Accessed 2005-03-21. <<http://users.ox.ac.uk/~prosop/>>

*The Prosopography of Anglo-Saxon England*. King's College London. Accessed 2005-03-21. <<http://www.kcl.ac.uk/humanities/cch/pase/>> ; database site <<http://maple.cc.kcl.ac.uk:8080/pase/index.jsp>> (not publicly launched).

---

## TM4DH (Topic Maps for Digital Humanities): Examples and an Open Source Toolkit

---

*John Walsh* ([jawalsh@indiana.edu](mailto:jawalsh@indiana.edu))

*Indiana University*

---

**M**y paper will discuss the use of Topic Maps to explore literary topics and navigate TEI-encoded literary texts. The presentation will include a brief introduction to Topic Maps and the XML Topic Map (XTM) syntax; an argument for the benefits of Topic Maps and similar metadata and ontology formats in electronic text and textual analysis applications; examples of fully-developed topic maps describing poetic genres and verse forms as well as the life and work of an individual author, Victorian poet Algernon Charles Swinburne. The presentation will conclude with a demonstration of an open source Topic Map toolkit, developed by the author and consisting of a Java-based web application and XSLT stylesheets for the presentation and navigation of XML-based Topic Maps.

Topic Maps have been the subject of recent attention in the digital humanities community. They are discussed briefly in John Bradley's "A Model for Text Analysis Tools" in a recent issue of *Literary and Linguistic Computing*. And at Digital Resources for the Humanities (DRH) 2003 and the 2003 TEI Members Meeting, Stuart Brown gave presentations on Topic Maps as a means for navigating the TEI Guidelines and comparing different sets of local TEI extensions. At DRH 2004, the author presented a paper on "Topic Maps and TEI-Encoded Literary Texts". The current paper will build on this previous work by exploring more fully developed and generally applicable topic maps in conjunction with the open source toolkit that can be used by others in the digital humanities community to deliver their own Topic Map applications.

Topic Maps are a powerful XML metadata format that may be used to create multi-dimensional indices and interfaces to humanities resources and TEI-encoded data. The basic building blocks of Topic Maps are topics. A topic may represent any subject or concept. A Topic Map about a poetry collection, for instance, may contain topics representing "poetry," "sonnet," "lyric," "poet," "Wordsworth, William," and "The Solitary Reaper." Topics may have multiple names and may be connected via associations. For instance, one poet may "influence" another, a poet may "author" a poem, or a poet may "mourn" a historical figure in an elegy. Topics may also be

typed and may have multiple types. So "Hamlet" may be an instance of the topic "play." Or "Rossetti, Dante Gabriel" may be an instance of both the "poet" and "painter" topics. Topic Maps allow one to build lists of concepts important to a collection or to a particular area of research and link those concepts to electronic resources, including TEI documents and individual elements within TEI documents. Topic Maps are powerful research and pedagogical tools that facilitate the organization, presentation, navigation, and visualization of conceptual and factual data.

To illustrate the discussion, I will use Topic Maps and TEI texts being developed as part of the *The Swinburne Project*, a digital collection of works by Victorian poet Algernon Charles Swinburne. Swinburne was an important cultural figure whose impact was felt beyond the domains of literature and poetry. He is an ideal central figure for the study of a wide range of nineteenth-century cultural and historical topics. Swinburne was an incredibly learned poet, and his range of form and allusion is extensive. His works include numerous and often obscure allusions to the bible, classical mythology, and Arthurian legend. He wrote a number of political poems addressing contemporary events. He wrote parodies of other contemporary poets, including Tennyson, Browning, and Rossetti. As Jerome McGann has noted, "[n]o English poet has composed more elegies than Swinburne." To address the breadth and range of form and allusion in Swinburne's work, the Swinburne Topic Maps include extensive lists of genre forms, people, biblical figures, mythological figures, Arthurian figures, as well as events and works by other artists and poets. These lists may then be used to build elaborate indices, navigation mechanisms, and data visualizations. The combination of TEI-encoded texts and XML Topic Maps allows the construction of a complex database of nineteenth-century British culture with Swinburne at its center. Supplementing the Swinburne-specific Topic Maps will be a more generally applicable Topic Map on poetic genres, meters, and verse forms. Swinburne was a versatile versifier who wrote in a great variety of forms. This "genre map," begun as a description of forms used by Swinburne, and since extended to include additional forms, may be used in conjunction with other digital collections.

The TM4DH (Topic Maps for Digital Humanities) open source topic map toolkit consists of a configurable Java-based Web application and bundled XSLT stylesheets for the presentation and navigation of topic maps conforming to the XTM 1.0 specification ( <http://www.topicmaps.org/xtm/1.0/> ), which describes an XML grammar for interchanging Web-based topic maps. TM4DH generates a homepage for the topic map. This homepage includes general metadata for the topic map (author, title, date, etc.) and a listing of major topic categories. From this homepage users can navigate to pages for individual topics, including detailed information about the

topic along with its occurrences (internal to the Topic Map or external on the Web) and associated topics.

## Bibliography

[No source references provided. Eds.]

# Exploring the Use of Term Proximity in Collocate-Ranking for Query Expansion

---

**Ying Wang** ([yingwang@engmail.uwaterloo.ca](mailto:yingwang@engmail.uwaterloo.ca))

University of Waterloo

**Olga Vechtomova**

([ovechtom@engmail.uwaterloo.ca](mailto:ovechtom@engmail.uwaterloo.ca))

University of Waterloo

---

The exponential increase in the amount of humanities information available in digital libraries and archives calls for better search techniques that can help information users to retrieve full-text documents matching their information need with high accuracy. *Information Retrieval (IR)* research can lead to improved search techniques that facilitate access to large collections of humanities literature.

IR researchers focus on various topics to improve the retrieval performance, such as the representation of documents, the formulation of queries, and document matching and ranking techniques. Many established retrieval models do not take into account relations between words in text. While they work well with short and semantically homogeneous documents, arguably they are less appropriate for long multi-topic and more semantically complex texts. Many documents in the humanities archives fall under the latter category. In this paper we report a study that was conducted to develop a new query expansion technique which uses term proximity information and statistical term association measures in selecting query expansion terms.

Query expansion is a technique commonly used in IR (e.g., Rocchio; Beaulieu) to improve the retrieval performance by reformulating the original query - either adding new terms or reweighing the original terms. Query expansion terms can be automatically extracted from the documents or taken from knowledge resources, such as thesauri. The main advantage of the former techniques is that they are collection-independent and cheaper to construct. Typically either top-ranked documents in the initially retrieved document set (blind or pseudo-relevance feedback) or documents judged relevant by the user in the retrieved set (relevance feedback) are used to extract query expansion terms. For short and incomplete queries, a substantial improvement can be achieved by using expanded queries (Sparck-Jones et al.). Terms added to the original query usually have the following common characteristics: a) they are semantically related to the original query terms; b) they are

good at discriminating between relevant and non-relevant documents. Computational linguists use various statistical association measures to extract significant word associations (or co-occurrences), as these measures can judge the degree of closeness between words. Association measures have been also used in query expansion to select words that are closely related to query terms (Ishikawa et al.; Vechtomova et al.).

The query expansion method proposed in this paper is used to select words which co-occur with the original query terms in a certain proximity (such as the same sentence, paragraph or a fixed-size window) in the documents judged as relevant by the user. We refer to such terms as collocates of the original query terms. We experimented with a number of parameters for selecting such terms, such as their distance from the original query term(s) in text and their degree of association with the query terms.

Previous related studies investigated the effect of using the distance information between query terms for document ranking, whereas we investigated the effect of using term proximity in collocate-ranking for query expansion. Under the similar assumption "if the distance between two words is closer, the pair is considered as more associated with each other", we proposed to use a distance factor in collocate-ranking formula to measure the association between collocates and query terms. The main contribution of our experimentation is that we combined the distance-weighting factor with the traditional word association measure of *Mutual Information (MI)*.

The following three hypotheses were explored in this study:

*Hypothesis 1:* The use of term proximity in collocate-ranking for query expansion results in a significant performance improvement over no query expansion.

*Hypothesis 2:* The use of term proximity in collocate-ranking formula for query expansion can lead to significant performance improvements over the current best-performing term selection values. We used *Offer Weight (OW)* of the Robertson/ Sparck Jones IR model as the baseline (Sparck Jones et al.).

*Hypothesis 3:* The collocate-ranking formula using distance information results in a significant performance improvement over the formula without the distance factor.

The experiments were conducted using the *Okapi IR* system (Sparck Jones et al.), and *TREC (Text REtrieval Conference)* evaluation framework (Voorhees).

The collocate-ranking method is comprised of several formulae – *Cohesion score*, *Similarity score*, *MI score* and *Distance factor* formulae. *Cohesion score* and *Similarity score* were formulated in a similar way to those proposed by Gao et al.. The cohesion score is the final score to select query expansion terms; the similarity score of a pair (x, y) is the multiplication

of the *MI score* and the *distance factor*. *MI score* was formulated similarly as the local *MI score* proposed by Vechtomova et al.. As the goal of this study was to explore the use of distance in collocate-ranking, different distance factors that might improve the retrieval performance were investigated. The best distance factor proved to be Formula 4.

The cohesion between a collocate *y* and query topic *T* is defined in Formula 1.

$$Cohesion(y,T) = \log(\sum_{set} SIM(x,y)) \tag{1}$$

The similarity score of a pair (*x, y*) is the multiplication of the *MI score* and the distance factor, shown in Formula 2.

$$SIM(x,y) = MI(x,y) * df(x,y) \tag{2}$$

Where *MI(x, y)* – Mutual Information score of pair (*x, y*);

*df(x, y)* - the distance factor of pair (*x, y*).

*MI(x, y)* is calculated using frequencies from the set of relevant documents, shown in Formula 3.

$$MI(x,y) = \log_2 \frac{\frac{f_r(x,y)}{R * V_s(D)}}{\frac{f_r(x)}{R} * \frac{f_r(y)}{N}} \tag{3}$$

Where *f<sub>r</sub>(x, y)* - the joint frequency of pair (*x, y*) in the set of relevant documents;

*f<sub>c</sub>(y)* – frequency of *y* in the corpus;

*f<sub>r</sub>(x)* – frequency of *x* in the relevant documents;

*V<sub>s</sub>(D)* – average document length in the relevant document set;

*N* – corpus size;

*R* – size of the relevant document set (in tokens).

The best distance factor proved to be Formula 4.

$$df(x,y) = fr(x,y) / D(x,y) \tag{4}$$

Where *fr(x, y)* - the joint frequency in the relevant document set;

*D(x, y)* - the average distance of the pair (*x, y*) within documents in the relevant set.

*Formulae*

We performed statistical analysis of the search results produced using (a) terms selected by our experimental collocate-ranking formula (b) original query terms only and (c) query expansion terms selected by using OW. Figure 1 shows the precision values at 11-Recall levels using the above three methods. The analysis results indicate that the experimental search run using our derived collocate-ranking formula significantly improved the retrieval performance compared with the no-expansion run, but did not outperform the OW run. The method using term proximity in collocate-ranking was proved to be effective. Hypothesis 1 was supported by the analysis, while Hypotheses 2 and 3 were not supported. The top 10 query expansion terms selected by MI, the best distance formula and OW for the query topic #432 are presented in Table 1.

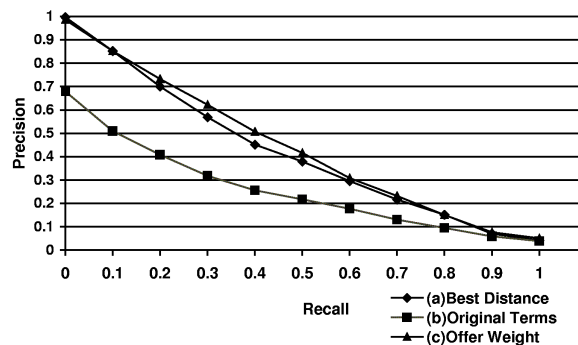


Figure 1: Precision Values at 11-Recall Levels

<num> Number: 432  
 <title> profiling, motorists, police  
 <desc> Description:  
 Do police departments use "profiling" to stop motorists?  
 <narr> Narrative:  
 A relevant document will report or discuss police department criteria for identifying motorists considered likely to be carrying contraband. Documents discussing the detention of individuals by foreign security forces are not relevant.

Best Distance Formula	MI	OW
Jackson	non-law	sherriff
officer	Hannon	checkpoint
Hawthorne	Mirabella	drunk
checkpoint	avocation	neighborhood
black	woodyard	non-law
Dickey	Dickey	enforcement
I	remade	search
complaint	D-hayward	Hannon
search	feeler	Mirabella
department	Hardeman	15300

Table 1: The top 10 query expansion terms for sample topic #432

Some of the findings and recommendations from this study are: the distance factor has to be compatible with the collocate-extraction process and the MI score itself is an effective collocate-ranking formula compared with no query expansion. Further studies on the use of term proximity for query expansion need to be carried on through integrating other promising techniques, such as part-of-speech tagging, into the query expansion process.

This study contributes to advancing high-accuracy retrieval of documents from large resources and archives in humanities through the investigation of the role of distance and association between words in text for selecting useful terms that can be added to the search formulations and help searchers find more relevant documents. Retrieval techniques which capture relations between words in text are particularly promising for the high-precision retrieval of long multi-topic texts. Large proportion of the humanities literature consists of such texts. Query expansion techniques that assume term independence in text are less appropriate for such collections. The techniques presented in this paper can also be used in interactive query expansion. Searchers often have difficulty in formulating their information need. Previous studies showed that searchers prefer to formulate short queries and then browse through the document space and reformulate queries manually. Finding related terms that co-occur with the query terms and suggesting them to the searcher can facilitate this process.

## Bibliography

- Beaulieu, M. "Experiments with interfaces to support query expansion." *Journal of Documentation*, 53.1 (1997): 8-19.
- Gao, J., J. Nie, H. He, W. Chen, and M Zhou. "Resolving query translation ambiguity using a decaying co-occurrence model and syntactic dependence relations." *Proceedings of the 25th annual international ACM SIGIR conference on Research and Development in Information Retrieval*. Tampere, Finland, 2002. 183-190.
- Ishikawa, K., K. Satoh, and A. Okumura. "Query term expansion based on paragraphs of the relevant document." *Proceedings of Sixth Text Retrieval Conference (TREC-6)*. Gaithersburg, MD, USA, 1997. 577-584.
- Rocchio, J.J. "Relevance feedback in information retrieval." *The SMART Retrieval System – experiments in automatic document processing*. Ed. G. Salton. Englewood Cliffs, New Jersey: Prentice-Hall, 1971. 312-323.
- Sparck Jones, K., S. Walker, and S.E. Robertson. "A probabilistic model of information retrieval: development and comparative experiments." *Information Processing and Management* 36.6 (2000): 779-808 (Part 1); 809-840 (Part 2).
- Vechtomova, O., S. Robertson, and S. Jones. "Query expansion with long-span collocates." *Information Retrieval* 6.2 (2003): 251-273.
- "Overview of TREC 2003." *Proceedings of the twelfth Text Retrieval Conference (TREC 2003)*. Gaithersburg, MD, USA, 2004. 1-13.

---

## User Centred Interactive Search: a Study of Humanities Researchers in a Digital Library Environment

---

*Claire Warwick* (c.warwick@ucl.ac.uk)

*School of Library, Archive and Information  
Studies, UCL*

*Ann Blandford* (a.blandford@ucl.ac.uk)

*UCL Interaction Centre*

*George Buchanan* (g.buchanan@ucl.ac.uk)

*UCL Interaction Centre*

---

This poster proposal describes research on humanities users of a digital library (DL). It seeks to understand their needs and behaviour both in digital and more traditional information environments, in order to develop and refine a digital library system, the better to support use in the humanities. This study of humanities users forms part of the larger *User Centred Interactive Search (UCIS)* project.

## Background

Large, structured information repositories such as digital libraries (DLs) are becoming commonplace. To realise their potential, they need to be usable and useful - by a range of users, in different situations, supporting a variety of information tasks. The current generation of DLs still poses substantial user difficulties: searches are often time-consuming and frequently unsuccessful (Blandford et al.), and the reasons for success or failure remain mysterious to most users. Within the broader information task, the information requirements are often poorly defined, as users are often trying to refine the information problem by using available information to understand what is possible, so that information acquisition is an evolving, highly interactive activity.

It is widely recognised that creating effective search criteria to achieve a particular information goal is a demanding and difficult task, particularly for less experienced users, and particularly when the goal is as yet under-defined. Shneiderman et al. observe the challenges of selecting a variety of search attributes, such as the words to be used in a query and the syntactic peculiarities of the system at hand. In addition, the



mapping of an information need to the use of metadata fields or full text search can prove difficult (Blandford et al.). Unlike the web, where the document text is the only possible target for a search, DLs provide a rich environment for information seeking: the user has a much wider potential range of selections (classification, author, publication date, etc.) to make. Effective searching relies on the careful selection and use not only of words or syntactic commands, but also of fields and information sources.

## Use in Context

Surprisingly little work on information seeking has set it within the context of the broader information work of which it is a component (Attfield et al.). While this divorce from the context may be valid when considering work in physical libraries, where the information seeking task is often a bounded activity delineated by arrival at and departure from the library building, it is less so for DLs that can be accessed from the user's normal place of work, removing the marked transitions between information seeking and other activities. One hypothesis this study will test is that users expect information seeking to flow more naturally into their broader information task when searching from their normal place of work.

## Humanities Users

Humanities researchers are the focus for studying use in context for several reasons: they typically have little technical or mathematical knowledge (e.g. for immediately understanding the designs of complex interactive systems or intuitively being able to construct the Boolean queries that are often key to successful query formulation); they often do not have a clear idea of what they are looking for, but will usually recognise it when they find it; and they have not been extensively studied, although they have substantial and sophisticated information requirements. In summary, humanities researchers are a particularly challenging population to design for, and many solutions that work for this user population are likely to also suit less demanding users. Studies of humanities researchers have tended to concentrate on needs or the types of resources used (*Library Trends*; *Open University*). Many of these are now relatively dated, and although their conclusions were important at the time, both the types of resources available and the technology used to find them have changed. Studies by Stone and Watson-Boone established that humanities users need a much wider range of resources than those in other disciplines; for example, they may need to refer to material which is much older than that used by researchers in the sciences and social sciences. They may still need to use

historical material in the form of manuscripts or early printed books even if digital surrogates are available (Warwick and Carty, Duff and Johnson). Relatively low levels of use of digital resources have been blamed on these particular needs, and on a lack of knowledge about their capabilities (Corns), but this has yet to be verified by an empirical survey.

Warwick's previous work suggests that humanities users may find it difficult to adapt search behaviour from a traditional to a digital library setting, and thus become discouraged by failed attempts to locate appropriate resources (Warwick, 1999a; 1999b) She has found that the patterns of use of digital resources in English literature have changed little since Corns' work in 1991, and argued that this may be because of lack of fit between the searching tasks users wish to carry out and the present capabilities of DLs. This was based on only a small sample of users, and on theoretical data. It is therefore important to test these hypotheses by studying a meaningful sample of humanities users in both a traditional library setting and a digital environment. This is one of the tasks that the present research is engaged in.

## Aims of the project

Overall, there are four strands of work in the UCIS project:

1. studying use of information in context, focusing on humanities researchers;
2. studying the development of expertise in searching (focusing on information management students);
3. identifying requirements on the design of digital libraries; and
4. developing and testing system modules for a digital library.

The proposed poster will describe the first strand of work, briefly outlining how it fits within the rest of the project. We believe that this work is important since very little work has studied use in context - particularly in the humanities - and translated findings into testable design requirements.

## Methods

Qualitative data (from interviews, observations, diary studies, transaction logs, etc.) will be gathered from academics and other researchers in the humanities regarding their activities with DLs and similar information resources. Two sub-issues will direct this work: how humanities researchers work with digital resources and how they integrate use of electronic and paper resources - both within the broader task context.

The first approach to data collection will be by user diaries, in which humanities users record their use of information resources (both traditional and digital) to support their research. This will provide base-line data to inform the use of techniques for subsequent study (depending on the patterns of resource use). The main approach to data collection will be contextual inquiry interviews, observing users as they work with digital libraries of their own choice and interviewing them on their perceptions of the usability of such electronic information sources. The focus of this data collection will cover what users currently do, their perceptions of the strengths and limitations of current technologies (including traditional resources), and requirements for future systems.

Data will be analysed in two different but complementary ways: first, using a Grounded Theory-style approach (Strauss and Corbin) to develop theory on the use of digital resources in context by humanities users; second, using design-oriented techniques to draw out requirements for design.

To enable us to focus on new technical challenges rather than needing to replicate work already done by others, technical developments will be based on the *NZDL Greenstone* software, for which a test collection material specific to the humanities has been developed. By grounding the work in empirical studies, we will be able to identify and present further requirements on the design of such systems. By basing system development on an established DL platform, we will be able to test candidate design solutions, deliver working components as part of the *Greenstone* system and provide examples for developers of other DL systems that illustrate tested approaches to improving user experience.

## Findings

The *UCIS* project began in August 2004 and the humanities phase will begin in early 2005. We therefore propose to use this poster to report on early findings of the research. It is for this reason that we have proposed a poster session, since this will be a report or work in progress.

## Acknowledgement

This work is funded by *EPSRC Grant GR/S84798*.

## Bibliography

Attfield, S.J., A. E. Blandford, and J. Dowell. "Information seeking in the context of writing: a design psychology

interpretation of the 'problematic situation.'" *Journal of Documentation* 59.4 (2003): 430 - 453.

Blandford, A.E., H. Stelmaszewska, and N. Bryan-Kinns. "Use of multiple digital libraries: a case study." *Proc. JCDL 2001*. Roanoke, VA, 2001. 179-188.

Corns, T.N. "Computers in the humanities: methods and applications in the study of English Literature." *Literary and Linguistic Computing* 6.2 (1991): 127-131.

Duff, W.M., and C.A. Johnson. "Accidentally found on purpose: Information-seeking behavior of historians in archives." *Library Quarterly* 72 (2002): 472-496.

1992.

Open University Library. *Safari: Skills in Accessing, Finding and Reviewing Information*. 2001. Accessed 2005-04-15. <<http://www.open.ac.uk/safari/>>

Shneiderman, B., D. Byrd, and B. Croft. "Sorting out searching." *Communications of the ACM* 41.4 (1998): 95-98.

Stone, S. "Humanities Scholars-Information needs and uses." *Journal of Documentation* 38.4 (1982): 292-313.

Strauss, A., and J. Corbin. *Basics of qualitative research: grounded theory procedures and techniques*. Newbury Park, CA: Sage Publications, 1990.

Warwick C., and C. Carty. "Only Connect, a Study of the Problems caused by platform specificity and researcher isolation in humanities computing." *Electronic Publishing 01, 2001 in the digital Publishing Odyssey. Proceedings of the 5th International ICC/IFIP Conference on Electronic Publishing*. Ed. Arved Hubler, Peter Linde and John W.T. Smith. 2001. 36-47.

Warwick, C. "English Literature, electronic text and computer analysis: an unlikely combination?" *Proceedings of the Association for Computers and the Humanities- Association for Literary and Linguistic Computing, Conference, June 9-13*. University of Virginia, 1999a. 71-74.

Warwick, C. "The lowest canonical denominator: Electronic literary texts, and their publication, collection and preservation." *New Fields for Research in the 21st Century, Proceedings of the Anglo Nordic Conference 1999*. Ed. M. Klasson, B. Loughridge and S. Loof. Boras, Sweden: Swedish School of Library and Information Studies, 1999b. 133-141.

Watson-Boone, R. "The Information Needs and Habits of Humanities Scholars." *Reference Quarterly* 34 (1994): 203-216.

---

## Annotating Electronic Texts of Shakespeare

---

*Philip Weller (pweller@ewu.edu)*  
*Eastern Washington U*

---

There are many texts of Shakespeare's plays online, but most are entirely without notes of any kind, which makes them only minimally useful to the general reader or beginning student of Shakespeare. There is a need for reliable and fully annotated electronic texts.

Currently there are only a handful of annotated Shakespeare texts online, and they all indicate the presence of a note by making a word or phrase in the text into a hyperlink. Some of these annotated texts are misleading because the links often point to inappropriate entries in a glossary. This problem could be solved by more careful work on the part of the editor, but a larger problem is that making words into hyperlinks is not the best way of annotating Shakespeare.

Notes should be unobtrusive; they should not tease a reader into looking at a note that he/she doesn't need. Notes should also be efficient; they should provide needed information at a glance, without the necessity of losing one's place in the text.

Print texts of Shakespeare exhibit various devices to make the notes unobtrusive, but at the sacrifice of efficiency. Bevington's *Complete Works of Shakespeare* avoids footnote numbers and markers; the presence of a footnote is indicated by the presence of a line number. *The Riverside Shakespeare* goes even farther; line numbers are given in intervals of five, but the reader has to look to the bottom of the page to see if there are any notes. The Folger Shakespeare editions put all notes on a facing page, with their line numbers, but without any indication in the text of the presence of a note. All of these make for slow going; it's up to the reader find the line number that's associated with the note, and it often takes several shifts of focus from text to notes and back again to make the correct connection. *The Norton Shakespeare*, in a compromise effort, puts notes of one to three words in the right-hand margin and flags the annotated word with superscript circle; longer notes are put at the bottom of the page and flagged by a footnote number.

Making words into hyperlinks is both obtrusive and inefficient. Hyperlinks are extremely obtrusive; the different-colored highlighting insists that the reader must be missing something if he/she doesn't click. And making repeated clicks to get information is very inefficient.

The scheme which I will demonstrate is not perfect, but is less obtrusive and more efficient than any other, either print or electronic. Most notes are presented in a column five pixels to the right of the column of text. Each of these notes is preceded by bolded word or phrase which indicates the subject of the note. Longer notes are indicated by a right-arrow which is a hyperlink; the hyperlink opens a smaller window, sized to the length of the note, and positioned to the right of the column of text. This formatting allows the reader to find most of the information he/she needs at a glance, and the pop-up windows allow for an unlimited amount of information, including images, without forcing the reader to lose his/her place in the text.

The technique that produces this formatting depends the use of the HTML table and JavaScript.

Each line of text is contained within the center cell of a three-cell table row. Within the left-hand cell of each fifth table row is an act, scene, and line number, so the reader always knows where he/she is, without having to scroll or look to another part of the page. The last cell of each table row provides the space for the notes. Tagging the data cells of each table row as `nowrap` keeps everything lined up, while at the same time allowing the reader to zoom in for more readability.

For longer notes, the use of JavaScript to open the windows means that each window can be a different size, according to need, and can be positioned in such a way that it does not obscure the column of text.

I am already applying these techniques to an online edition of *Julius Caesar*. To see my work so far, go to this address: [http://www.clicknotes.com/Julius\\_Caesar](http://www.clicknotes.com/Julius_Caesar).

## A New Methodology for Parts of the Dutch Price History, Based on Analysis of the Paalgeld Portbooks, 1771-1778

---

*George Welling* ([welling@let.rug.nl](mailto:welling@let.rug.nl))  
*University of Groningen*

---

In the early thirties of the last century an international group of economic historians gathered to set standards for the writing of national price-histories of the participating countries. In the Netherlands Nicolaas Posthumus started the Herculean task of gathering the data for the two volumes he would publish: *Nederlandsche Prijsgeschiedenis (Dutch Price-History)*, which was published in 1943 and the second volume that was published in 1964 after his death in 1960. The last volume was completed by F.Ketner. Both volumes would become cornerstones of the Dutch economic and social history.

However, the Second World War made contacts between the historians involved very difficult and most of them decided to continue their work based on the agreements made earlier. Posthumus was very strict in sticking to the rules set by the committee, although the some of the decisions made by the group would have far reaching effects on the reliability of his work. Still, since then these price histories have been seen as the solid foundations for various discussions, a.o. for the standard of living debate.

A crucial aspect of the price-histories are the price indexes, which allow us to see developments over longer periods. There are two critical aspects in creating indexes: the choice of goods and their weights in the indexes, and the reference period for the indexes. It now seems that on both accounts the choice were unlucky.

There is good reason to question the solidity of the work done by Posthumus. First of all, the manual compilation of all the data lead to a selection of the data, since not all of the data from the Price Currents of the Amsterdam exchange could be processed. I will show that this selection has been carried out on a misinterpretation of the sources. For quite a number of goods there are more than one price notations in the Price Currents: the meaning of the various prices is not yet fully known. Posthumus just averaged these prices and used the average. It is quite unlikely that in cases where the lowest price notation and the highest price notation for the same good can

vary more than 100% that the meaning of these notations is identical.

Secondly, the methodology used by Posthumus to fill the gaps in his data is questionable. His sources were lacking and his treatment of missing data was quite creative if we put it mildly. His interpolation of missing data is based on averaging the two closest available data, not taking into account seasonal fluctuations. Sometimes the real data were years apart and interpolations were made for all months in between, creating a very static picture for some prices. Next to that, Posthumus did not make clear which prices were real and which were interpolations. Next to that for some years, for which the data in Dutch were missing, he used Italian versions of the Price Currents of the Amsterdam Exchange, forgetting however that they were dated in the Italian style of Venice, which put them in the wrong year.

Thirdly, most of the goods chosen for the indexes were imported to the Netherlands. But, because very little large scale research had been done about the composition of the total of goods going around, the choice was mainly based on pre-conceptions. Posthumus published two forms of indexes: unweighted and weighted indexes. The choice of goods for the indexes does not seem to be a real reflection of the importance of the goods that were traded on the exchange. Likewise the weights attached to the goods in computing the indexes seems to have been quite arbitrary.

Based on my recent research on the Portbooks of the levy of the Paalgeld in Amsterdam, a tax register on incoming shipping from overseas to the port of Amsterdam, which offer a complete oversight of all the goods imported to Amsterdam at the end of the 18th century, I will argue that the composition of the group of goods for the indexes does not reflect the realities of the trade. Some of the goods that Posthumus used for his indexes were not imported at all in the 18 year period, for which the data of the Paalgeld Portbooks were digitized.

I will offer a statistical analysis of the composition of the imports to Amsterdam, which will show that a different set of goods as a basis for indexes will allow a much more reliable view of the economical fluctuations. There have been plans to make a machine readable version of the *Nederlandse Prijsgeschiedenis (Dutch Price-History)*, but I suggest that a complete rethinking of the whole project is necessary.

### Bibliography

Posthumus, Nicolaas Wilhelmus. *Nederlandsche Prijsgeschiedenis*. Leiden: E.J. Brill, 1943-1964.

## Approaches to Searching for Language and Diversity in a 'Whitebread City' Digital Corpus: The Charlotte Conversation and Narrative Collection

*Stephen Westman* (srwestma@email.uncc.edu)

University of North Carolina at Charlotte

*Boyd Davis* (bdavis@email.uncc.edu)

University of North Carolina at Charlotte

**M**acaulay comments that

Dialects, like languages, have both a unifying and a separatist function. We speak the way we do to be like those we wish to associate with and to distinguish ourselves from others. When that association is based on where we live. . . a distinctive form of speech is likely to survive. However, we need to look at the whole configuration of linguistic features and not a few features which may or may not be the critical ones for the speakers.

(239)

Why? In addition to the "grammatical, phonetic, and lexical" features traditionally posited as characterizing a dialect, Macaulay adds

prosodic features and possibly also voice quality and discourse characteristics. There is no reason to believe that dialects have fewer features than other forms of language, and we do not know in advance which features will be important to distinguish the dialect.

(229)

In a discussion of features of southern style that warrant further investigation, Barbara Johnstone cites several which can be searched at word- and text-level: these lexicogrammatical features contribute to the reader/hearer's assessment of style, and include rhetorical genres triggered by particular discourse markers; style shifts into regional colloquialism, stylization and self-parody signaled by shifts into nonstandard verbs, for example, or a judicious sprinkling of double modals to suggest temporary intimacy. She asks for an investigation of regional styles of interacting that "makes strategic use of nostalgia for neighborhood, local community, or region." (206) Well, they said southerly, we have a gracious plenty of data that

accommodates such investigation; the issue is, of course, how to access, identify, and retrieve it.

The corpus that we are using to investigate these questions is *Project MORE's* expanded *Charlotte Conversation and Narrative Collection (CNCC)*, which is part of the 11.5 million-words in the First Release of the *American National Corpus (ANC)*. Considered a satellite corpus to the core of the *ANC*, which parallels the organization of the *British National Corpus* (Reppen et al.), the *CNCC* goal is far more modest, but still one of difficulty: to create a corpus of conversation and conversational narration in a New South city at the beginning of the 21st century. And that, of course, brings us smack up against issues of region (Macaulay) and representativeness (Douglas), of dialect diversity (Wolfram & Dannenberg), and distinctions between rural and metropolitan features (Tillery, Bailey & Wikle).

The *CNCC* is hybrid in some ways; similar to the *ONZE* corpus in its evolution through multiple formats and purposes (Gordon et al.). In addition to being a part of the *ANC*, it is also included in the *New South Voices (NSV)* digital resource housed at the University of North Carolina at Charlotte Library. *NSV* includes interviews that cover a wide range of historical subjects, from African American churches and Billy Graham crusades to women's basketball and World War II. Other interviews, narratives and conversations document the experiences and language of recent immigrants to the area. As such, it seeks to address a wide variety of audiences from local historians and historic preservationists to public school students. By using *NSV*, we are able to expand the number and range of interviews available for linguistic as well as for historical analysis.

If the corpus is to be inclusive of the range of spoken styles that conglomerate in the elastic borders of a New South city, it must begin by identifying what they are. In today's Charlotte, today's North Carolina, this is no longer simple. As Tillery, Bailey & Wykle note, metropolitanization, foreign and domestic migration, and expanding ethnic diversity have "eliminated many of the vestiges of traditional regional culture and . . . are radically reshaping the United States" (228). Their painstaking study of what they see as the impact of demographic change on American speech is keyed to 22 socio-demographic and linguistic variables: 14 phonological features, 3 lexical, and 5 that are lexicogrammatical. They see a balkanization (241; cf 244) with increased divergence of rural and urban ways of speaking; they ask will "old towns with new populations" — such as Charlotte — create new communities and new ways of speaking?

Investigation of these phenomena within a database environment requires a variety of tools and approaches if we are to extract the information contained in these transcripts. The reason for this is due to the nature of the types of

information we need to obtain from these interviews and then to the question of how we can best obtain that information.

On the one hand, we need to be able to perform textual analysis on the interviews and to examine subjects' speech patterns and linguistic characteristics. This in turn requires that we be able to extract information that is embedded within discursive text — looking at how they use language *'in situ'*. On the other hand, there are discrete pieces of information — metadata if you will — about the participants (place of origin, current residence, gender, ethnicity, etc.) to which we need access if we are to do anything meaningful with what we discover from our textual analysis. This dichotomy pertains to any area doing textual analysis. As noted by Ronald Bourrett, the roots of this dichotomy lie in the two types of information with which we are dealing: document-centric and data-centric.

Due to the different nature of the two types of information — linguistic analysis and descriptive metadata — we have found that a single approach does not allow us to fully explore the types of correlations we were seeking. While our XML database allows us to find useful things in searching tagged information within the interviews, it does not provide sufficient flexibility in searching data-centric information. On the other hand, with relational database technology we have exactly the opposite situation.

In addition, during the course of our investigation, we discovered that there were certain types of textual information — such as word- and phrase-frequency; retrieving and isolating particular words and phrases within their context in a document; and looking for particular words and/or phrases within certain proximity of each other — that were amenable neither to an XML database, nor a classic relational database, approach. To address this need, we decided that a third option — inverted indexes — was needed to allow us to look for such patterns. As noted in Zaïane, this technique greatly enhances the ability to search textual-based information.

Therefore, in designing our database system, we decided to adopt a mixed approach, one that allowed us to utilize the strengths of each system without running afoul of its limitations. In doing so, we use both XML and inverted indexing to do textual analysis and then a relational database to correlate that information with relevant demographic criteria. The result is a hybrid that allows us to do more than any single approach could provide.

This paper will first present how we are using readily available tools to implement a searching system that supports demographic correlation with textual features (including some features of proximity search and frequency of occurrence). These tools, all of which are part of the Open Source tools, allow us to build and configure with ease a system that not long

ago would require extensive and non-trivial programming. They include:

- *eXist* (XML) and *MySQL* (relational) database managers
- php, perl and Java programming languages
- *Apache* Web server

As a way of concluding, we will then earnestly solicit assistance on how we can best make this collection of roughly 1,000 transcribed oral interviews, conversations and narratives more useful to any researcher, particularly in the area of text-based, online searching.

## Bibliography

- Douglas, Fiona. "The Scottish Corpus of Texts and Speech: problems of corpus design." *Literary and Linguistic Computing* 18 (2003): 23-37.
- Gordon, Elizabeth, Margaret Maclagan, and Jennifer Hay. "The ONZE corpus. Manuscript." *Models and Methods in the Handling of Unconventional Digital Corpora. Volume 2: Diachronic Corpora*. Ed. J.C. Beal, K.P. Corrigan and H. Moisl. Houndsmills: Palgrave, Forthcoming.
- Hudson-Ettle, Diana. "Nominal that clauses in three regional varieties of English: a study of the relevance of text type medium, and syntactic function." *Journal of English Linguistics* 30 (2002): 258-273.
- Johnstone, Barbara. "Features and uses of southern style." *English in the Southern United States*. Ed. S. Nagle and S. Sanders. Cambridge: Cambridge University Press, 2003. 189-207.
- Kjellmer, Goeran. "A modal shock absorber, empathizer/emphasizer and qualifier." *International Journal of Corpus Linguistics* 8 (2003): 145-168.
- Macaulay, Ronald. "I'm off to Philadelphia in the morning." *American Speech* 77 (2002): 227-241.
- Reppen, Randi, and Nancy Ide. "The American National Corpus: Overall goals and the first release." *Journal of English Linguistics* 32 (2004): 105-113.
- Tillery, Jan, Guy Bailey, and Tom Wikle. "Demographic change and American dialectology in the twenty-first century." *American Speech* 79 (2004): 227-249.
- Wolfram, Walt, and Clare Dannenberg. "Dialect identity in a tri-ethnic context: The case of Lumbee American Indian English." *English World-Wide* 20 (1999): 179-216.
- Zaïane, Osmar. *Inverted Index for Information Retrieval (Slides keyed to Chapter 22 of unlisted textbook: CMPUT 391:*

*Database Management Systems*). University of Alberta, 2001. Accessed 2005-04-11. <<http://www.cs.ualberta.ca/~zaiane/courses/cmput391-02/slides/Lect7/>>

# Action and Interaction in Music and New Media Art: Exploration of Musicians' Performative and Interactive Decisions as Evidenced by Annotated Musical Scores

---

*Megan Winget* ([winget@email.unc.edu](mailto:winget@email.unc.edu))  
*UNC-Chapel Hill*

---

## Introduction

Preservation and representation of digital art presents a significant challenge for curators, archivists and artists. The most notable problem is the fundamentally interactive nature of this evolving art form. Because there are no clearly defined frameworks for the authentic representation of the variations resulting from interactions, curators and archivists are currently deciding on a case by case basis which interactions should be preserved, who should be able to interact with a given piece, and how to represent those interactions in a meaningful and objective way. Not only is this method problematic from a representational standpoint, it's very time-consuming. My thesis will attempt to resolve some of these complexities by making connections between this new art form and an established one, namely music; assessing the representational characteristics of music's notation system and exploring the methods by which musicians and conductors handle musical interaction.

## Background

One of the most significant challenges for the new media art community is development of a representation framework for preservation purposes. There are currently three preservation models under consideration. The first two have technical origins, and should be familiar to the general digital preservation community: migration, the premeditated upgrade of file formats; and emulation, which focuses on development of ur-operating systems able to run obsolete media. The third option, much more radical, and developed by and for the new media art community, is re-interpretation (Depocas, Ippolito & Jones); a method intimately related to the presentation,

exhibition, and performance of an interactive new media art object.

While re-interpretation is essential to success in the performing arts, in the fine arts and electronic preservation communities, the idea of re-interpretation as a valid preservation strategy challenges beliefs that are at the heart of these fields. Specifically, for the purpose of archival preservation, the characterization of a 'reliable' or 'authentic' object will need to undergo a significant transformation if these new media art objects; highly variable, interactive, and data-rich cultural artifacts, stand a chance of meaningful survival. Although radical, curators, archivists and conservators could regard the idea of re-interpretation as a way to honestly address the realities of presenting a variable work over time. Both migration and emulation effect subtle re-interpretations of a work, whether we perceive those changes or not. For example, the processor speed of an emulated computer will display an interactive work from the mid-1970s differently than the original operating system would. Whether that difference fundamentally changes the work's meaning can only be determined if a representational framework exists in which the creator/cataloger has the ability to address the realities of this medium's artistic creation and representation process: specifically, the highly interactive nature of the work, the variable quality of output, and technically complex interactions within the work itself must be addressed. By embracing the paradigm that includes re-interpretation as a valid preservation strategy, digital preservationists will need a thorough and consistent representation of the original from which to work.

In mid-2004 Richard Rinehart delivered a paper arguing that new media art is more like a musical performance than it is like a painting or a book, and therefore more appropriately represented by a scoring system than the text-centric methods used today (Rinehart). His proposal of a media art notation system (MANS), based on the MPEG-21 framework, is a welcome step forward in the development of a viable preservation schema for these highly variable and ephemeral objects. Rinehart's system, however, is more of a metadata framework or ontology than a scoring system, and therefore runs into the same problems inherent in any text-based representational framework that attempts to describe or define a non-textual entity (Svenonius), and is based on the 'conduit-metaphor' model of communication (Davis).

## Musical Scores

**T**here are numerous examples of existing art forms that employ re-interpretation as the de facto means of delivery and representation: namely; drama, for which scripts represent the work; dance, which often makes use of a notational system to describe and record body movements across space and time

(Labanotation being the primary example); and music, which is usually represented in the West using the common notation system (CNS). In order to develop a less text-centric representational model for new media art, I chose to explore the representation and interaction techniques of these art forms, which have interpretable or variable output.

Music, drama and dance each share characteristics with new media art. They're temporal: meaning they take place and change over time; they're performance based and ephemeral: the performance is the instantiation of the work, and once that performance is over, it's gone; and they're open to interpretation within some pre-determined set of values: although it is possible and expected to interpret freely, each form has a framework within which the director/choreographer/conductor must work. However, music has traits uniquely shared with new media art. For example, music and new media art both share a 'unity of artistic vision', basically meaning that all instruments, in the case of an orchestra; and libraries, programs, data sources, in the case of a new media art object, interact in a specific way to produce the final product. Another shared trait between music and new media art is the existence of a level of abstraction that is not necessarily present in drama or dance. Both the composer and the programmer use instruments, or tools, to achieve his or her artistic vision, whereas a playwright or choreographer works with the more immediately available bodies in motion and/or words as their means of expression. Finally, a play or dance does not inherently depend on the availability of tools to exist, whereas an orchestral / new media art performance most often does.

In the hopes of developing a set of characteristics to include in a new media art notation system, I have started exploring the information contained in musical scores. In addition to recording the fixed musical elements like pitch, rhythm, tempo, dynamics, and articulation; this research also seeks to understand musicians' interpretative decisions, as well as their interactions with each other, and with the score. Interpretation and interaction are particularly interesting for the purpose of this research, because these are the primary means by which musicians achieve artistic, reliable performances, and that is the ultimate goal of any new media representation framework. Although it might be difficult to completely understand musicians' interpretative choices and interactions, we believe that the personal notes (annotations) musicians' make on the scores themselves can provide valuable information regarding these transient and personal decisions (Marshall).

## Research questions

**S**ome of the questions this phase of the research seeks to answer: Which musical elements must be regulated, which can be improvised, and which must be freely interpreted? Is



this dependant on context of presentation? Can any of these musical elements transfer usefully to representation of new media art? Under what circumstances does a musician decide to do something that is not in the published score? How do composers communicate over time, space, and cultures, their intent and goals regarding performance? Are there different models of interpretation based on different musical styles or genres? How do composers, conductors, and musicians react to the idea of official representation of interpretation?

## Methodology

With these issues in mind, we developed an experimental framework consisting of a musician/score collection methodology; a coding schema, which will help categorize the annotations; and a method for the systematic exploration of annotations.

The musician/score hierarchy defines from whom and what kinds of musical scores this study will consider. It is a structural arrangement of four parallel interests and skills of each musician: first is the three 'levels' of music-makers: musician, conductor, and composer; second is the level of proficiency: amateur, college-level, and professional; third, we take into consideration the presence of a conductor: orchestras versus quartets, for example; and finally, the hierarchy includes a consideration of style of music – jazz, classical, and musical theater.

After collecting the scores, we are marking up any annotations the musicians/composers/conductors might have made on them. There are two types of mark-up: structural, and content-based. At the structural level, we decided to mark up at the bar level, delineating the extent of bars and phrases, where appropriate. At the content level, there are three types of written notes: textual, where the musician has actually written a word in the margins ( "Less Bow!!!" or "FROG!" ); symbolic, where the musician has written non-textual symbols (stars, exclamation points, and glasses); and numeric, where the musician has put numbers above or below notes for fingering or timing instructions, or numbered the bars if that information isn't included in the published score.

At this point in the process we conduct interviews with the participating musicians, asking questions regarding the process, and context of annotation behavior (MacMullen). The purpose of the interview is to get a deeper understanding of musicians' attitudes toward interpretations and whether their annotation behavior is in fact an important element in understanding that interactive quality of musical performance.

The final step in this process is to analyze the annotated scores, looking for n-way consensus, investigating any 'important' or consistently documented sections or elements within a piece.

I will normalize at the basic unit of annotation (in this case at the bar and phrase level); I'll record all instances of annotation: who, where, what kind; and all count all instances of annotation to provide percentages at the bar and phrase level between and among musicians and musical types. Finally, I'll conduct consensus analysis to determine how often annotations concur on selections.

## Initial Findings

After concluding a pilot study, exploring the initial data from a university-level orchestra, and interviewing several musicians and conductors, initial findings indicate that the premises upon which this study are based are valid. Annotation of the score provides insight into a number of different issues relevant to the development of a notation schema for new media art: annotation analysis identifies those characteristics of musical notation important across musician types and skill level, allowing me to make recommendations for the inclusion of certain characteristics in the new notation system. Annotation also evidences performance-related interaction between and among musicians. Finally, annotation is a physical sign of an individual musician's interaction with, and interpretation of, the score itself.

This work is ongoing. Data collection will end in April 2005, when more comprehensive findings will be available.

## Acknowledgments

This work was partially funded by an unrestricted research gift from *Microsoft Research* to the *Annotation of Structured Data* research team in the School of Information and Library Science at the University of North Carolina at Chapel Hill, whose members contributed to this work: Gary Marchionini, Paul Solomon, and Catherine Blake, co-PIs; with team members Tom Ciszek, Xin Fu, Lili Luo, W. John MacMullen, Cathy Marshall, Mary Ruvane, and David West. The project's website is available at: <http://ils.unc.edu/annotation/>.

## Bibliography

Davis, Marc. "Theoretical foundations for experiential systems design." Paper presented at ACM SIGMM Workshop on Experiential Telepresence (ETP) 2003. New York: ACM Press, 2003.

Depocas, Alain, Jon Ippolito, and Caitlin Jones. *Permanence through change: The variable media approach*. New York, NY: Guggenheim Museum Publications, 2003.

MacMullen, W.J. "Annotation as Process, Thing, and Knowledge: Multi-domain studies of structured data annotation." Paper delivered at ASIST Annual Meeting, Charlotte, NC (October 28 - November 2, 2005). In Review.

Marshall, Catherine C. "Toward an ecology of hypertext annotation." Paper presented at Hypertext98, Pittsburgh, PA, 1998. New York: ACM Press, 1998.

Rinehart, Richard. "A system of formal notation for scoring works of digital and variable media art." Paper presented at AIC Digital Media Group 2004 Annual Meeting, (Portland, Oregon: June 14, 2004). 2004.

Svenonius, Elaine. "Access to nonbook materials: The limits of subject indexing for visual and aural languages." *Journal of the American Society of Information Science and Technology* 45.8 (1994): 600-606.

---

## Texttechnologie in der Universitären Lehre

---

*Andreas Witt* ([andreas.witt@uni-bielefeld.de](mailto:andreas.witt@uni-bielefeld.de))

*Bielefeld University*

*Dieter Metzling* ([dieter.metzling@uni-bielefeld.de](mailto:dieter.metzling@uni-bielefeld.de))

*Bielefeld University*

---

### Einleitung

An der Universität Bielefeld (eine deutsche Universität mit ca. 16 000 Studierenden) ist an der Fakultät für Linguistik und Literaturwissenschaft seit wenigen Jahren der Bereich Texttechnologie in der universitären Lehre vertreten. Die Studienmöglichkeiten dieses Faches, die in dieser Form einzigartig sind, sollen auf einem bilingualen Poster (Englisch und Deutsch) erstmals auf einer wissenschaftlichen Tagung vorgestellt werden. Wir sind davon überzeugt, dass unsere Erfahrungen für die Fachkolleginnen und Kollegen an anderen Instituten sehr interessant sind, da es sich um eine innovative curriculare Umsetzung von Inhalten aus dem Bereich Humanities Computing handelt.

Die Texttechnologie an der Universität Bielefeld beschäftigt sich unter anderem mit Eigenschaften von Texten sowie der Modellierung und der Strukturierung textueller Information. Institutionell ist die Texttechnologie dem Arbeitsbereich Computerlinguistik und damit mittelbar dem Fach Linguistik zugeordnet. In der Forschung ist die Texttechnologie seit der Mitte der 1990er Jahre an der Universität Bielefeld vertreten. Zum Wintersemester 1999/2000 begann die Etablierung der Texttechnologie in der Lehre, und zwar in Form eines Magisternebenfachstudiengangs. Dies bedeutete, dass Texttechnologie mit einem geisteswissenschaftlichen Hauptfach und einem weiteren Nebenfach verbunden werden musste. Ein vollständiges Magisterstudium dauert in der Regel 9-10 Semester. Vor ca. zwei Jahren wurden u.a. auf Grund von Internationalisierungsbestrebungen die Magisterstudiengänge in Bielefeld abgeschafft und die Studienmöglichkeiten auf Bachelor- und Masterstudiengänge umgestellt. Dies ist Teil einer generellen Entwicklung in Europa, die 1999 in der Bologna Deklaration von 29 europäischen Ländern fest vereinbart worden ist. Im Zuge dieses Prozesses sollen bis 2010 alle Hochschulstudiengänge auf Bachelor und Master umgestellt worden sein.

Texttechnologie kann seit dem Wintersemester 2002/2003 als eigenes B.A. Nebenfach studiert werden. Dieses Nebenfach kann mit unterschiedlichen B.A.-Kernfächern, z.B. Germanistik, Anglistik, Geschichte, Sozialwissenschaften oder Philosophie kombiniert werden. Die Lehrinhalte des Nebenfachstudiengangs sind modular organisiert.

Nachfolgend sollen diese Module kurz beschrieben werden.

## Module des Studiums

### Grundlagen der Texttechnologie

*Computerpropädeutikum (3 Leistungspunkte)*

*Einführung in die Texttechnologie (4 Leistungspunkte)*

*Textstruktur und Textsatz (4 Leistungspunkte)*

*Hypertext (4 Leistungspunkte)*

Im ersten Studiensemester findet ein Computerpropädeutikum statt, das grundlegende praktische Kenntnisse im Umgang mit Computern vermittelt und dabei auf die Verarbeitung von Textdaten ausgerichtet ist (Betriebs- und Dateisystem, Texteditoren, reguläre Ausdrücke). In der Veranstaltung *Einführung in die Texttechnologie* werden die Eigenschaften elektronischer Texte und die Methoden ihrer Erstellung behandelt. Der Begriff des elektronischen Dokuments, die Trennung von Struktur und Layout, die Zeichencodierung sowie Typographie werden ebenfalls angesprochen. Diese Thematiken werden in der zweistündigen Veranstaltung *Textstruktur und Textsatz* weiter vertieft. Das Seminar *Hypertext* behandelt u.a. Grundlagen dieser Textsorte (von den verschiedenen Begriffsdefinitionen und bis zur Geschichte), konkrete, d. h. praktisch angewendete, Hypertextsysteme und textlinguistische Themen wie z.B. Kohärenz in Hypertexten.

### Formale Methoden der Linguistik

*Formale Methoden I, II, III (je 3 Leistungspunkte)*

Methoden aus der Mathematik, der Informatik und der Logik werden in drei teilweise aufeinander aufbauenden Seminaren besprochen. Zu den Methoden zählen die Mengentheorie als Sprache zur Formulierung von Strukturen, der Umgang mit der Hierarchie der formalen Sprachen und deren Grammatiken (der Chomsky-Hierarchie) und die Verwendung formaler Automaten. Des Weiteren wird in die Aussagen- und Prädikatenlogik eingeführt. Darüber hinaus werden Attribut-Wert-Strukturen und deren Relationen vorgestellt und deren Operationen (vor allem die Unifikationsoperation) eingeübt.

### Programmierung für die Texttechnologie

*Einführung in die Programmierung (12 Leistungspunkte)*

Am Beispiel einer konkreten, jeweils aktuellen und für das Humanities Computing relevanten Programmiersprache wird sich in diesem Modul mit den grundlegenden, allgemeinen Prinzipien der Programmierung auseinandergesetzt. Die Vermittlung der programmiersprachlichen Paradigmen steht neben dem Erlernen einer konkreten Programmiersprache (z.B. Java, Perl, PHP oder Prolog) im Zentrum des Moduls. Anwendungsschwerpunkt der Beispiele und der Übungen bildet die Verarbeitung (z.T. strukturiert annotierter) Textdaten. Eine sehr umfangreiche Programmieraufgabe muss *nach Abschluss* des Seminars von Studierenden in kleinen Gruppen gelöst werden.

### Informationsstrukturierung und Auszeichnungssprachen

*Auszeichnungssprachen (4 Leistungspunkte)*

*Informationsstrukturierung (8 Leistungspunkte)*

Die Auszeichnungssprachen (engl. *Markup Languages*) zählen zu den grundlegenden Standards der Texttechnologie. In dem Seminar *Auszeichnungssprachen* wird XML, aber auch SGML, besprochen, wobei die Themen maschinelle Verarbeitung von ausgezeichneten Dokumenten, Schemasprachen, Parser, Transformation und Formatierung von besonderer Relevanz sind.

In dem Seminar *Informationsstrukturierung* werden die dann bekannten Auszeichnungssprachen verwendet, um Informationen, die aus sehr unterschiedlichen Domänen stammen, zu strukturieren. Hierbei wird den Studierenden die Möglichkeit geboten, die Inhalte ihres Kernfachs (bei nahezu allen Studierenden ist dies eine Geisteswissenschaft) als allgemeine Information aufzufassen und zum Gegenstand ihrer Informationsmodellierung werden zu lassen. Die Richtlinien der Text Encoding Initiative werden in diesem Zusammenhang genauer angesehen. In dem Seminar werden Kleingruppen gebildet, die — *kollektiv* — bestimmte Informationsmodellierungsaufgaben lösen und der gesamten Gruppe präsentieren.

### Wahlpflichtmodule

Neben den bisher genannten Modulen, die alle studiert werden müssen, ist von den Studierenden noch eines von drei weiteren Modulen zu wählen. Sie können sich entscheiden, ob sie für ca. 3 Semester Sprachkurse (insbesondere Sprachen mit nicht-lateinischen Schriftsystemen, wie Japanisch oder Arabisch) besuchen möchten, ob sie sich mit Empirischen Methoden (u.a. Statistik) oder mit den linguistischen

Beschreibungsebenen (z.B. Semantik und Syntax) intensiv beschäftigen möchten.

## Studieninfrastruktur und weitere Studienmöglichkeiten

Die Mehrzahl der Seminare enthält sehr praxisnah ausgerichtete Lehrkomponenten. An der Fakultät für Linguistik und Literaturwissenschaft ist in engem Zusammenhang mit dem Aufbau des Studiengangs eine Infrastruktur errichtet worden, die eine derartige praxisnahe Ausbildung ermöglicht. So stehen zwei Multimediaseminarräume (mit 40 und mit 80 Arbeitsplätzen an 20 bzw. 40 Rechnern, einem didaktischen Netz, einem Intranet und selbstverständlich mit Internetanbindung), Rechnerarbeitsplätze für das Selbststudium, sowie Arbeitsplätze im Hochschulrechenzentrum zur Verfügung. Erst durch diese Infrastruktur wurde es möglich, die Praxisnähe auch bei der derzeitigen durchschnittlichen Seminargröße von 60 bis 80 Personen sicherzustellen.

Neben dem B.A.-Nebenfachstudiengang ist es auch möglich, Texttechnologie als so genanntes Profil im B.A. Studiengang Linguistik zu studieren. Darüber hinaus wird an der Universität Bielefeld der M.A.-Studiengang "Interdisziplinäre Medienwissenschaft" angeboten, der fakultätsübergreifend verankert ist. Dieser modular organisierte Masterstudiengang, der auch texttechnologisches Modul beinhaltet, stellt eine von mehreren weiterführenden Studienmöglichkeiten dar, die von den Absolventen des Texttechnologiebachelorstudiengangs gewählt werden können. Es ist natürlich auch möglich nach dem B.A.-Abschluss einen Beruf zu ergreifen. Da sich insbesondere auch die klassischen Berufsfelder für Geisteswissenschaftler/innen (z. B. in den Verlagen oder im Bereich Journalismus) durch den massiven Einzug der Informationstechnologie in den vergangenen Jahren sehr stark verändert haben, sind die Absolventen der Texttechnologie auch für diese (bei den Studierenden der Geisteswissenschaften besonders beliebten Berufe) sehr gut vorbereitet. Es existieren zwar bisher keine Verbleibestatistiken der Studierenden, erste Erfahrungen zeigen aber, dass das Nebenfach Texttechnologie den Studierenden tatsächlich das berufsrelevante Wissen vermittelt, das sich viele Firmen von Absolventen geisteswissenschaftlicher Studiengänge wünschen.

## From Text to Topics — Zigzagging Towards the Knowledgebase of Tang Civilization

---

*Christian Wittern* ([wittern@zinbun.kyoto-u.ac.jp](mailto:wittern@zinbun.kyoto-u.ac.jp))  
*Kyoto University*

---

### Abstract

A few years ago, the prospect of having access to a large amount of digitized data promised to give a completely new direction to the field of Chinese Studies. Although today we have such databases as the Siku Quanshu (四庫全書) Fulltext Database<sup>1</sup>, as well as many other texts, some of them even freely available on the Internet, the benefits of this has been limited. There are many reasons for this, not all of them technical. Of the technical reasons, the limited, idiosyncratic interface that each of these database provides, and the unstructured data it operates on are probably the most important ones.

The *Knowledgebase of Tang Civilization* is an attempt to remedy this situation, at least for material relating to the Tang, by providing a comprehensive electronic archive of information about China during the period of the Tang dynasty (618-907 A.D.) in a way that allows new ways to access, analyze and expand the information. Work on this Knowledgebase started in 2003<sup>2</sup> with initial funding for 5 years. This presentation will present some of the experiences gained in the first development phase.

The design of the Knowledgebase uses a two layer model, that distinguishes the 'information layer' from the 'resource layer'. The organization of the information layer is based on the topic map paradigm.<sup>3</sup> to allow for the expression of ontology subtrees, with links from the information layer back to the resource layer, which will hold primary sources.

Its main point of access for researchers will be a web application, but other interfaces will be developed.

Initially most of the information to be included will be textual, but will in due time be enhanced by images, visual reproductions of objects, digital maps and animations of events. The distinguishing feature of the knowledgebase is the way

information items are interconnected in a flexible and innovative way.

The information in the knowledgebase will be organized along the following information axis:

- Personal names, dates and activities of people of the Tang.
- Placenames and georeferences to their locations, administrative geographical units, digital maps.
- Works created during the Tang, including texts, artefacts and buildings
- Calendar and time
- Events of importance and influence

Obviously, many if not all information items will be accessible through more than one of these axes; internally they are cross-linked and form more of a web-like structure. Additionally, these items are organized in hierarchical ontologies. This allows to access the information also based on their position within the hierarchy, or on the relation with other items. For geographical locations, like a city for example, such a hierarchy would consist of the upper administrative units it belongs to; for persons this could consist of the family line, but also the region of origin, the school or tradition of thought, in the case of monks also the ordination line and line of transmission.

The challenge in the first phase of the development, which will be concluded by the time this presentation will be given, was to design a way to bootstrap the Knowledgebase. For this purpose, two dynastic histories (the *Jiu Tang Shu* (945) and the *Xin Tang Shu* (1060)) and one chronologically arranged historical account by Sima Guang (*Zizhi Tongjian*, 1084) have been chosen to provide a basic set of information about the Tang period. This idea relies on the fact, that the dynastic histories do not only provide a day to day chronicle of court affairs and other events, but also include monographs on a variety of subject matters, including geography (with detailed accounts of administrative units, their changes in size and denomination, local production, population etc.), calendar (including accounts of the calendar systems in use), ritual observances, music, astronomy, offices, state finances, law and a detailed bibliography of works known to have written in that period. In addition to that, more than half of the text of the official histories is taken up by biographic accounts. In the case of the Tang, there are two such histories, since in the eyes of Ouyang Xiu, the editor of the second, "new" history, the first one had some defects in style and presentation.

The texts are encoded in XML using the TEI vocabulary. In a first phase, only structural encoding was applied, so that the texts could be accessed using XML technologies<sup>4</sup> and could be further processed. It was then started to add semantic markup to allow for automatic extraction of information.

It should be obvious, that this collection provides rich material that could be mined for inclusion in the Knowledgebase, but the challenge was to find an efficient way to mine that information, generate topics from them and relate them to each other in the way outlined above. The presentation will focus on the strategies employed and results achieved and will then try to look at how to generalize these methods. It is also planned to show a prototype of an interface, that allows further enhancement of the data.

1. A database published by Chinese University Press, that includes on 176 CD-ROMS an electronic text of the anthology *Siku Quanshu*, which was compiled in China in the 17th century and takes 1500 volumes in the modern reprint.
2. More information on this project can be found at <http://coe21.zinbun.kyoto-u.ac.jp/> the website of the Institute for Research in Humanities, COE 21 section
3. As defined in ISO 13250 (International Organization for Standardization)
4. Most of this had been done semi-automatically. Just to give an idea of the amount of the material, the size of the files with only very basic encoding applied runs at this moment to well above 30 MB.

## Bibliography

ISO. *ISO/IEC 13250, Information technology - SGML Applications - Topic Maps*. Geneva: ISO, 2000.

Liu Xu. *舊唐書 (Jiu Tang Shu) (945)*. Beijing: Zhonghua Shuju, 1975.

Ouyang Xiu. *新唐書 (Xin Tang Shu) (1060)*. Beijing: Zhonghua Shuju, 1975.

Sima Guang. *資治通鑑 (Zizhi Tongjian) (1084)*. Beijing: Zhonghua Shuju, 1956.

*Wenyuange Siku quanshu dianziban*. Hong Kong: Chinese University Press, 1998.

## Reading Potential: The Oulipo and the Meaning of Algorithms

---

Mark Wolff (wolffm0@hartwick.edu)  
Hartwick College

---

Recent efforts to reconceptualize text analysis with computers in order to broaden the appeal of humanities computing have invoked the example of the Oulipo, a group of writers in France that invent 'potential' ways to create literature using rigorous formal constraints. Rejecting the practice of using computers as tools for objective, empirical research with texts, Stephen Ramsay envisions an algorithmic criticism that transforms texts for "the purpose of releasing what the Oulipians would call their 'potentialities'" (Ramsay 172). Stéfan Sinclair has developed HyperPo as a web-based tool for helping scholars read and play with texts using procedures inspired by the Oulipo. The idea of playing with texts using computers is pursued further by Geoffrey Rockwell who calls for the creation of web-based playpens where scholars can experiment with tools and discover the potentialities inherent in the practice of humanities computing.

Although there are similarities between the activities of the Oulipo and the new approach to computer-assisted literary analysis, the development of tools for the express purpose of encouraging scholars outside of humanities computing to play with texts does not follow the model of Oulipian research into potentialities. For the Oulipo, the invention of procedures for playing with texts is not necessarily a means to greater engagement with literature: it is its own end, an intellectual activity that invites application but does not require adoption by others as an indication of success. According to Raymond Queneau, one of the founding members of the Oulipo and author of the *Cent mille milliards de poèmes*,

The word 'potential' concerns the very nature of literature; that is, fundamentally it's less a question of literature strictly speaking than of supplying forms for the good use one can make of literature. We call potential literature the search for new forms and structures that may be used by writers in any way they see fit.  
(Oulipo 1986, 38)

Queneau makes it clear that what the Oulipo does relates to but does not constitute literary creation. Writing is a derivative activity: the Oulipo pursue what we might call speculative or theoretical literature and leave the application of the constraints to practitioners who may (or may not) find their procedures useful. According to François Le Lionnais, another founding member, a method for writing literature need not produce an

actual text: "method is sufficient in and of itself. There are methods without textual examples. An example is an additional pleasure for the author and the reader" (Bens 81, my translation).

The Oulipo did not articulate a clear statement explaining potential methods for reading literature, but we can extrapolate a definition from how they described their efforts to invent methods for writing literature. *Potential text analysis is less a question of interpreting literature than of supplying algorithms for the good use one can make of reading. Producing exemplary interpretations with algorithms is a secondary consideration.* It follows that the interpretation of texts using a computer should not be in and of itself the objective of the new computer-assisted text analysis. The objective should be the invention of algorithms that scholars may (or may not) use, according to their own interests. The potentiality (as opposed to the reality) of computers as tools for text analysis implies that scholars engaged in the derivative activity of interpreting literature may not find such methods useful.

When the Oulipo formed in 1960, one of the first things they discussed was using computers to read and write literature. They communicated regularly with Dmitri Starynkevitch, a computer programmer who helped develop the IBM SEA CAB 500 computer. The relatively small size and low cost of the SEA CAB 500 along with its high-level programming language PAF (Programmation Automatique des Formules) provided the Oulipo with a precursor to the personal computer. Starynkevitch presented the Oulipo with an "imaginary" telephone directory composed of realistic names and numbers generated by his computer. He also programmed the machine to compose sonnets from Queneau's *Cent mille milliards de poèmes*. In both cases the Oulipo was impressed but did not believe these computer applications had 'potential'. What worried the Oulipo was the aleatory nature of computer-assisted artistic creation: they sought to avoid chance and automatisms over which the computer user had no control (Bens 147-148). In 1981 the Oulipo published *Atlas de littérature potential* where they described some of the computer applications they devised for reading literature. Their early experiments included machine-assisted readings of the *Cent mille milliards de poèmes* and Queneau's *Un conte à votre façon*. The algorithms used to read these texts provided a certain degree of interaction between the user and the machine but did not reveal unforeseen potentialities. Some members of the Oulipo formed ARTA (*Atelier de Recherches et Techniques Avancées*) and ALAMO (*Atelier de Littérature Assistée par la Mathématique et les Ordinateurs*) to explore computer-assisted writing, but the Oulipo itself has not further pursued methods for reading texts with machines.

This is not to say the Oulipo abandoned the idea of potentialities in reading. There are at least two examples of original

algorithms developed by Oulipians for reading texts. The first is Harry Mathews's Algorithm, which consists of combinatoric operations over a set of structurally similar but thematically heterogeneous texts. These operations generalize the structure of the *Cent mille milliards de poèmes* and allow for the production of new texts. Mathews notes that the algorithm works not only with letters, words and phrases but with entire works, entire oeuvres, entire literatures, entire worlds. Creating a computer program based on this algorithm ( <http://bumppo.hartwick.edu/Oulipo/Mathews.php> ) is relatively simple, but its interest does not lie in its application. According to Mathews, the aim of the algorithm "is not to liberate potentiality but to coerce it" (Oulipo 1986, 139). A 'new' reading of a text (or a reading of a 'new' text) through the algorithm is not the objective. The use of the algorithm is meaningful in that the apparent unity of texts can be dismantled by the algorithm and give way to a multiplicity of meanings. Mathews invented a system of constraints that illustrates what deconstructionists have maintained for decades.

The second example is Raymond Queneau's matrix analysis of language, published in *Etudes de linguistique appliquée* and discussed at length during one of the Oulipo's early gatherings. Using principles of linear algebra, Queneau devised a mathematics of the French language that could describe the structure of texts and provide statistical "indices of an author's style that may be interesting, for they escape the conscious control of the writer and doubtless depend on several hidden parameters" (Queneau 319, my translation). Queneau himself provided analyses of a number of short sample texts. His ability to apply the algorithm to lengthy texts was limited, however, because he did his calculations 'by hand': he did not use a computer. With the availability of part-of-speech taggers such as Helmut Schmid's *TreeTagger*, it is easy to use a computer to perform a matrix analysis of any text written in French ( <http://bumppo.hartwick.edu/Oulipo/Matrix.html> ). Matrix analysis may prove useful for authorship attribution in combination with other techniques, such as the use of Markov chains proposed by Khmelev and Tweedie. Queneau, however, expressed greater interest in the algorithm's mathematical properties: he proved several theorems on the behavior of matrices and identified similarities between them and the Fibonacci series. The members of the Oulipo were intrigued by matrix analysis but looked forward to the creation of poems written in columns and rows rather than the transformation of existing poems into matrices (Bens 236-237).

Mathews and Queneau offer two algorithms we can operationalize with computers for literary analysis, but the interest of the algorithms lies not in what they help us see in a given text but in the way they invite us to play rigorously for play's sake. Oulipian constraints on reading are better understood as toys with no intended purpose rather than as tools we use with some objective in mind. These procedures for

making sense of texts provide for their own interpretation: they are not instruments for meaning but reflections on meaning itself.

## Bibliography

- Bens, Jacques. *Oulipo: 1960-1963*. Paris: Bourgois, 1980.
- Khmelev, Dmitri V., and Fiona J. Tweedie. "Using Markov Chains for Identification of Writers." *Literary and Linguistic Computing* 16.3 (2001): 299-307.
- Oulipo. *Atlas de littérature potentielle (1981)*. Paris: Gallimard, 1988.
- Oulipo. Ed. F. Motte Warren Jr. *A Primer of Potential Literature (1986)*. Trans. F. Motte Warren Jr. Normal, IL: Dalkey Archive Press, 1998.
- Queneau, Raymond. *Cent mille milliards de poèmes*. Paris: Gallimard, 1961.
- Queneau, Raymond. "L'Analyse matricielle du langage." *Etudes de linguistique appliquée* 3 (1964): 37-50.
- Queneau, Raymond. *Bâtons, chiffres et lettres*. Paris: Gallimard, 1965.
- Ramsay, Stephen. "Toward an Algorithmic Criticism." *Literary and Linguistic Computing*. 2003.
- Rockwell, Geoffrey. "What is Text Analysis, Really?" *Literary and Linguistic Computing* 18.2 (2003): 209-219.
- Schmid, Helmut. *TreeTagger - a language independent part-of-speech tagger*. Institute for Natural Language Processing, University of Stuttgart. Accessed 2005-03-03. <http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/DecisionTreeTagger.html>
- Sinclair, Stéfan. "Computer-Assisted Reading: Reconceiving Text Analysis." *Literary and Linguistic Computing* 18.2 (2003): 175-184.
- Sarynkevitch, Dmitri. "The SEA CAB 500 Computer." *Annals of the History of Computing* 12.1 (1990): 23-29.

## **The *Online Nahuatl Dictionary*: A Model for Interdisciplinary Multicultural Collaboration**

---

**Stephanie Wood** (*swood@uoregon.edu*)

*Co-Director*

**Judith Musick** (*musick@uoregon.edu*)

*Co-Director*

**William Henderson**

(*willhen@darkwing.uoregon.edu*)

*Graduate Fellow*

---

**N**ahuatl, one of the world's indigenous languages, is facing serious threats to its survival. Unjust rural land use patterns, the lure of perceived urban employment opportunities, and lingering colonial prejudices are but some of the factors contributing to the rapid loss of the Nahuatl language in highland central Mexico today. Classical Nahuatl, as it was written and spoken until about 1800, is already extinct. At the end of the eighteenth century native scribes and notaries who were proficient in written Nahuatl had shifted to the use of written Spanish, the language of the colonizers. Nahuatl became an oral language, and fewer and fewer Nahua individuals were able to read the thousands of manuscripts written in Nahuatl (using the Roman alphabet, after about 1540) their forebears had left on archive shelves. The *Online Nahuatl Dictionary* project proposes a resource to help reverse these trends: a reference for modern Nahuas and other interested parties to gain access to the indigenous language, Classical and modern, for self-education or other scholarly purposes.

This trilingual dictionary project (Nahuatl, Spanish, and English) has many facets. It involves the participation of an institute in Zacatecas, Mexico, where John Sullivan, a teacher of Spanish and Nahuatl recruits Nahua university students from the Huasteca region. These students receive training in written Nahuatl (Classical and modern) and Spanish and collaborate in the building of the dictionary, entering terms from their current, everyday speech. They are also entering sixteenth and seventeenth-century Nahuatl vocabularies compiled by Spanish priests and their informants, entering them into the dictionary database. Finally, they are taking courses in cultural history, paleography, and linguistics, which help prepare them for their Classical Nahuatl manuscript studies.

The plan is to obtain funding (we have applied for the NEH/NSF grant Documenting Endangered Languages) to underwrite a larger number of scholarships, to expand the student body and diversify it, inviting the participation of Nahuas from additional regions. This, in turn, will diversify the dictionary base, with vocabulary, pronunciation variations, and broadening perspectives from places such as the modern state of Guerrero. It will also help spread a Nahuatl literacy movement to different parts of the country. It will prepare more people to better understand their histories by studying manuscripts from their communities of the past. These various students will not be consultants or human subjects but rather full participants in the project.

The dictionary project also involves the participation of ethnohistorians at the University of Oregon, Stephanie Wood and Robert Haskett, who have training in Classical Nahuatl. Wood and Haskett have considerable experience translating manuscripts and have access to additional colleagues working in the same field. They are selecting terminology and contextual information from recent manuscript translations and are adding this material to the dictionary, with full citations that will point philologists back to the original sources. They enter terms just as they find them in the original manuscripts, capturing a range of orthographic variation, but they also include in the database any known standardized versions to facilitate the greatest possible success for searching. (Because of the diverse ways of writing Nahuatl, we often lack a single lemma that all would agree upon as a universal dictionary form.) Wood and Sullivan will also be adding a linguistically marked-up rendition of each word, showing vowel length and glottal stops, for instance, in order to facilitate linguists' searches and research needs. They will follow the markup used in Frances Karttunen's *Analytical Dictionary of Nahuatl* (1983).

Sullivan, Wood, and Haskett will work together, in consultation with linguists such as Frances Karttunen, Johnathan Amith, and other colleagues in Mexico, to ensure that the English and Spanish-language search interfaces for the dictionary will meet the interests of multiple disciplines. Luis Reyes García, who recently passed away, was a model Nahua scholar who led translation workshops in Mexico City and Tlaxcala that we wish to emulate. One of his students, Raul Macuil, and another Nahua college graduate, Ignacio Silva Cruz, in Mexico City, are ready to help us test our materials. Sullivan and Wood will open up dialogue between the student and faculty groups and encourage a growing appreciation for what each is contributing. Sullivan will additionally guide the Nahua students in building the Nahuatl-language search interface.

The Nahuatl interface will be accessible from the trilingual dictionary home page but it will also work as a stand-alone site, as a monolingual dictionary aimed at modern speakers. The intent is to encourage written literacy and to document current



vocabularies and word meanings. It will help modern speakers recapture terms from Classical Nahuatl that have been lost, whether this might result in an expanded vocabulary in modern-day usage or only an aid to help with translating cultural heritage materials. Although there will be an emphasis on the written language, the Nahuatl portion of the dictionary will nevertheless offer users access to audio files, as well. Sound recordings will serve to preserve and document regional differences in pronunciation and provide access to vision-impaired users and to those who are still building their literacy.

As we are proposing it, the Online Nahuatl Dictionary project offers a unique model for interdisciplinary work that will help preserve an endangered indigenous language and, at the same time, create some of the tools needed for the analysis of neglected cultural heritage materials. It will bring together modern language consultants with linguists and ethnohistorians of various cultures (Spanish-speaking and English-speaking, to start) to document a rapidly disappearing language, reinstate a lost literacy for Nahuas, and establish a methodology for international and cross-cultural collaboration in both dictionary creation and manuscript transcription, translation, and analysis.

Our project has created an online environment for this international collaboration, allowing native students to work with native and non-native professors in building a multifaceted reference database of the Nahuatl Language. This resource can then be leveraged to provide features that will serve the participants' multiple purposes while simultaneously forging a relationship of greater equality and respect. A key tool to help facilitate this 'data collection' effort is the most recent version of *FileMaker Pro* (Version 7.0) and its 'instant web publishing' functionality. This technology gives us not only the flexibility to allow any of the project participants to remotely enter data, but also enables us to create various database interfaces on the fly from anywhere via the internet. With the collaborative nature of our project, this ability to alter the database and the various interfaces through which different audiences interact with it is crucial to addressing the needs of all the participants. In addition, the use of this technology saves us the cost of purchasing multiple database licenses, helps ensure the validity of the data by having just one 'collection point' for information and eases regular data backup. Furthermore, *FileMaker Pro 7.0* supports the use of Unicode, facilitating the typing of special characters that are so essential for written Nahuatl and Spanish and handles audio files easily, allowing sound to be uploaded from various locations.

Because the *Online Nahuatl Dictionary* is in the developmental phase, our presentation for the ACH/ALLC will highlight our goals and explore our intended methodology. We will appreciate and benefit from suggestions from the audience. We see our approach as unique in the way it combines linguistic with

humanities uses, in a distance research environment, to build a lexicon that will document and preserve an endangered language and help resurrect the voices of an extinct language.

## Bibliography

Karttunen, Frances. *Analytical Dictionary of Nahuatl*. Oklahoma: University of Oklahoma Press, 1983.

# English Usage Comparison between Native and non-Native English Speakers in Academic Writing

---

*Bei Yu (beiyu@uiuc.edu)*

*University of Illinois at Urbana-Champaign*

*Qiaozhu Mei (qmei2@uiuc.edu)*

*University of Illinois at Urbana-Champaign*

*Chengxiang Zhai (czhai@cs.uiuc.edu)*

*University of Illinois at Urbana-Champaign*

---

## Introduction

Discovering the differences in the language usage of native and non-native speakers contributes to contrastive linguistics research and second language acquisition. Unlike grammatical errors, the language usage differences are grammatically correct, but somehow do not conform to the native expressions. For example, if an English phrase is used popularly in a native English speaker group (G1) but not in a Chinese speaker group (G2), or vice versa, one possible reason may be first language (L1) impact on second language (L2). Such an impact may be due to grammar differences or even culture differences.

Isolating the language usage differences is never an easy task. There are generally two approaches to detect such differences: controlled experiments and corpus-based analysis. Many English as a Second Language (ESL) studies were conducted in a controlled environment, for example, through manually comparing the short in-class writing samples by a small number of college students in G1 and G2. The small data sets and the small subject groups undermined the result generalization because conflicting results were sometimes produced regarding the same language usage.

In contrast, corpus-based approaches (for example, computational stylistics) facilitate analysis of larger sample sets written by many subjects. Such approaches involve three major steps: 1) building a corpus of two comparable subsets; 2) automatically extracting a language usage feature set; 3) selecting a subset of the features that distinguishes G1 and G2.

Given a corpus and a feature set, we can transform the above task into a text categorization problem. The goal is to categorize the texts by the authors' language background. Picking the subset is then transformed into the feature selection problem in text categorization.

Oakes used Chi-square test to find vocabulary subsets more typical of British English or American English. Oakes' work focused on identifying two feature categories: (1) the features common in G1 but not in G2; and (2) the features common in G2 but not in G1, while we hypothesize the existence of a third discriminative feature category (3): features common in both G1 and G2, but with different usage frequencies. All the other features are considered irrelevant. Distinguishing these different categories of features allows us to discover subtle differences in the language usage.

In this paper, we propose a simple approach of comparative feature analysis to compare the language usage between native English speakers and Chinese speakers in academic writing. We first select candidate features that are common in at least one group and then categorize them into the above three feature categories. Features in category (1) and (2) are ranked by their differences in document frequency, and features in category (3) are ranked by their difference indices as defined in the next few paragraphs.

We use two heuristic constraints to select a "good" candidate discriminative feature:

- Constraint (1): The feature should be common within at least one of the groups.
- Constraint (2): The feature should be contrastive across the groups.

We use the normalized document frequency (DF) to measure the feature commonality within a group. DF means the number of documents containing this feature. Denote  $DF_1$  and  $DF_2$  as the document frequencies in G1 and G2, respectively, and  $T=0.3$  as a DF threshold. A feature falls into

- category (1) if  $(DF_1 - DF_2) \geq T$ ;
- category (2) if  $(DF_2 - DF_1) \geq T$ ;
- category (3) if  $|DF_2 - DF_1| < T$  and  $(DF_1 \geq T)$  and  $(DF_2 \geq T)$ .

Intuitively, the features in category (3) are those that are sufficiently popular in both G1 and G2 and have comparable DFs in G1 and G2.

After categorizing the features, we then define a Difference Index (DI) to measure the feature discriminating power and rank the features in category (3) by DI. We define TF as the number of occurrences of a feature in a document. Let  $m_1$  and  $m_2$  be the mean of the TF value of a feature in G1 and G2, respectively, we define DI as  $DI = \text{sig}(m_1 - m_2) * \max(m_1,$

$m_2)/\min(m_1, m_2)$ .  $\text{sig}(m_1 - m_2)$  is 1 if  $m_1 > m_2$  and is -1 otherwise. A positive DI means the feature is more heavily used in G1 and a negative DI means it is more popular in G2. The larger the  $|\text{DI}|$  is, the bigger the difference is.

## Corpus Construction

We propose the following fairness constraints for corpus construction: 1) the subjects in each group should have similar English proficiency; 2) the text genre and the topic should not interfere with the language usage analysis.

We collect two datasets that satisfy the constraints above. The first has 40 selected electronic theses and dissertations (ETD) from the ETD database at Virginia Tech., in which 20 are from Chinese students and 20 from American students, all from computer science and electronic and engineering departments to avoid genre interference. The second has 40 selected research articles downloaded from Microsoft Research (MSR), in which 20 are contributed by Chinese researchers in Beijing, China and 20 by their British colleagues in Cambridge, UK. The documents in ETD collection are long and each strictly attributed to one author, while the documents in MSR collection are much shorter and many are co-authored. We restrict the co-authors to the same language background. The biographies in theses and the resumes on MSR website help us identify the authors' first and second language.

## Feature Extraction

We extracted plain text from the original pdf and ps files. We limit our search for features at the lexical level because syntactic parsing does not perform well in such a technical writing corpus due to formulas, tables and figure captions, etc. In order to avoid topic interference, we choose a feature set popular in computational stylistic analysis (Koppel et. al.), the n-gram (we set  $0 < n < 4$ ) common word sequences (CWS). A common word list consisting of 626 functional words and some common content words for technical writing (e.g. "problem") is used to generate CWS. For example, in "SW4 is among the few known wireless system tools for in-building network design.", the following 3-gram CWS features are extracted: "is among the", "among the few", "the few known".

## Experiments and Results

We used the aforementioned procedure to analyze both ETD and MSR corpora. The results show that most "differences" found in one corpus do not repeat in the other one, but there are some "stable" features across the corpora.

Figure 1 lists the 1-gram, 2-gram and 3-gram CWS features in all the three categories.

Category	1-gram CWS	2-gram CWS	3-gram CWS
1	fact, might, placed, appropriate, course, seem, indeed, unfortunately, mean, allow, particularly, yet, could, look, never	this would, until the, to an, we would, fact that, note that, these are, to give, the fact, such that, would be, with this, could be, given the, along with, as to, and so, for which, which the, and not, to allow, it to, the appropriate, not the, a particular, can then, the form, then be, if a , of which, this provides, that all, where there	the fact that, can then be, this is a, to give a
2	according		
3 (positive)	specifying, being, would, itself, must, able, produce, useful, rather, question, were, allows, particular	a way, in which, form of , a given, must be, the use, in any, use of, a more, rather than	in which the, the use of
3 (negative)	novel, respectively, build	of different, show that, we may, and more, according to, and it, than that, we use, see that, is still, or the, is very, but also	used in the, show that the

Figure 1: 1-gram, 2-gram and 3-gram CWS in category 1, 2, and 3.

- G1: native English speakers (British/American); G2: Chinese speakers.
- Category 1: CWS common in G1 but not G2.
- Category 2: CWS common in G2 but not G1.
- Category 3: CWS common in both G1 and G2 but the frequencies differ.
- Category 3 (positive): CWS with mean frequency in G1 at least twice as that in G2.
- Category 3 (negative): CWS with mean frequency in G2 at least twice as that in G1.

An example in category (1) is the word "never". It appears in 75% American students theses in ETD but just in 20% Chinese student theses, and its frequency mean in the American group is six times as that in the Chinese group. Similarly, "never" appears in 45% MSR papers from UK, but just in 10% MSR papers from China, and British researchers use it five times more often than their Chinese colleagues.

Category (2) is surprisingly small with only one stable feature "according".

Category (3) includes features common in both groups. It has a positive subset consisting of features more heavily used in the British/American groups, and a negative subset consisting of features more popular in the Chinese groups. Examples in the positive subset are "specifying", "must", "were", "rather than", "the use of", etc. Examples in the negative subset include "novel", "respectively", "build", "show that", "according to", "used in the", etc.

We noticed that some feature groups are worth further study, such as the negation words, modals, personal pronouns, and parallelism indicators as listed in figure 2.

language usage aspects	representative features
negation	"nothing", "cannot", "not", "none", "no", "nobody", "nothing", "nowhere", "neither", "never"
modal	"can", "could", "may", "might", "will", "would", "shall", "should"
personal pronoun	"I", "my", "me", "you", "your", "us", "our", "we"
parallel	"and", "or", "nor", "but"

Figure 2: special groups of interesting features

As shown in figure 3, the native English speakers always use more negation words than the Chinese. It is probably due to culture difference rather than grammar impact.

Negation	ETD	MSR
cannot	1.53	MAX
not	1.77	1.09
never	MAX	MAX
is not	1.42	1.16
do not	1.88	MAX
not the	MAX	MAX
and not	MAX	MAX

Figure 3: comparison of negation words usage

Note: In figure 3, 4, and 5, a number (positive or negative) in a cell means the corresponding feature belongs to category (3) and the number is the feature's difference index (DI) value as defined in the paper. "MAX" instead of a number in a cell means this feature belongs to category (1) and thus it does not have a DI value. Similarly, "-MAX" means the feature belongs to category (2).

As shown in figure 4, the native speakers also use more modals such as "might", "would" and "could". The Chinese use "will" more often. There is no big difference between their uses of "can".

Modal	ETD	MSR
will	-1.41	-1.31
it will	-1.97	-1.50
we will	-1.72	-2.62
will be	-2.16	-1.61
we may	-4.04	-2.21
can	-1.08	-1.00
we can	-1.08	-1.16
which can	-1.24	-1.11
this can be	1.10	MAX
can be used	1.39	MAX
can be used to	1.60	MAX
can then	MAX	MAX
can then be	MAX	MAX
might	MAX	MAX
could	MAX	MAX
would	5.60	MAX
this would	MAX	MAX
we would	MAX	MAX
would be	MAX	MAX
could be	MAX	MAX

Figure 4: Comparison of Modal Usage

As shown in figure 5, "us", "our" and "we" are the three mostly used personal pronouns for both groups, but the native speakers use "us" more often while the Chinese use "our" and "we" more often.

Personal Pronoun	ETD	MSR
us	1.55	1.61
we would	MAX	MAX
we are	1.29	MAX
that we	1.16	1.10
our	-1.05	-1.87
we	-1.16	-1.13
if we	-1.44	-1.36
we may	-4.04	-2.21
we can	-1.08	-1.16
we use	-3.19	-MAX
in our	-1.81	-2.96
we also	-1.04	-2.29
we will	-1.72	-2.62
we use the	-1.86	-MAX

Figure 5: Comparison of personal pronoun usage

"And" and "but" are two parallel structure indicators commonly used by both groups. The Chinese group use "and" slightly more than the American/British group, but they use "but" almost twice as less than the native English speakers.

## Conclusion and Future Work

We use a simple comparative feature analysis method to compare the differences in the English common word usage between the native British/American English speakers and the Chinese speakers in their academic writing. The proposed method helps us find some interesting or even surprising differences between these two groups. We also see that common words are a very limited feature set. We shall explore more meaningful linguistic features to find more useful differences.

## Bibliography

- Koppel, M., S. Argamon, and A.R. Shimoni. "Automatically Categorizing Written Texts by Author Gender." *Literary and Linguistic Computing* 17.4 (2003): 401-412.
- Oakes, M. "Text Categorization: Automatic Discrimination between US and UK English using the Chi-square Text and High Ratio Pairs." *Research in Language* 1 (2003): 143-156.

## Play and Code in Humanist Research

Vika Zafrin (zafrin@brown.edu)  
Brown University

### 1. VHL: An Introduction

The Virtual Humanities Lab is a new humanities computing project at Brown University. We focus on two areas of research. First, we are designing and building a web-based engine for the presentation of semantically encoded primary texts, and for further annotation of these texts by invited scholars. Together with this engine we will be publishing several annotated texts. This engine is complemented by a weblog and by a discussion forum; both of these invite input from anyone interested.

We are in the process of semantically encoding, annotating, and publishing online three early modern Italian texts. The first and largest, Giovanni Boccaccio's *Esposizioni sulla Comedia di Dante* is the vernacular, originally oral text of Boccaccio's unfinished lecture series on Dante's *Commedia*. Giovanni Villani's *Nuova Cronica* (of which we are publishing a part) is an extensive account of Florentine history up to 1348. It is written in lively Italian, valuable not only for its record of events but also for its historiographical methods and political commentary. Finally, Giovanni Pico della Mirandola's *Conclusiones Nongentae* is an aphoristic Humanist text currently being developed as part of the Pico Project (<http://www.brown.edu/pico/>). All three of these will be presented electronically for the first time.

The sheer amount of information present in these texts — as well as their size, relative obscurity and general importance for the humanities — lend themselves to semantic encoding, collaborative annotation and electronic dissemination.

The number and variety of electronic tools being built for humanities research is ever-increasing. So is the learning curve for taking full advantage of these tools. Since semantic markup plays an increasingly important role in electronic humanities scholarship, having an idea of what it looks like and — broadly — how it functions seems to be an advantage for academics in the humanities.

Scholars who have never practiced semantic encoding of texts as a research tool, or performed complicated searches on semantically encoded texts, may find themselves reluctant to

spend time learning an unfamiliar way of working, even if the result of such learning may prove useful to them. Such researchers are a large part of our intended audience, and we are putting significant effort into writing clear, concise documentation and hands-on tutorials. The documentation is aimed at academics relatively new to humanities computing and, as such, will include a brief overview of the principles of semantic encoding as well as a guided tour of the VHL toolset. Our goal is to make these supplementary materials enjoyable and concise: we want scholars to receive just enough technical information to enable them to play with their texts.

## 2. Playing and Modeling

**M**ichael Mahoney says that a sufficiently complex idea for a piece of machinery cannot be described; the thing must be made or modeled. In order to be understood, complex texts should also be modeled.

A usable model of a text need not be comprehensive, but may rather address one or more specific issues. VHL researchers read the texts and use semantic encoding to arrange their parts (linguistic entities, recurrent themes and imagery, rhetorical devices etc.) in sets of metadata. These sets may overlap and intersect, and function as scholarly arguments. Encoding once does not preclude a division of the same text into a *different* set of parts, with another purpose or from another angle, or in response to an argument made through previous encoding. Each variant model contributes to a deeper understanding of the text at hand.

### 2.1 Collaboration

At last year's joint conference, Siemens et al. reported: "In terms of mark-up, respondents appear to be a bipolar group with half expecting to acquire text with no mark-up and half with rich XML." This no-middle-ground report seems to imply that once a user of electronic humanities resources is at all familiar with semantic encoding, rich markup becomes preferable to weak markup. Marking up large texts and corpora, common units of literary study, is a challenge both in terms of resources and required expertise. Such work calls for collaboration. VHL's toolset for presenting and working with primary texts (in development) provides several ways to contribute. A complex annotation engine and an opportunity to view the encoding behind any given segment of text are in place. In development is a tool for suggesting corrections to our encoding (intended to replace it), or submitting variations on it (intended to be viewed as alternate encodings of the same text).

While providing increased potential for new forms of communication, this toolset does not force scholars to change their preference for working mostly in solitude: Siemens et al.

do warn us that most of the humanists they surveyed "do not [currently] see the need for collaborating with other scholars."

### 2.2 Atomic Approach to Research

Freehand semantic encoding allows us to construct our own set of elements, based on prior knowledge of sources both primary and secondary, modifiable at will. Eventually this set must be regularized, perhaps later transcribed into a standardized form. But in the beginning stages such constraint would be detrimental, limiting the scope of analysis at the outset.

So we have begun to model without these constraints, permitting ourselves the spontaneity of a ludic approach. In doing this, we adapt Edward Hall's 1976 objective in examining culture — "look at the way things are actually put together" (13) — to text analysis. The encoding structure emerges bit by bit out of the primary source itself, which frees the researcher's critical eye to note interesting aspects of the text that might have eluded a pre-existing DTD.

Combining such an atomic approach to gathering research results with a web-based presentation implies a lot of flexibility for participating scholars: work may be done in smaller segments by individuals who live far apart. Here lies a strong driving force behind our work: similarly to already-successful electronic means of communication (email, weblogs, discussion lists), VHL allows small information packets to be published and discussed. Being unsuitable for the essay format because of their seemingly incomplete, fragmentary nature, these bits of information might not otherwise be expressed at all. Reducing the minimum size of a contribution to the knowledge base from an article to a paragraph or sentence, provided a review process is still employed, increases the net amount of useful knowledge available for discussion. We hope that it will actively encourage researchers to branch out and participate in more conversations, perhaps creating a distributed version of the editing process.

Stripping critical expression down to the essentials as expressed through semantic tagging will either highlight or address (or perhaps both) the difficulty Willard McCarty sees humanists having "with any intellectual culture whose cognitive activity is expressed in things rather than in words" (168). Thinking about a text by encoding criticism directly into it bridges the gap between the two, allowing multi-media corpora (literature, sculpture, films, drawings) to be encoded within the same electronic framework. Emphasis is shifted from the prose that delivers ideas (which consumes time and energy and often dilutes the argument) to precision in presenting the argument itself.

## 2.3 Humanists and Code

The encoding process requires considerable resources; writing up separate documentation is a significant enough amount of additional work that it isn't often done well. For humanist academics, it is absolutely necessary to be able to look at semantic encoding and more or less understand it.

Mahoney, and Henry Ford before him, are right: the masses are not mechanics. Yet, these days a certain amount of common knowledge about how machines work is necessary. Since semantically encoded electronic texts will only multiply as time goes on, humanists must know what code is and understand how it works. Knowledge of the underlying principles of encoding is not yet widespread, and VHL has taken it as a goal to present these principles in such a way that they become tacit knowledge for the humanist. We are making all of our XML code transparent -- any unit of text is viewable with all its code, and the XML itself is easily human-readable and well documented. Thus code remains an argument meant to be discussed and challenged as necessary, not an implicit, uncontested premise.

Learning to read code may require non-trivial effort, but carries with it an important additional benefit: it opens the door to a format of academic expression markedly different from the essay. Both have their uses, but mastering the basics of semantic encoding is a learnable and improvable skill that is likely to become tacit knowledge more readily than the much more difficult natural-language rhetorical approach of essay writing. Sentence structure, flow and finding the right word are essential to the humanist; but encoding makes it easier to learn and practice critical, in-depth analysis of texts.

## 3. Summary

Putting small bits of information together and hoping that a larger picture will emerge is arguably risky. There is no guarantee that the results will be interesting or useful. That said, this risk is inherent in all academic discussion, and recent experience indicates a movement (back?) toward tinkering with primary sources directly. Stephen Ramsay's call to go in "with a hunch borne of our collective musings" (171) encourages play, frightening though it may be to dedicate extremely scarce resources to the endeavor. This is where a playground like VHL shines. It is a tool for collaborating, community building and education that does not require a significant commitment of finances or time from its participants. In fact, for it to function, there need only be interest in the subject matter, and the willingness to record a single thought.

## Bibliography

*Decameron Web: A Growing Multimedia Archive of Materials Dedicated to Boccaccio's Masterpiece*. Accessed 2005-03-04. <<http://www.brown.edu/decameron/>>

Hall, Edward. *Beyond Culture*. Garden City, NY: Anchor Press, 1976.

Mahoney, Michael. "Keeping In Touch With the World." Commencement Address delivered at Brevard College. 16 May 1998. Accessed 2005-03-04. <<http://www.princeton.edu/~mike/brevard.html>>

McCarty, Willard. "As It Almost Was: Historiography of Recent Things." *Literary and Linguistic Computing* 19.2 (2004): 161-80.

*Pico Project*. Accessed 2005-03-04. <[http://www.brown.edu/Departments/Italian\\_Studies/pico/](http://www.brown.edu/Departments/Italian_Studies/pico/)>

Ramsay, Stephen. "Toward an Algorithmic Criticism." *Literary and Linguistic Computing* 18.2 (2003): 167-174.

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." Paper presented at ALLC/ACH 2004, Gothenburg, 2004.

*Virtual Humanities Lab*. Accessed 2005-03-04. <[http://www.brown.edu/Departments/Italian\\_Studies/vhl/](http://www.brown.edu/Departments/Italian_Studies/vhl/)>

# Index of Presenters

Adell, Joan-Elies.....	3
Akama, Hiroyuki.....	148
Ammon, Prof. Dr. phil. Ulrich.....	170
Anderson, Jean.....	157
Archer, Dawn.....	214
Argamon, Shlomo.....	4
Baayen, Harald.....	100
Baker, Paul.....	214
Balsamo, Anne.....	7
Barbera, Michele.....	8
Barrière, Caroline.....	10
Baskerville, Peter.....	69
Bauman, Syd.....	66
Beaubian, Rick.....	70
Bellavance, Claude.....	69
Best, Michael.....	13
Beynon, Meurig.....	135
Birnbaum, David J.....	55, 66
Blake, Jonathan.....	17
Blandford, Ann.....	250
Bonnett, John.....	131
Borràs, Laura.....	18, 26, 152
Bradley, John.....	20
Buchanan, George.....	250
Burghart, Alex.....	244
Butler, Diane.....	221
Butler, Terry.....	23, 24
Canadell, Roger.....	18, 26, 152
Canfield, Kip.....	28
Cantara, Linda.....	30
Carreras Riudavets, Francisco Javier.....	201
Catapano, Terence.....	174
Caws, Catherine.....	24
Cayless, Hugh.....	32
Chartrand, James.....	185
Chatterjee, Amal.....	35
Chesher, Chris.....	36
Choi, Yunseon.....	176
Clement, Tanya.....	38

Coburn, Aaron.....	40
Cole, Creagh.....	42
Cunningham, Richard.....	44
D'Ercole, Nicolò.....	8
Da Sylva, Lyne.....	45
Dahl, Susan.....	70
Daneker, Ingrid.....	47
Darroch, Gordon.....	69
Davis, Boyd.....	255
Deegan, Marilyn.....	214
Dekhtyar, Alex.....	82, 84, 102
Deshaye, Joel.....	220
Downie, J. Stephen.....	50, 53
du Cassé, William.....	206
Dubin, David.....	50, 55
Duke, David.....	44
Dunae, Patrick.....	131
Durand, David.....	58
Ernestus, Mirjam.....	100
Eustace, John.....	44
Eva, Buchi.....	198
Evans, Jennifer.....	120
Evans, Richard.....	161
Fegan, Michael.....	60
Flanders, Julia.....	66
French, Amanda.....	68
Friesen, Norm.....	24
Fuchs, Brian.....	218
Fuhr, Prof. Dr.-Ing. Norbert.....	170
Furuta, Richard.....	233
Gaffield, Chad.....	69
Galey, Alan.....	13, 191
Galloway, Patricia.....	156
Galway, Anna.....	44
Gartner, Richard.....	70
Gervás, Pablo.....	123
Gibson, Matthew.....	72
Gollan, Andrew.....	206
Gosciniak, André S.....	73
Grove-White, Elizabeth.....	164
Gueguen, Gretchen.....	208
Gurevich, Olga.....	87
Gurney, Lyman W.....	75
Gurney, Penelope J.....	75



---

Haas, Felicitas.....	77	Meijer, Piet.....	35
Hart-Davidson, Bill.....	60	Meister, Jan Christoph.....	123
Henderson, William.....	266	Merrilees, Brian.....	109
Hernández Figueroa, Zenón.....	204	Meschini, Federico.....	146
Hoover, David.....	79, 80	Metzing, Dieter.....	260
Hswe, Patricia.....	68	Michalak, Sarah.....	68
Hunyadi, Laszlo.....	214	Miyake, Maki.....	148
Huot, Wendy.....	13	Moll Soldevila, Isabel Clara.....	18, 26, 152
Iacob, Ionut Emil.....	82, 84, 102	Moore, Neil.....	90, 102
Isaksen, Leif.....	218	Musick, Judith.....	266
Janssen, Bill.....	87	Mylonas, Elli.....	154
Jaromczyk, Jerzy W.....	90, 102	Nakagawa, Masanori.....	148
Jessop, Martyn.....	91	Narayan, Vidya.....	156
Jockers, Matthew.....	215	Nelson, Jennifer.....	206
Johnson, Andrea.....	93	Nutter, Susan.....	68
Juola, Patrick.....	95, 97	Opas-Hanninen, Lisa Lena.....	157
Juuso, Ilkka.....	157	Palmer, Joy.....	60
Karttunen, Lauri.....	87	Patrik, Linda E.....	159
Kennedy, Oliver.....	212	Patterson, Erin.....	44
Keune, Karen.....	100	Peinado, Federico.....	123
Kiernan, Kevin.....	84, 102	Pekar, Viktor.....	161
Kumar, Amit.....	120, 208	Pérez Aguiar, José Rafael.....	201
Lancashire, Anne.....	109	Petter, Chris.....	164
Lancashire, Ian.....	115	Pierazzo, Elena.....	167
Langerth Zetterman, Monica.....	116, 119	Piez, Wendell.....	169
Lee, Jin Ha.....	50, 176	Pilz, Thomas.....	170
Leslie, Scott.....	24	Pitti, Daniel.....	174
Levine, Jennie A.....	120	Porter, Dorothy.....	102
Levitan, Shlomo.....	4	Posgate, Jessica.....	164
Lönneker, Birte.....	123	Pound, Chris.....	221
Losada García, Luis.....	204	Price, Kenneth.....	174
Losada García, Luis Javier.....	201	Radzikowska, Milena.....	191
Luther, Prof. Dr., Wolfram.....	170	Ramsay, Stephen.....	191
Lutz, John.....	131	Ramsay, Stephen J.....	53
Makoshi, Nobuyasu.....	148	Rehberger, Dean.....	60
Martin, Shawn.....	133	Renear, Allen.....	50
Mateas, Michael.....	123	Renear, Allen H.....	176
Mathes, Adam.....	50	Rentfrow, Daphnée.....	68
McCarty, Willard.....	135, 210	Rheault, Sylvain.....	180
McDonough, Jerome.....	70	Richards, Griff.....	24
McEnery, Tony.....	214	Roberts, Linda.....	164
McKeag, Michael.....	142	Roberts-Smith, Jennifer.....	109
Medina, Karen.....	50	Robey, David.....	182
Mei, Qiaozhu.....	268	Rockwell, Geoffrey.....	53, 182, 185, 210, 215

---

Rodríguez Rodríguez, Gustavo.....	204	Vechtomova, Olga.....	248
Romary, Laurent.....	66, 198	Vetch, Paul.....	20
Rose, Spencer.....	164	Walda, Hafed.....	244
Rudman, Joseph.....	187	Walsh, John.....	246
Ruecker, Stan.....	191	Walter, Katherine.....	174
Ruotolo, Christine.....	197	Wang, Ying.....	248
Russ, Steve.....	135	Wardrip-Fruin, Noah.....	58
Saddlemeyer, Ann.....	208	Warwick, Claire.....	47, 210, 250
Salmon-Alt, Susanne.....	198	Weller, Philip.....	253
Santana Suárez, Octavio.....	201, 204	Welling, George.....	254
Scaife, Ross.....	206	Westman, Stephen.....	255
Schreibman, Susan.....	120, 208, 210, 215	Willett, Perry.....	210
Schröder, Bernhard.....	77	Winget, Megan.....	257
Scifleet, Paul.....	42	Witt, Andreas.....	260
Seppänen, Tapio.....	157	Wittern, Christian.....	262
Shawver, Gary.....	212	Wolff, Mark.....	264
Shoichet, Jillian.....	164	Wood, Stephanie.....	266
Short, Harold.....	214	Xiang, Xin.....	176
Siemens, Ray.....	24, 210, 215	Yu, Bei.....	53, 268
Sinclair, Stéfan.....	185, 191	Zafrin, Vika.....	271
Slights, Jessica.....	13	Zhai, Chengxiang.....	268
Smith, Abby.....	210	Zimmerman, Matthew.....	66
Smith, Amy.....	218		
Smith, Jeff.....	220		
Smith, Martha Nell.....	210		
Smith, Steven Escar.....	233		
Spiro, Lisa.....	221		
St-Hilaire, Marc.....	69		
St-Jacques, Claude.....	10		
Steggle, Matthew.....	223		
Stoicheff, Peter.....	220		
Tabata, Tomoji.....	224		
Tcheng, David.....	53		
Terras, Melissa.....	227		
Thomas, Stephanie F.....	230		
Tingle, Brian.....	70		
Turner, James.....	45, 232		
Unsworth, John.....	53, 68, 182, 210		
Urbina, Eduardo.....	233		
van Dalen-Oskam, Karina.....	237, 240		
van Hardenberg, Peter.....	13		
van Hout, Roeland.....	100		
van Zundert, Joris.....	237, 240		
VandeCreek, Drew.....	243		

# Index of Topic Keywords

3-D reconstruction in humanities.....	131
Abraham Lincoln.....	243
acoustic reduction.....	100
agglomerative clustering.....	156
AIML.....	35
algorithms.....	264
análisis sintáctico.....	201
Anglo-Saxon.....	244
annotation.....	13, 20, 240, 253, 257
approaches to humanities research.....	221
archive digitisation.....	93
archiving.....	13, 156, 223
arts informatics.....	36
authorship attribution. 4, 75, 79, 80, 97, 187, 237	
autoguidage.....	180
automated textual analysis.....	255
automatic text processing.....	23
Biblical software.....	148
case-based storytelling.....	123
Cervantes.....	233
chat bot.....	35
civic discourse.....	243
classical music.....	20
CLIR Post-Doctoral Fellowship.....	68
collaboration.....	13, 174, 266, 271
collaborative research.....	102
collocation.....	212
comprehensive tagging.....	102
computational creativity.....	123
computational lexicography.....	10
computational linguistics.....	201
computational stylistics.....	268
computationally augmented reading.....	87
computer wordscoring content analysis.....	73
computer-assisted reading.....	264
computing.....	91
Computing Science.....	135
concurrent hierarchies.....	82
contemporary academy.....	221
content reuse.....	218
conversion.....	120
corpus.....	157
corpus analysis.....	10
correspondence analysis.....	224
criticism.....	18
CS & humanities.....	102
culture.....	7
curriculum definition.....	227
curriculum development.....	227
cybertextuality.....	115
data management framework.....	82
data mining.....	53
database.....	13, 244
databases.....	109
Delta.....	79, 80
desambiguación.....	201
descriptive markup.....	119
design considerations.....	55
development.....	44
diachronics.....	198
dictionaries.....	109
digital editing.....	240
digital editions.....	20
digital humanities.....	210
digital libraries.....	60, 146, 233
digital library.....	42, 70, 250
digital literature.....	18
digital preservation.....	30, 257
digitization.....	254
disambiguation.....	201
disciplinary integration of computing tools..	215
docencia.....	152
DTD.....	167
Dublin Core.....	50
e-journals.....	8, 223
e-learning.....	26
e-literature.....	3
e-publishing.....	223
e-text.....	13
EAD.....	47, 174
EAD (Encoded Archival Description).....	120
EAD encoding.....	156
editing tools.....	84, 102

educational settings.....	119	interdisciplinary education.....	102
educational technology.....	24	interface design.....	191
electronic editing.....	191	Internet Shakespeare Editions.....	13
electronic edition.....	13, 72	Java.....	120
emerging digitization efforts.....	3	keyword extraction.....	214
encoding.....	167	knowledge representation.....	262
encoding & electronic editions.....	230	language acquisition.....	10
encoding standards.....	198	language cognition.....	115
evaluating digital texts.....	38	language resources.....	161
eXist XML database.....	164	Latin.....	75
experience.....	135	lexicography.....	198
feature selection.....	268	lexicon.....	28, 266
Federalist Papers.....	187	libraries.....	197
fiscal conservativeness of political actors....	73	Library Science.....	68
fixed-phrase.....	212	libretto.....	167
français écrit.....	180	linguistically-directed reading.....	87
FRBR.....	50	lingüística.....	204
function words.....	4	lingüística computacional.....	201
games.....	58	linguistics.....	142
geometry.....	142	literary criticism.....	215
German Nietzsche reception.....	170	literary studies.....	18
GIS in humanities.....	131	literatura.....	26, 152
GODDAG.....	82	literature electronic archive.....	220
guardian-spender budgetary framework.....	73	Logic Programming.....	55
heraldry.....	142	manuscripts.....	90, 102, 266
hierarchies.....	13	markup.....	116, 262
higher education.....	260	markup languages & programming.....	260
hipertexto.....	26	metadata.....	30, 70
historical indexes.....	109	metadata for information management.....	232
history.....	227, 254	methodology.....	237
humanities.....	91, 135	METS.....	70
humanities computing.....	182	minorité.....	180
humanities teaching.....	24, 221	minority languages.....	157
humanities users.....	250	modeling.....	40, 176, 271
hypermedia.....	44	morfología computacional.....	204
hypertext.....	58, 154	multicultural.....	159
image retrieval.....	45	multilevel modeling.....	100
image-based electronic editing.....	84, 102	multilingual.....	206
image-based electronic editions.....	90	multilingual metadata.....	45
index generation.....	95	multimedia.....	13
indexing multimedia objects.....	45	multimedia modelling.....	50
information retrieval.....	40, 214	multiple-texts.....	230
interactive story.....	123	multivariate analysis.....	116
interdisciplinary data integration.....	69	narrative corpora.....	255

narrative intelligence.....	123	Semantic Web.....	8
narrative ontology.....	123	semantics.....	212
narratology.....	123	Shakespeare.....	13, 191, 253
national support.....	182	sociolinguistic & register variation.....	100
natural language processing.....	148, 161	standards.....	174
Navajo.....	28	statistical stylistics.....	79, 80
neolatin.....	206	story generators.....	123
new media.....	24, 58	story-based games.....	123
new media art.....	257	stylistic variation.....	224
non-standard orthography.....	170	stylometry.....	224
online delivery of primary text.....	23	SVG.....	32, 169
online tools.....	157	syntactic analysis.....	201
Open Access.....	8	system architecture.....	123
open source tools.....	255	teaching.....	91
oral history.....	30	teaching humanities.....	131
Oulipo.....	264	technological imagination.....	7
pedagogical applications of new media.....	3	technology design.....	7
pedagogy.....	36, 44	TEI.....	32, 38, 42, 66, 146, 159, 169, 208, 246
philosophy.....	42, 77	term co-occurrence.....	248
plugin architecture.....	90, 102	term proximity.....	248
poetry.....	17, 55	text analysis.....	115, 185, 191, 215
portals.....	185	text analytics.....	95, 97
preservation.....	13	text and markup.....	72
project management.....	102	text categorization.....	268
proppian functions.....	123	text classification.....	4
prosopography.....	116, 244	text encoding.....	23, 66, 176, 208
public humanities.....	243	text mining.....	53
publishing.....	210	text technology.....	77, 260
query expansion.....	248	text-processing software.....	95, 97
rdf syndication.....	218	textbase.....	28
recuperación morfológica.....	204	texts.....	206
reflexivity.....	36	textual analysis.....	75, 214
representation of non-text objects.....	257	textual editing.....	38
rhetoric.....	154	textual iconography.....	233
Robert Graves diary.....	164	textual variants.....	230
role of new media.....	221	theory of text.....	240
rule based search.....	170	Tibetan.....	159
scansion.....	17	tools.....	185
scholarly communities.....	197	tools development.....	53
scholarly editing.....	208	Topic Maps.....	246
scholarly information resources.....	68	topic maps.....	262
scholarly publishing.....	72	TopicMap.....	146
second/foreign language acquisition.....	35	trade.....	254
secondary repositories.....	60	usability.....	250

use of electronic resources.....	133
user interfaces.....	93
user-performance data.....	60
variorum.....	191
Virtual Humanities Lab.....	271
virtualidad.....	152
visual aids.....	232
visual collections.....	218
visualization.....	32, 40, 169, 191
vocabulary analysis.....	237
Web.....	17
Web interfaces.....	164
web publication.....	154
Web technologies.....	161
XML.....	13, 66, 120, 246
XML database eXist.....	164
XML DTD.....	47
XML tools.....	119
XSL schema.....	47