ALLC/ACH 2003

The Association for Computers and the Humanities
The Association for Literary and Linguistic Computing
The 2003 Joint International Conference

Conference Abstracts

Posters and Demonstrations

The University of Georgia
Athens, Georgia
May 29–June 2, 2003

ALLC/ACH 2003
Conference Abstracts

The Association for Computers and the Humanities
The Association for Literary and Linguistic Computing
The 2003 Joint International Conference

The University of Georgia

Plenary Speakers

John Maeda, Massachusetts Institute of Technology
William Potter, University of Georgia
Marie-Laure Ryan, Independent Scholar

Program Committee

Elisabeth Burr, Gerhard-Mercator-Universität Duisburg (Germany)
Lorna Hughes (Chair), New York University
Laszlo Hunyadi, University of Debrecen (Hungary)
Martha Nell Smith, University of Maryland
Natasha Smith, University of North Carolina at Chapel Hill
Ray Siemens, Malaspina University College (Canada)
Michael Sperberg-McQueen, World Wide Web Consortium
Simon Horobin, University of Glasgow (Scotland)

Editors

Eric Rochester
William A. Kretzschmar, Jr.

Created using XSLT from XML source documents.

TABLE OF CONTENTS

Posters

Conference Program

WEDNESDAY, MAY 28TH

WORKSHOPS

9:30 AM – 6:00 PM **Introduction to XML and the TEI** Computer Lab A

QUANTITATIVE LINGUISTICS CONFERENCE

1:15 – 1:30 PM **Welcome**

1:30 – 2:00 **Petra Steiner**
International Computer Science Institute, Berkeley, USA
"Morphosyntactic diversification of German words."

2:00 – 2:30 **Shu-Kai Xie**
University of Tuebingen, Germany
"Revisiting the word length problems of Chinese."

2:30 – 3:00 **Break**

3:00 – 3:30 **Relja Vulanovic**
Kent State University, USA
"Fitting periphrastic do in affirmative declaratives."

3:30 – 4:00 **Kris Heylen, Dirk Speelman**
University of Leuven, Belgium
"The use of stratified statistical analysis in quantitative corpus research: The case of word order variation in the German middle field."

4:00 – 4:30 **Break**

4:30 – 5:00 **Liang Chen, John Oller**
University of Louisiana, Lafayette, USA
"Episodic organization is essential for valid and reliable language assessment tools."

5:00 – 5:30 **Hanjung Lee**
"Quantitative variation in style shifting: A stochastic OT analysis of case ellipsis in Korean."

5:30 – 6:00 **Harald Baayen**
University of Nijmegen, The Netherlands
"Word Frequency and its formal and semantic correlates: a multivariate approach to the interpretation of word frequency"

THURSDAY, MAY 29TH

WORKSHOPS

9:30 AM – 6:00 PM **Introduction to XSLT** Computer Lab A

QUANTITATIVE LINGUISTICS CONFERENCE (QUALICO)

9:00 – 9:30 AM **Anatoliy Polikarpov**
Moscow University, Russia
Evolutionary model as a basis for revealing quantitative regularities of word formation process.

9:30 – 10:00 **M. Hubey**
An interlanguage metric for historical linguistics for dependent comparanda.

10:00 – 10:30	**John Nerbonne, Peter Kleiweg** University of Groningen, The Netherlands A dialectological yardstick.
10:30 – 11:00	**Break**
11:00 – 11:30	**Sheila Embleton, Dorin Uritescu, Eric Wheeler** York University, Canada Romanian Online Dialect Atlas.
11:30 – 12:00	**Marjatta Palander, Lisa Lena Opas-Hanninen, Fiona Tweedie** University of Joensuu, University of Oulu, University of Edinburgh
12:00 – 12:30	**Patrick Juola** Duquesne University, Pittsburgh, USA Becoming Jack London.
12:30 – 12:45	Farewell

ALLC/ACH 2003 CONFERENCE

9:00 AM – 5:00 PM	**ALLC Executive Meeting**	Room T/U
9:00 AM – 5:00 PM	**ACH Executive Meeting**	Room V/W
2:00 PM – 5:00 PM	*Excursion (Walking tour)*	Gather in the Lower Lobby
6:00 PM – 7:00 PM	**Plenary:** "To Digital or Not To Digital" **John Maeda** Massachusetts Institute of Technology Muriel Cooper Chair Professor of Media Arts and Sciences; Associate Professor of Design and Computation; Director of the Aesthetics & Computation Group (ACG)	Masters Hall
7:00 PM – 8:00 PM	**Reception**	Georgia Center
8:00 PM – midnight	Shuttles to Downtown	

FRIDAY, MAY 30TH

PARALLEL SESSION I: 8:30–10:00 AM

Room K/L

1A: Session: 3 papers
Kevin Kiernan *session* The ARCHway Project: Architecture for Research in Computing for Humanities through Research, Teaching, and Learning
Participants:
Dorothy Carr Porter, Introduction
Jerzy Jaromczyk and Sandeep Bodapati, "An Architecture Promoting Collaborative Research, Teaching, and Learning";
Alexander Dekhtyar and Ionut Emil Iacob, "Management of Data for Building Electronic Editions of Historic Manuscripts";
Kevin Kiernan and Kenneth Hawley, "An Image-Based Electronic Edition of Alfred the Great's Old English Version of Boethius's *Consolation of Philosophy*"

Room Q

1B: Authorship
Chair: **Ray Siemens** Malaspina University-College
Patrick Juola and Harald Baayen *paper* "A Controlled-Corpus Experiment in Authorship Identification by Cross-Entropy"
Joseph Rudman *paper* "On Determining a Valid Text for Non-Traditional Authorship Attribution Studies: Editing, Unediting, and De-Editing"
David Hoover *paper* "Collocations, Authorship Attribution, and Authorial Style"

Room R

1C: Session: Panel
Nelson Hilton *session* EMMA: Re-forming Composition with XML
Participants:
Nelson Hilton, "Imagining an English Markup and Management Application";
Ron Balthazor, "EMMA's Genesis: Building the Client";
Alexis Hart, "EMMA's Development: From Software to Students";
Robert Cummings and Angela Mitchell, "EMMA at Work in the Writing Classroom";
Christy Desmet and Angela Mitchell, "Observing EMMA: The First Year"

PARALLEL SESSION II: 10:30 AM – NOON

Room K/L

2A: New representations.
Johanna Drucker and Bethany Nowviskie *paper* "Temporal Modeling"
David Saltz *paper* "Virtual Vaudeville: A Live Performance Simulation System"
Jan Christoph Meister *paper* "Tagging Time in Prolog"

Room Q

ALLC Open Session.
Chair: **Harold Short**

Room R

2C: Session, 3 papers.
Susan Brown *session* "Deep Encoding"
Participants:
Susan Brown, "Delivering the Depths: Representing the Orlando Project's Interpretive Markup";
Willard McCarty, "Modeling the Depths of a Literary Encoding, with an Example from Ovid";
Wendell Piez, "Whither Deep Markup?"

Noon – 1:30 PM	**Lunch**
Noon – 1:30 PM	**ALLC General Meeting** Room K/L (lunch provided for the first 30 participants)

PARALLEL SESSION III: 1:30–3:00 PM

Room K/L

3A: Digital Libraries and Museums.
Chair: **Lorna Hughes** (New York University)
María Blume, Elaine Westbrooks, Cliff Crawford, James Gair, Tina Ogden, and Barbara Lust *paper* "Creating a Virtual Center as an International Web-Based Interactive Infrastructure for Research and Teaching in the Language Sciences: A New Research and Library Collaboration."
Hsin-liang Chen *paper* "Chinese Collections in Museums on the Web: Current Status, Problems, and Future."
Marilyn Deegan and Harold Short *paper* "New Technologies, New Strategies for Integrating Information and Knowledge: Forced Migration Online"

Room Q

3B: Interoperability and preservation
Chair: **Martha Nell Smith** (University of Maryland)
Elizabeth J. Shaw *paper* "Data or Document? Migration of Descriptive Metadata for Medieval and Renaissance Manuscripts Between Data-Centric and Document-Centric Models: A Case Study"
Linda Cantara *paper* "The Charles W. Cushman Collection: Enhancing Visual Resource Discovery Through Descriptive Metadata Based on Subjective Image Analysis"
Megan Winget *paper* "Preservation of the New Media Arts

Room R

3C: Session: Panel
Peter Scharf *session* Linguistic Issues in the Text-Encoding of Sanskrit
Participants:
Peter Scharf, "Linguistic Issues in the Entry, Character-Encoding, Processing, and Rendering of Sanskrit";
Malcolm Hyman, "Applications of a Sound-Based Encoding Scheme for Sanskrit";
Venu Govindraja, "Truthing Scanned Sanskrit Documents";
Ralph Bunker, "User-Customizable OpenType Fonts for Devanagari"

PARALLEL SESSION IV: 3:30–5:00 PM

Room K/L

4A: Representations on the Screen.
Chair: **Willard McCarthy** (King's College London)
Michele White *paper* "The Screen or the Window: A Critical Proposal for Reading Computer Representations"
Eunice Johnston *paper* "Visual or Verbal: Two Approaches to Creating an Immersive Virtual Environment"

Room Q

4B: Session, 3 papers.
John Paolillo *session*
"Computational Approaches to Linguistic Variation"
Participants:
John Nerbonne, "Vocabulary and Pronunciation in Linguistic Variation";
John Paolillo, "Zooming in on Longitudinal Variation";
William Kretzschmar and Jean-Claude Thill, "Self-Organizing Maps as an Approach to GIS Analysis of Linguistic Data"

Room R

4C: Allied organization panel: ELO Session
"PAD: Preservation, Archiving, and Dissemination of Electronic Literature"
Participants:
David Durand, Marjorie Coverley Luesebrink, Nick Montfort, Jessica Pressman, Scott Rettberg

---

5:00 – 6:00 PM	**ELO Special Session**: "Writers Reading Electronic Literature: A Creative Performance"	Room K/L
6:00 – 7:00 PM	**Plenary:** "A New Library Model in the Digital Age: the UGA Student Learning Center" **William Potter** The University of Georgia University Librarian	Memorial Hall
7:00 – 8:00 PM	**Reception (with Mayor Heidi Davison)**	Memorial Hall
5:30 PM – midnight	Shuttles to Memorial Hall, Downtown	

SATURDAY, MAY 31ST

POSTER SESSION: 8:30 AM – 10:00 AM

- **Rafael Alvarado and Sarah-Jane Murray,** "Figura: A Tool for the Collaborative Editing of Non-Nesting Content"
- **Syd Bauman,** "On the Content Model for <respStmt>: Newer Is Not Necessarily Better"
- **Hanno Biber,** "The Austrian Academy Corpus, an Extensive Corpus of German Literature and Language—The AAC Literary Journals Subcorpora"
- **Laura Borras,** "Teaching Literature Through the Net: An Answer to the Caos or the Construction of the Self"
- **Hugh Cayless**, "Digital Epigraphy"
- **Patricia Clements, Renée Elio, Sharon Balazs, Susan Brown, and Isobel Grundy,** "Orlando on the Web: From Development System to Web-based Delivery of a Content-Encoded Textbase"
- **Julia Flanders,** "ACH Mentoring"
- **Christiana Fordyce,** "Electronic Esposizione"
- **Alice Hickock, Chuck Spornick, Julia Leon,** "The Development of the Poetry Portal at the Beck Center, Woodruff Library, Emory University"
- **John Paolillo,** "Solving the Legacy-Encoding Debacle with On-line Transliteration"
- **Geoffrey Rockwell, Lian Yan, and Stéfan Sinclair,** "TAPoR Tools: Portal Text Analysis Tools and Other Primitives"
- **Yan Tian,** "Web Prompts the Increase of Chinese Non-English Majors' Speaking, Writing and Translating Abilities"
- **Helen Tibbo,** "Historians Primary Search Materials"

10:00 – 11:00 AM	**Plenary:** "Metaleptic Machines" **Marie-Laure Ryan** Independent Scholar	Masters Hall
11:00 AM – 7:00 PM	*Excursion (Plantation / Madison)*	Gather in the Lower Lobby
6:00 PM – midnight	Shuttles to Downtown	

SUNDAY, JUNE 1ST

PARALLEL SESSION V: 8:30–10:00 AM

Room K/L

5A: Allied organization panel: RSA Session.
William Bowen *session* "Present and Future Directions in Developing Online Resources for Renaissance Studies"
Participants:
William Bowen, "Iter: Building an Effective Knowledge Base";
Raymond Siemens, "Algorithmic Approaches to an Electronic Scholarly Edition of Early Modern Materials";
Stephanie Thomas, Chris Roast, and Innes Ritchie "The Exploration and Development of Tools for Active Reading and Electronic Texts"

Room Q

5B: Linguistics
Chair: **Simon Horobin** (University of Glasgow)
Mark Arehart *paper* "Identifying Multiword Tokens Using POS Tagging and Bigram Statistics"
Hermann Moisl and Val Jones *paper* "Cluster Analysis of the Newcastle Electronic Corpus of Tyneside English: A Comparison of Methods"
Greg Lessard *paper* "Computational Generation of Limericks"

Room R

5C: Session: 3 papers.
Bethany Nowviskie *session* "Constraint, Practice, and Interpretation"
Participants:
Bethany Nowviskie, "Llullian Method and Interpretation in Humanities Computing";
Andrea Laue, "Rules for Reading";
Stephen Ramsay and Geoffrey Rockwell, "Programming as Writing as Programming"

PARALLEL SESSION VI: 10:30 AM – NOON

Room K/L

6A: Session.
Joanne Kaczmarek (University of Illinois) *session* "Applications of the Open Archives Initiative"
Participants:
Martin Halbert, Joanne Kaczmarek, David Seaman, Stephen Schwartz

Room Q

6B: Session, 3 papers.
Susan Hockey *session* "Integrating TEI and EAD to Create Usable and Re-usable Archival Resources"
Participants:
Susan Hockey, Elizabeth Hallam-Smith, Anna Sexton, Chris Turner

Room R

6C: ACH Open Session.
John Unsworth *session*

Noon – 1:30 PM	**Lunch**
Noon – 1:30 PM	**ACH General Meeting** Room K/L (lunch provided for the first 30 participants)

PARALLEL SESSION VII: 1:30–3:00 PM

Room K/L

7A: Coding/Editing.
Chair: **Julia Flanders** (Brown University)
Barbara Bordalejo *paper* "Textual Critical Encoding"
Chris Tiffin, Graham Barwell, Phill Berrie, and Paul Eggert *paper* "Annotation and Electronic Scholarly Editions"
Paul Caton *paper* "Theory in Text Encoding"

Room Q

7B: Computing and language.
Chair: **László Hunyadi** (University of Debrecen)
Jan Rybicki *paper* "Burrowing into Translation: A Case Study"
Joachim Gerich and Roland Lehner *paper* "A Computer-Based Questionnaire for Hearing Impaired People"
Talan Memmott *paper* "Beyond Taxonomy: Digital Poetics and the Problem of Reading"

Room R

7C: Session, 3 papers.
Steve Guynup *session* "Beyond the Archive: Immersive Textuality for William Blake's Poetry"
Participants:
Steve Guynup, Web3D Immersive Illustration of Blake's "Crystal Cabinet";
Marcel O'Gorman, "The Fourfolds of William Blake and Martin Heidegger: Minds, Bodies, Technologies";
Nelson Hilton, "Golgonooza Songs, or, Blake in a Flash";
Ron Broglio, MOO Demonstration

PARALLEL SESSION VIII: 3:30–5:00 PM

Room K/L

8A: Session, 3 papers.
Chair: **Nancy Kushigian** (UC Davis) *session* "Research Library Collection Descriptive Frameworks"
Guenter Waibel *paper* "Museums in the Mix: Collections Across Communities"
Jared Campbell *paper* "Collection Descriptive Frameworks"
Nancy Kushigian *paper* "Developing a Descriptive Framework for Legacy Collection Descriptions"

Room Q

8B: Encoding.
Chair: **John Unsworth** (University of Virginia)
Michael Sperberg-McQueen *paper* "XML Schema 1.0: A Language for Document Grammars"
Allen Renear *paper* "Text Markup—Data Structure vs. Data Model"
Peter Robinson *paper* "Anastasia: A New XML Publication System"

Room R

8C: Session, 3 papers.
Matthew Gibson *session* "New Ways in Using and Creating Lexicographical Resources"
Participants:
Matthew Gibson and Ute Recker-Hamm, "Middle High German Interlinked: A Comprehensive Digital Text Archive";
Frank Queens and Ute Recker-Hamm, "Tools for Lexicography, Retrieval, Middle High German";
Thomas Schares, "Electronic Dictionaries and Metalexicography: The Digital Version of the Deutsche Wörterbuch by Jacob and Wilhelm Grimm as a Basis for Metalexicographical Research";

6:00 – 9:00 PM **Conference Banquet** Botanical Garden
Buses will begin departing from the Georgia Center at 5:45
Entertainment provided by "Truckfire"

MONDAY, JUNE 2ND

PARALLEL SESSION IX: 8:30–10:00 AM

Room K/L

9A: Developing Projects.
Chair: **Matthew Zimmerman** (New York University)
Lara Vetter and Jarom McDonald *paper* "Confronting the Challenges in Collaborative Editing Projects: *The Dickinson Electronic Archives* File Management System"
John Bradley and Harold Short *paper* "Texts into Databases: The Evolving Field of New-style Prosopography"
Ross Scaife and Raphael Finkel *paper* "The Suda On Line: Applying Computer Technology to Ancient and Byzantine Studies"

Room Q

9B: Allied organization Session: COSH session
Ray Siemens, COSH Panel *panel* "Great Expectations, Expectant Implementations—or, What We Expect of Our Electronic Resources and How We Meet Those Expectations"
Participants:
Ray Siemens, Geoffrey Rockwell, Patricia Clements, Andrew Mactavish, Michael Best

Room R

9C: Session, 3 papers.
Wendell Piez *session* "Ambiguity, Technology, and Scholarly Communication"
Participants:
Wendell Piez, "Scholarly Transgressions";
Julia Flanders, "Ambiguity and Text Encoding";
John Lavagnino "Ambiguity, Language, and the Scholarly Economy"

PARALLEL SESSION X: 10:30 AM – NOON

Room K/L

10A: Discovering texts.
Chair: **Marilyn Deegan** (University of Oxford)
Mark Olsen *paper* "Écriture féminine: Searching for an Indefinable Practice?"
Sabine Harwardt and Stefan Büdenbender *paper* "Chasing DTDs. The Digital Edition of the 'Repertorium Biblicum Medii Aevi'"

Room Q

10B: Corpus-based research(?).
Chair: **Elisabeth Burr** (University of Bremen and University of Duisburg)
Clayton Darwin, William Kretzschmar, and Donald Rubin *paper* "The Tobacco Documents Corpus: Archiving the Industry"
Hermann Moisl *paper* "Linguistic Corpus Construction and Analysis before and after the IT Revolution: the Newcastle Electronic Corpus of Tyneside English in the 1960s and Now"
Gary Simons *paper* "Developing Markup Metaschemas to Support Interoperation Among Resources"

Room R

10C: Panel Session.
Stefan Sinclair *session* "Peer Review of Humanities Computing Software"
Participants:
Stéfan Sinclair, John Bradley, Stephen Ramsay, Geoffrey Rockwell, Ray Siemens

---

Noon – 1:00 PM **Closing General Session**

PAPERS

The ARCHway Project: Architecture for Research in Computing for Humanities through Research, Teaching, and Learning

JERZY W.I. JAROMCZYK
Computer Science, University of Kentucky
jurek@cs.uky.edu

J. ADAM TURNER
IBM, CS University of Kentucky
aturner@qx.net

ALEXANDER DEKHTYAR
Computer Science, UK
dekhtyar@cs.uky.edu

IONUT EMIL IACOB
Computer Science, UK
ionut@ms.uky.edu

KENNETH HAWLEY
English, UK
kchawl0@uky.edu

KEVIN KIERNAN
English, UK
kiernan@uky.edu

PART I
SESSION STATEMENT

Kevin Kiernan, Research in Computing for Humanities, Department of English, University of Kentucky

With the help of the NSF's Information Technology Research program, an unusual alliance under the ARCHway Project is developing an Edition Production Technology (EPT), a technological infrastructure for collaborative research, teaching, and learning between computer scientists and specialists in Old English. Our goal is to identify and solve problems of mutual importance in building image-based electronic editions of significant cultural materials. EPT will allow us to implement and integrate both new and already available software applications, to construct a digital library of previously unedited Old English manuscripts as a testbed for our solutions, and to distribute the digital library to the public.

To establish a durable infrastructure, we will design formal methodologies for collaborative teaching and research, based on our practical goals. We will maintain an open-standards architecture with modular, extensible, interoperable components to coordinate research and development of novel methods, tools, and associated technologies. The EPT will guide the definition and coordination of well-encapsulated collaborative student projects using our testbed from semester to semester, student to student, year to year. The PIs will lead teams of students in both disciplines in focussed research projects related to documenting, encoding, editing, storing, accessing, searching, and disseminating digital libraries of image-based electronic editions.

ARCHway will in the process produce a system for building digital libraries of image-based scholarly editions for the humanities. This system will be used to produce electronic editions of a number of previously

unedited or inadequately edited Old English manuscripts. The architecture should enable both the creation of image-based electronic editions by editors in other fields and the use of those editions by research scholars, students, and the general public. The research results of this project will lay the groundwork for sophisticated technical tools to interpret, assemble, disseminate, and maintain image-based scholarly editions on a continuing basis by humanities scholars with limited or no access to programming resources.

The papers comprising this session assemble three of these collaborative teams of teachers and students to introduce some of the tools and technologies currently under development as we build this interdisciplinary teaching and research infrastructure. The first paper, "An Architecture promoting Collaborative Research, Teaching, and Learning," by Jerzy W. Jaromczyk, associate professor of Computer Science, and his masters student, J. Adam Turner, describes and demonstrates formal methodologies for collaborative teaching and research and the development of tools, based on our practical goals. The second paper, "Management of Data for Building Electronic Editions of Historic Manuscripts," by Alexander Dekhtyar, assistant professor of Computer Science, and his doctoral student, Ionut Emil Iacob, presents new ways for maintaining the integrity of highly complex, layered, XML markup, combined with image integration and for allowing a research team to work simultaneously on the same texts without impeding or destroying each other's work. The third paper, "An image-based electronic edition of Alfred the Great's Old English version of Boethius's Consolation of Philosophy," by Kevin Kiernan, professor of English, and his doctoral student, Kenneth Hawley, presents some of the practical applications of our collaboration.

PART II
AN ARCHITECTURE PROMOTING COLLABORATIVE RESEARCH, TEACHING, AND LEARNING

Jerzy W. Jaromczyk, Department of Computer Science, University of Kentucky, J. Adam Turner, IBM, and Department of Computer Science, University of Kentucky

The organization of a complex system like the Edition Production Technology (EPT) presents a set of serious design problems. Its workbench must offer comprehensive services, which the editor and co-researchers need to prepare and maintain DTDs or X-Schemas, encode texts with searchable descriptive and editorial markup in XML, integrate images, and design interfaces. It must also prepare powerful interactive displays for people to make full use of the completed image-based edition. Unlike most software projects, moreover, whose primary objective is to produce a functional and robust application, ARCHway's goals are more complicated, requiring an infrastructure that integrates Research, Teaching and Learning between the very different disciplines of computer science and the humanities. To meet this principal purpose, we must carefully plan the underlying architecture of EPT, which will guide the development process and map the functionality into the overall system. Our architectural model promotes the development of modular, extensible, interoperable components in a collaborative research and teaching environment that actively involves researchers and students in both disciplines.

RESEARCH

EPT is an innovative system whose strengths hinge on a number of ideas that require novel solutions and utilization of emerging technologies. Among them are questions of effective, convenient, and complete mapping of the phenomena of the editorial domain into a software system. They lead to research problems of image archiving, image access and manipulation, methods for automated linking of images with the text, storage, access, secure transactions, data persistence and consistency (algorithms), and virtual remorphing. The intellectual integrity of the project will be achieved by continuous exchange of "what" and "how" knowledge and domain specific-phenomena between humanities scholars and computer and technology specialists. The workbench that we create will mirror and augment the processes, methodologies, and software tools the editor uses to develop electronic editions from original manuscripts.

Reminiscent of the currently prevailing Windows manager interface introduced by Doug Englebart and his Stanford team in the late sixties, the workbench will attempt to recreate a virtual environment for studying codicology. As Frederick Brooks, an authority in software engineering puts it, "The Windows, Icons, Menus and Pointing interface is a superb example of a user interface that has conceptual integrity, achieved by the adoption of a familiar mental model, the desktop metaphor, and its careful consistent extension to exploit a computer graphics implementation." However, the editors' workbench will become more than a metaphor as we aim to build a system that on the functional, hardware, and almost haptic (i.e. hands-on, touchable) levels reproduces, augments, and organizes work with manuscript images into image-based editions.

TEACHING AND LEARNING

As in any endeavor of developing a complex system the knowledge, creativity, and skills of the people

involved are critical to its ultimate success. Students at various levels of their studies will be participating in the project, and thus will bring with them different levels of skill. To accommodate and integrate students and researchers from interdisciplinary environments, and to ensure that the architecture is flexible and open for creative solutions and responds well to potential errors, unavoidable in any learning process, EPT will use the end-to-end skeleton system advocated by Harlan Mills. In this approach the system grows from a basic loop with stubs (subroutine calls) by adding functionality to gradually increasing components. The original skeleton is developed by the most advanced researchers. This approach will help maintain the correctness of the system throughout the entire process and at the same time will separate and divide design and implementation tasks among different groups of students. The encapsulation and implementation of Object Oriented Programming allows developers to define self-contained modules such as parsers, search facilities, and image manipulation tools, while limiting interaction between separate components to interfaces, thus reducing the probability of errors. Ultimately all these modules are plugged into the skeletal system and become available through a portal to serve the user.

We will illustrate this approach with an implementation of the customizable graphical search tool (CGST-XML) for XML documents that is a part of the development for ARCHway. CGST-XML is an application built around the Model-View-Controller design pattern, wherein a user is provided with a number of options (buttons) to control the operations of the system. Specific to CGST-XML is that the operation of the system is organized into three steps: preprocessing, action, and post processing. Each step can be executed by a number of modularized components that are available to the user. Through dynamically loaded configuration files, the user not only can decide what particular programs, tools, and implementations are used, but also can change the Graphical User Interface to one that best suits the current stage of the editorial process. Additional implementations of separately developed modules can be added to expand the system, multiplying the functionality of the system. The system is reminiscent of a plug-in oriented Web browser, but instead of selecting specific applications based on a suffix, users have the ability to create their own suites of operations. As a specific example, we will show CGST-XML in action with several possible configurations to demonstrate the extensibility of the project.

REFERENCES

F. P. Brooks. *The Mythical Man-Month: Essays on Software Engineering*, 1995.

M. S. Brown, W. B. Seales, K. Kiernan, and J. Griffioen. “3D Acquisition and Restoration of Medieval Manuscripts.” *Communications of the ACM: Special Issue on Digital Libraries*, May 2001.

E. Gamma, R. Helm, Ralph Johnson and John Vlissides, *Design Patterns: Elements of Reusable Object-Oriented Software*, Addison-Wesley, 1996.

Handbook of Discrete and Computational Geometry, edited by J. E. Goodman and J. O’Rourke, CRC Press, 1997.

A. Hunt and D. Thomas, *The Pragmatic Programmer*, Addison-Wesley, 1999.

PART III

MANAGEMENT OF DATA FOR BUILDING ELECTRONIC EDITIONS OF HISTORIC MANUSCRIPTS

Alex Dekhtyar, Ionut Emil Iacob, Department of Computer Science, University of Kentucky,

The process of preparing electronic editions [SGK00] is long, time-consuming, and arduous for an editor or editors. Some of the work cannot be automated—editors, for example, must scrutinize every single letter of a document numerous times to come to fully informed decisions concerning script, meaning, and spelling [Hay01]. The success of the ARCHway project, designed to alleviate unnecessary tedium of the editorial process lies, in major part, in correct manipulation of the data that forms the electronic edition.

At the outset of the process of preparing an electronic edition, an editor takes raw data in the form of digital images of the manuscript folios and precise transcriptions of them and proceeds to encode the transcriptions with descriptive markup by:

- identifying the folios and folio lines, and the prose and/or verse lines;
- associating folio images with the text that they contain;
- creating a full glossary entry for every word in the text;
- recording a wide range of features of the manuscript, such as (among many other things) its script, its legibility, any damage to the manuscript, any technological means used to restore damaged readings, as well as editorial emendations and conjectural restorations.

This encoding must be stored in a way that

1. allows efficient retrieval of information (e.g., “Display all characters written by the second scribe that are visible only under ultra-violet light”);

2. ensures efficient and convenient manipulation of the data;
3. provides support for editorial work by more than one person at the same time.

The challenges the editorial process presents can be broken into three broad categories:

- management of XML markup produced during the editorial process [BSK01],
- maintenance of associations between XML markup and manuscript images [YS01]
- support for a multi-user environment

MANAGEMENT OF XML MARKUP FOR ELECTRONIC EDITIONS

Because the editor and research team will record a highly diverse and extensive set of manuscript features, the markup of the edition text is bound to be complex. Different features subject to description may not conveniently follow valid or well-formed XML nesting patterns in the actual manuscript, making the XML document of the edition quite eccentric with potentially clashing hierarchies. There are two approaches to overcoming this problem. One is to maintain an extremely complicated DTD (or XML Schema) [BPS00, BM01] for all markup and use specific, at times cumbersome, markup conventions to overcome clashing hierarchies in the tagging. While this approach has the benefit of storing all annotation in the same XML document, the ad hoc solutions adopted to keep the markup well-formed reflect negatively on the clarity of the XML and may adversely affect subsequent searches and retrieval of information. The tagging software for the editor may also become too tied to a particular XML Schema to be of generic use.

Another approach is to separate markup for different features into different DTDs (or XML Schemas) and maintain parallel markup of the edition text automatically. This approach results in clear, concise, easily maintained DTDs or X-Schemas. It also is an excellent stepping stone for supporting simultaneous work of many editors on one edition: for example, paleography markup created by one editor and damage markup created by another at the same time can be stored separately and thus will create no conflicts in data. This approach shifts most of the data maintenance burden from the shoulders of the editorial team to the software that supports the process. It is currently being implemented in ARCHway.

BUILDING ELECTRONIC EDITIONS AROUND IMAGES

The images of manuscript folios [Bse01] are, by far, the most important component of the Electronic Editions as they provide the opportunity for researchers to see and study the actual manuscript. The lion's share of the editor's time in preparation of the edition is spent studying the images and creating annotations in the form of searchable descriptive markup based on close scrutiny of the manuscript images. We must therefore ensure that the XML markup in our electronic editions is pervasively associated with the images or parts of images on which the markup is based. To support this association, we introduce some supplemental data into the edition dataset.

First come folio layouts, which store spatial associations between different images available for the same folio. These images include full folio images under different lighting conditions as well as higher-resolution images of certain important folio fragments. An individual layout is created for every manuscript folio. Next, we establish a more detailed association between the manuscript text and its recorded features and the folio layouts. This association is achieved by introducing indexing conventions that tie parts of XML documents and the folio layouts. These conventions can be implemented either by creating and maintaining special "linking" XML markup or by storing linking information in index structures (such as quad-trees and inverted indices). The final solution adopted for the project will be the combination of the two approaches shown in tests to provide the right balance between convenience of use and efficiency of retrieval of information.

HELPING RESEARCH TEAMS WORK TOGETHER

Creation of an electronic edition of a manuscript can be significantly sped up if all members of a research team, rather than a single editor, can work on parts of it simultaneously. As a collaborative process leads to situations where different researchers contribute to the editing process at the same time, edition production software must support concurrent work and be capable of helping the editor resolve data conflicts and prevent resulting loss of data. Otherwise known in database research as concurrency control [EGLT76], this problem presents a number of interesting challenges for our framework, stemming from the complexity of the edition data set and the need to provide as much flexibility as possible to the editor and the research team.

While leaving concurrency control to the editor and the research team may well be the easiest solution to this problem, it is excessively prone to human error and puts an unacceptable burden on the editor. It is instead the task of the back-end of the EPT workbench to automatically assure data integrity at every step of the editorial process and provide adequate means for concurrency control. This support is facilitated, in part, by having multiple DTDs or XML Schemas for editorial markup of the manuscript. Beyond that, we must design and implement specific concurrency control protocols to work with the electronic edition dataset.

These protocols must assure consistency of all changes made to the XML documents and provide significant flexibility for the editor and research team to work simultaneously.

REFERENCES

[BM01] P. V. Biron and A. Malhotra, eds. “XML Schema Part 2: Datatypes. W3C Recommendation.” 2 May 2001. <**http://www.w3.org/tr/xmlschema-2/**>.

[BPS00] T. Bray, J. Paoli, C. M. Sperberg-McQueen, and E. Maler, eds. “Extensible Markup Language (XML) 1.0 (Second Edition). W3C Recommendation.” 6 October 2000. <**http://www.w3.org/tr/rec-xml**>.

[BSe01] M. S. Brown and W. S. 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.

[BSK01] M. S. Brown, W. B. Seales, K. Kiernan, and J. Griffioen. “3D Acquisition and Restoration of Medieval Manuscripts.” *Communications of the ACM: Special Issue on Digital Libraries*. May 2001.

[EGLT76] K.P. Eswaan, J.N. Gray, R.A. Lorie, I.L. Traiger. “The Notions of Consistency and Predicate Locks in a Database System,” *Communications of the ACM*, vol 19, No. 11, Nov. 1976.

[Hay01] D. Hayes. “Glossing Damaged Manuscripts: an Example from AElfric's *Lives of Saints*.” *Digital Resources for the Humanities* (DRH01). University of London, London, UK. 10 July 2001.

[SGK00] W. B. Seales, J. Griffioen, K. Kiernan, C. J. Yuan, and L. Cantara. “The Digital Atheneum: New Technologies for Restoring and Preserving Old Documents.” *Computers in Libraries* 20:2 (February 2000), 26-30. <**http://www.infotoday.com/cilmag/feb00/seales.htm**>.

[YS01] C. J. Yuan and W. B. 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.

PART IV

AN IMAGE-BASED ELECTRONIC EDITION OF ALFRED THE GREAT'S OLD ENGLISH VERSION OF BOETHIUS'S CONSOLATION OF PHILOSOPHY

Kevin Kiernan, Kenneth Carr Hawley, Department of English, University of Kentucky

Alfred the Great’s Old English translation of Boethius’s *Consolation of Philosophy* survives in two manuscripts, one from the tenth and one from the twelfth century, and in an indispensable seventeenth-century transcript and collation of the earliest manuscript, which was badly damaged by fire in the eighteenth century. The confusing manuscript tradition, the catastrophic damage to the principal manuscript, and the easy recourse to the late transcript, have conspired to lead modern editors into producing two major editions, one in prose and one in verse, neither one remotely approximating the manuscript testimony. An image-based electronic edition makes it possible to restore much of the damaged tenth-century manuscript and produce the first prose-and-verse edition of it.

British Library Cotton Otho A. vi, the earliest manuscript of Alfred’s Boethius, is also one of the three manuscript testbeds for the ARCHway Project. The collaboration of computer scientists and specialists in Old English will produce not only practical, modular, extensible tools for non-technical humanities editors to construct complex electronic editions, but also complementary interfaces for scholars, students, and other interested people to make full use of them. Using digital images as the foundation for the Electronic Boethius, we will use these tools to encode searchable glossaries, transcripts, and edition, all linked to the images. Users at all levels will thus have unprecedented access to high-resolution representations of the damaged and fragmentary folios that survive.

The Glossary Tool allows the editor and research assistants to create from a preliminary XML source encoded only for folio lines and edition lines (prose or verse), an exhaustive sorted wordlist with line locations. We then compile the glossary from this wordlist using custom-made templates for each part of speech. These templates permit us to perform complex, comprehensive, XML tagging without even seeing any tags. The resulting XML-encoded glossary is thus fully searchable, and may be reformatted for HTML or any other kind of useful display through XSL transformations.

A complementary Tagger software now under production similarly allows the editor and research assistants to provide pervasive, extremely complex, XML encoding for the transcript and edition, again without the necessity to consult the encoded XML file. The custom-made software, specifically designed to relieve scholars from having to learn markup or face a forest of angle brackets, correctly nests all tags behind the scenes, silently avoiding on behalf of the humanities editors the creation of invalid or non-well-formed

XML encoding. The editor views in one window the more ultimate source of all the markup in the manuscript folio, and tags a transcript of it in another window using clickable element buttons. The tagging comprehensively covers everything from paleographical features to minute physical description of the fire damage to editorial conjectural restorations and emendations. The resulting tagged file is, like the glossary, fully searchable and open to any number of displays. By including coordinates for all tagged parts of an image, searching the xml and the image proceed simultaneously.

Other custom-made tools will provide methodical tagging for other specialized purposes. For example, the editor can easily prepare paleographical descriptions of the scribal letterforms by using templates for specific letters. These individually tagged letterforms can provide students with easy access to paleographical illustrations, which are usually omitted in scholarly editions and sparsely included even in paleographical treatises. Another facility, an image "overlay" device, allows the editor to superimpose any combination of images digitized, for instance, by daylight, fiber-optic, and ultraviolet, to disclose readings rendered obscure or illegible by fire-damage. Users of the completed edition will have access to this device to conduct their own research of these collected images. These specialized tools are all intended to function together, and ultimately to transform the editor's EPT workbench into a virtual reading room for the intended users of the electronic edition.

We will illustrate the process of assembling a scholarly edition with the workbench, tools, and materials we are designing and developing, and discuss how we intend to make it available for widespread use for other humanities projects. The image-based scholarly edition of Alfred's Boethius will comprise a complete collection of all manuscript images intricately linked to edition, transcript, glossary, and apparatus files to allow users to view, read, compare, study, and search in an electronic environment that maintains and encourages connections between text and image. Editorial glosses, emendations, restorations, punctuation marks, paleographical notes, codicological analyses, and bibliographical materials should all be linked to relevant portions of the manuscript from which the scholarship emerges, taking advantage of the strengths of the electronic medium rather than remaining bound by the typical structure and organization of traditional scholarly editions in printed books. The apparatus, for example, should be pervasively available "Help" whenever and wherever pertinent information relating to the edition, transcript, and glossary files, and their associated images, is required. In this way editorial interventions become completely transparent by the availability of high-resolution images alongside corresponding textual notes, explanatory notes, and bibliographical materials.

While it should provide a variety of views and configurations of the edition, transcript, and image files, the user interface for the scholarly edition should also allow users to search exhaustively the collection for words, substrings, grammatical properties, poetic features, and in fact anything that may be of interest. The search facilities for the Electronic Boethius and all the other texts under the aegis of the ARCHway Project will be as functional as the ones developed for the Electronic Beowulf, but will not be bound to a closed system. Like all the other tools in the EPT workbench, these facilities too will be modular, interoperable, and extensible.

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.

REFERENCES:

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. NY: 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., ed. *King Alfred's Anglo-Saxon Version of Boethius, de Consolatione Philosophiae*. Oxford: Clarendon Press, 1899.

Secondary Sources

- Beagrie, Neil and Daniel Greenstein. "A Strategic Policy Framework for Creating and Preserving Digital Collections." Arts and Humanities Data Service (AHDS). Version 5.0, July 1998. Online. <**http://ahds.ac.uk/strategic.pdf**>. Updated July 2001, Christopher Pressler.
- CEDARS Project. "Metadata for Digital Preservation: The CEDARS Project Outline

Specification." March 2000. Online. <**http://www.leeds.ac.uk/cedars/MD-STR~5.pdf**>.

- Committee on Scholarly Editions. Modern Language Association. "Guidelines for Electronic Scholarly Editions." Online. <**http://sunsite.berkeley.edu/MLA/guidelines.html**>. 1 December 1997. Lat updated 18 June 2002. A draft of a proposed revision of the Guidelines for both print and electronic editions is available at: **http://jefferson.village.virginia.edu/~jmu2m/cse/CSEguidelines.html**
- Cover, Robin, ed. *The XML Cover Pages*. Online. <**http://www.oasis-open.org/cover/sgml-xml.html**>. Last modified 24 August 2001.
- Digital Library Federation. "Metadata Encoding and Transmission Standard: an Overview & Tutorial." 14 June 2001. Online. <**http://www.loc.gov/standards/mets/METSOverview.html**>.
- Godden, M. R. "Editing Old English and the Problem of Alfred's Boethius." *The Editing of Old English: Papers from the 1990 Manchester Conference*. D. G. Scragg and Paul E. Szarmach, eds. Cambridge: Brewer, 1994. 163–76.
- Kenney, Anne R. and Paul Conway. "From Analog to Digital: Extending the Preservation Tool Kit." *Collection Management* 22:3/4 (1998), 65–79.
- Kenney, Anne R. and Oya Y. Rieger. *Moving Theory into Practice: Digital Imaging for Libraries and Archives*. Mountain View CA: Research Libraries Group, 2000.
- Kiernan, Kevin. "Alfred the Great's Burnt Boethius." *The Iconic Page in Manuscript, Print, and Digital Culture*. George Bornstein and Theresa Tinkle, eds. Ann Arbor: University of Michigan Press, 1998, 7–32.
- ---. "Creating Electronic Editions from Medieval Manuscripts," forthcoming in *Digital Preservation of Medieval Manuscripts and Early Printed Books*. Milena Dobreva, Serguey Ivanov, Seamus Ross, and Kevin Kiernan, eds. Sofia, Bulgaria: The Institute of Mathematics and Informatics of the Bulgarian Academy of Sciences.
- ---. "Digital Facsimiles in Editing: Some Guidelines for Editors of Image-based Scholarly Editions," forthcoming in *Electronic Textual Editing*, 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'Keeffe, and Lou Burnard.
- ---., with Andrew Prescott, David French, Elizabeth Solopova, Linda Cantara, Michael Ellis, and Cheng Jiun Yuan. *Electronic Beowulf*. 2 CD-ROMS. London: British Library Publications and Ann Arbor: University of Michigan Press, 1999.
- ---. Brent Seales, and James Griffioen. "The Reappearances of St. Basil the Great in British Library MS Cotton Otho B. x," *Image-based Humanities Computing*. ed. Matthew Kirschenbaum, forthcoming in a special issue of *Computers and the Humanities*.
- Lupovici, Catherine and Julien Masanés. "Metadata for Long-Term Preservation." NEDLIB Consortium. July 2000. Online. <**http://www.kb.nl/coop/nedlib/results/preservationmetadata.pdf**>.
- OCLC/RLG Working Group on Preservation Metadata. "Preservation Metadata for Digital Objects: A Review of the State of the Art." 31 January 2001. Online. <**http://www.oclc.org/digitalpreservation/presmeta_wp.pdf**>.
- 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*. C. J. Wright, ed. British Library Publications, 1997. 391–454.
- Research Libraries Group. Commission for Preservation and Access. "Preserving Digital Information: Report of the Task Force on Archiving of Digital Information." May 1996. Online. <**http://www.rlg.org/ArchTF/tfadi.index.htm**>.
- Ross, Seamus. "Changing Trains at Wigan: Digital Preservation and the Future of Scholarship." JISC/NPO Digital Preservation Workshop, 3/4 March 1998, University of Warwick. Online. <**http://www.leeds.ac.uk/cedars/OTHER/SRoss.htm**>.
- Seales, W. Brent, James Griffioen, and Kevin Kiernan. "The Digital Atheneum—Restoring Damaged Manuscripts." *RLG DigiNews* 3:6 (15 December 1999). Online. <**http://www.rlg.org/preserv/diginews/diginews3-6.html#technical1**>.
- 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. Online. <**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*. XML conversion by Syd Bauman, Lou Burnard, Steven DeRose, and Sebastian Rahtz. Chicago and Oxford: TEI P4, 2001.

- Stauffer, Andrew. “Tagging the Rossetti Archive: Methodologies and Praxis.” *Journal of Electronic Publishing* 2 (December 1998). Online. <**http://www.press.umich.edu/jep/04-02/stauffer.html**>. First published in *Revue Informatique et Statistique dans les Sciences Humaines*.
- 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. Online. <**http://www.iath.virginia.edu/~jmu2m/sts2001.html**>.
- ---. “Supporting Digital Scholarship: a Project Funded by the Andrew W. Mellon Foundation.” 1 November 1999. Online. <**http://www.iath.virginia.edu/sds/proposal.html**>.
- World Wide Web Consortium (W3C). “Extensible Markup Language.” Online. <**http://www.w3.org/XML/**>. Last modified 8 November 2002.
- World Wide Web Consortium (W3C). “The Extensible Stylesheet Language (XSL).” Online. <**http://www.w3.org/Style/XSL/Overview-table.html**>. Last modified 10 July 2001.

A Controlled-Corpus Experiment in Authorship Identification by Cross-Entropy

PATRICK JUOLA
Duquesne University
juola@mathcs.duq.edu

HARALD BAAYEN
Max Planck Institute for Psycholinguistics
baayen@mpi.nl

This paper describes an authorship, and more generally document classification, experiment on a difficult Dutch corpus of university writings. By measuring linguistic distances using a cross-entropy technique, a technique sensitive not only to the distributions of language features, but also to their relative intersequencing, classification judgments can be made with great sensitivity, significance, confidence, and accuracy. In particular, despite the designed difficulty of the Dutch corpus used, the technique was still able to reliably detect not only authorship, but also subtle features of register, topic, and even the educational attainments of the author.

Authorship attribution has received significant attention in recent years as a testbed and touchstone for new theories of authorial markers and linguistic statistics; (Holmes, 1998) presents a brief historical outline of some major theories. One of the current front-runners, proposed in (Burrows, 1989), suggests that a strong cue to authorship can be gleaned from a principle components’ analysis (PCA) of the most common function words in a document. Like most statistical techniques, a scholar’s ability to apply this technique is limited by the usual features of sample size, sample representativeness, and test power. Another popular technique, linear discriminant analysis (LDA), may be able to distinguish among previously chosen classes, but as a supervised algorithm, it has so many degrees of freedom that the discriminants it infers may not be “clinically” significant. An alternative technique using measurements of cross-entropy (Wyner, 1996; Juola, 1997, 2003) might be a better tool under difficult circumstances because it is capable of extracting more information (and thus distinguish more readily) along perceptually salient lines from a given data set.

The corpus studied by Baayen, Van Halteren, Neijt, and Tweedie (Baayen, et al. 2002) has been specifically developed to be a difficult problem in authorship attribution. This corpus consists of small writing samples taken from undergraduate students at the University of Nijmegen on carefully and closely controlled topics. Eight students, four first-year and four fourth-year, were asked to write nine essays (for a total of 72) on the same set of closely-specified topics. These topics included three samples of fiction (for example, a retelling of the fairy tale of “Roodkapje,” the Dutch equivalent of “Little Red Riding Hood”), three of argumentative writing, and three of descriptive writing. Document sizes varied from approximately 3600 to 7600 words. These documents were analyzed using cross-entropy and the results compared to analysis via PCA and linear discriminant analysis (LDA) (Baayen et al., 2001) to see both what the appropriate dimensions of variation were, and whether or not authorship attribution was practical in this tightly controlled

a setting.

Analysis using function word PCA showed a marked lack of significant and useful results. Baayen et all found "no authorial structure," although education level and to a lesser extent genre were separated. LDA also could not reliably identify author, and in most experiments performed at chance level. First, as expected, the overall cross-entropy distances showed an extremely strong ability to sort documents by topic (grouping all "Roodkapje" stories together, for instance). Statistical analysis, however, shows a strong (and significant) tendency to within-group homogeneity across all dimensions of interest : topic, genre, author, and even year in school. In particular, the average within-group distance, excluding identity measurements, is significantly less than the average without-group measurement. This holds true, irrespective of whether groups are defined by authorship ($p < 0.000001$), register ($p < 10^{-15}$), or even education level ($p < 0.0084$). This suggests that cross-entropy can be usefully deployed to author identification in this corpus.

A further test revealed that, for every document in the corpus and for every pair of authors, a potential authorship dispute can be resolved approximately 73% of the time using cross-entropy. Furthermore, applying a fractionation technique to restrict attention to only the function words (Juola, 1998; Juola and Baayen, in preparation) could improve these results to 86% correct identification. This can be compared to a best published result of 82% obtained by Baayen et al. using an enhanced LDA model.

Although the enhanced version of LDA performed substantially better, it should be noted, however, that LDA is a supervised technique, and itself makes strong assumptions about the nature of the authorial problem that make it unsuitable to exploratory, unsupervised study.

Even when one restricts the samples to a mere 500 words each, analysis still yields 63% of the pairwise comparisons correctly made, better than any result using PCA or standard (unenhanced) LDA. These results improve upon principle components analysis and standard linear discriminant analysis, with remarkable economy of calculation.

The improved performance can be attributed to information used by cross-entropy but not by word-frequency histograms, specifically in ordering and sequencing of (function) words. Function word PCA performs well by using the relative presence or absence of function words as a cue and/or proxy for idiosyncratic syntactic patterns. Cross-entropy can distinguish not only a difference in presence/absence, but also a difference in relative sequencing. This additional source of information can allow more reliable inferences to be drawn from corpora and more subtle distinctions to be made. In theory, the idea that sequencing is more distinguishing than "mere" presence or absence could be applied to almost any form of linguistic data and any task. Irrespective of what feature sets are eventually found to be informative for authorship attribution, it is to be expected that sequences of features will probably outperform unordered bags of features in making fine distinctions.

REFERENCES

Baayen, H., van Halteren, H., Neijt, A, Tweedie, F. (2002) "An experiment in authorship attribution." *JADT* 2002.

Burrows, J.F. (1989) "An Ocean where each Kind...: Statistical Analysis and Some Major Determinants of Literary style." *Computers and the Humanities*, 23(4–55), pp. 309–21.

Holmes, D.I. (1998) "The Evolution of Stylometry in Humanities Computing." *Literary and Linguistic Computing*, 13(3), pp. 111–7.

Juola, P. (1997) "What can we do with small corpora? Document categorization via cross-entropy." *SimCat* 1997.

Juola, P. (1998) "Measuring Linguistic Complexity : The Morphological Tier." *Journal of Quantitative Linguistics*, 5(3), pp. 206–14.

Juola, P. (2003) "The Time Course of Language Change." *Computers and the Humanities* 37(1).

Juola, P, and Baayen, H. (in preparation). 'Fractionation as a technique for improving document analysis."

Wyner, A. (1996) "Entropy estimation and patterns." 1996 Workshop on Information Theory.

On Determining a Valid Text for Non-Traditional Authorship Attribution Studies: Editing, Unediting, and De-Editing

JOSEPH RUDMAN
Carnegie Mellon
jr20@andrew.cmu.edu

INTRODUCTION:

The work's material history since its inception, the vast and largely uncharted alterations imposed by the history and by the mediation of generation upon generation of printers, editors, publishers—this is a relativism we are prone to ignore, but ignore at our peril. (Marcus 1996)
The literary texts often are not homogenous since they may comprise dialogues, narrative parts, etc. An integrated approach, therefore, would require the development of text sampling tools for selecting the parts of the text that best illustrate an author's style. (Stamatatos et al. 2001)

Most non-traditional authorship attribution studies place too much emphasis on statistics, stylistics, and the computer and not enough focus is given to the integrity and validity of the primary data— the text itself.

It is intuitively obvious and easily shown empirically that if you are conducting a study of the patterns of an author's stylistic usage (e.g. Daniel Defoe), the study will be systematically denigrated by each interpolation of non-Defoe text and even by each interpolation of Defoe text of a different genre or significantly different time period.

The crux of this paper is about one important element in the empirical methodology of a valid non-traditional authorship attribution study—the preparation of the text for stylistic and statistical analysis: unediting, de-editing, and editing.

The general emphasis of this presentation is on prose analysis with some peripheral treatment of drama and poetry.

I. BACKGROUND AND DEFINITIONS

A. Why a valid text is necessary should not even be asked. No valid experiment can be done if the input data is flawed—garbage in, garbage out!

Too many practitioners simply grab a text from any available source—without any thought to its pedigree. (e.g. Khmelev and Tweedie's "Using Markov Chains for the Identification of Writers.")

Are undertakings such as Project Gutenberg or the Oxford Text Archive with their easily available machine readable texts a boon or a bane to non-traditional authorship atudies? This question is explored in some detail.

B. Selecting a starting text

The validity of using texts from the oral tradition and the scribal tradition is discussed.

Before any manipulation and analysis of a text is carried out, a valid starting text must be acquired that fulfills many necessary requirements. This selection is primarily bibliographically driven. If a practitioner is not savvy in the bibliographical arts, a collaborator who is should be recruited.

Examples of bad starting texts causing problems are given (e.g. Peng and Hengartner's "Quantitative Analysis of Literary Styles.")

If you cannot obtain a valid text, do not do the study.

C. **Unediting**—getting back to the state of "not yet edited"
De-editing—removing selected text
Editing—changing (preparing) a text for statistical analysis

II. EXPLICATION

The statement, "each age, each author, each study demands a different mixture of the following particulars," is discussed.

A. Unediting

As a rule, the closest text to the holograph should be found and used.

1. Editorial interpolation
    a. Filled in lacunae
    b. Marginal notation
    c. ‘Changes’ in the text
    d. Critical editions
2. Printer interpolation

    For the Printer is a beast, and understands nothing I can say to him of correcting the press. Dryden (Ward p. 97)

    a. Catchwords (the first word of the next leaf or gathering)
    b. Signatures (combinations of letters and numerals used something like catchwords)
    c. Removing obvious typesetting mistakes (a slippery slope)
    i. ‘f’ for the long ‘s’
    ii. Double words (e.g. ‘the the’ ‘was was’

B. De-editing
1. Quotes
    a. Factual, unattributed
    b. Factual, attributed
    c. Self quotes from earlier writings
2. Plagiarism
    a. Direct copy
    b. Paraphrasing
    c. Imitation
3. Collaboration
    a. Sectional
    b. Phrasal
    c. Word level
    d. Ghostwriting
4. Genre
    a. Poetry, prose, drama, letters, etc.
    b. Mixture (e.g. verse drama)
5. Graphs and Numbers
    a. Tables
    b. Lists
    c. Arabic and Roman numerals
6. Guide words
    a. Titles—chapter headings—the end word ‘Finis’
    b. Marginal annotation
7. Foreign Languages
    a. Sentence level and greater
    b. Phrase or word level
8. Translations
    a. Verbatim
    b. Concepts
9. Examples of items de-edited (or not de-edited) incorrectly by practitioners
    a. Biblical quotes
    b. Titles in direct apposition
    c. Numbers that are spelled out
    d. Words with an initial capital

De-editing

C. Editing
1. Encoding the text
    a. Why (e.g. homographic forms)
    b. TEI
2. Regularizing
    a. Spelling
    b. Contracted forms (simple, compound)
    c. Hyphenation
    d. Masked words (e.g. ‘D_ _ _ e’ for ‘Defoe’)

3. Lemmatizing
    a. Pro
    b. Con

D. Special Problems in Drama and Poetry
1. Stage directions
2. The ‘age’ dependency of transmission and technique.

III. SOME EXAMPLES

Studies that are compromised by mistakes of commission and/or omission in editing, unediting, or de-editing.

A. Historia Augusta
1. Twelve individual studies

B. Shakespeare
1. Eliott and Valenza
2. Foster
3. Horton

C. Defoe
1. Hargevik
2. Rothman

IV. CONCLUSION

1. Some items that are de-edited are valid style markers in their own right (e.g. latin phrases, different genre) and should be treated as such in a parallel study.
2. No matter which text is selected, the practitioner must disclose which text was used and everything that was done to it.
3. The same care must be taken with every text in the study—the anonymous text, the suspected author’s text, and all of the control texts.
4. If valid texts cannot be located and correctly edited, unedited, and de-edited, do not do the study
5. A valid text does not guarantee a valid study. However, a non-valid text guarantees a non-valid study.

REFERENCES

Altick, Richard D., and Fenstermaker, John J. *The Art of Literary Research* (Fourth Edition). New York: W.W. Norton & Company, 1993.

Burrows, John. “Questions of Authorship: Attribution and Beyond. A Lecture Delivered on the Occasion of the Roberto Busa Award.” ACH-ALLC01 Conference. New York University, New York, June 14, 2001.

Eliott, Ward E.Y., and Robert J.Valenza. “So Many Hardballs, So Few Over the Plate: Conclusions From Our ‘Debate’ With Donald Foster.” *Computers and the Humanities* 36 (2002): 455–460.

Foster, Don. *Author Unknown: On the Trail of Anonymous*. New York: Henry Holt and Company, 2000.

Goldgar, Bertrand A. “Imitation and Plagiarism: The Lauder Affair and Its Critical Aftermath.” *Studies in Literary Imagination* 34.1 (2001): 1–16.

Geetham, D. C. *Textual Scholarship: An Introduction*. New York: Garland, 1992.

Grefenstette, Gregory, and Pasi Tapanainen. “What is a Word, What is a Sentence? Problems of Tokenization.” In *Proceedings of the 3rd International Conference on Computational Lexicography*. Budapest: Research Institute for Linguistics, Hungarian Academy of Sciences, 1994. pp. 79–87.

Hargevik, Steig. *The Disputed Assignment of “Memoirs of an English Officer to Daniel Defoe”*. (Part I and Part II) Stockholm: Almqvist and Wiksell, 1974.

Holmes, David I., et al. “A Widow and Her Soldier: Stylometry and the American Civil War.” *Literary and Linguistic Computing* 16.4 (2001): 403–420.

Horton, Thomas B. *The Effectiveness of the Stylometry of Function Words in Discriminating between Shakespeare and Fletcher*. Thesis. university of Edinburg, 1987.

Khmelev, Dmitri V., and Fiona J. Tweedie. “Using Markov Chains for Identification of Writers.” *Literary and Linguistic Computing* 16.3 (2001): 299–307.

Lindey, Alexander. *Plagiarism and Originality*. New York: Harper and Brothers, 1952.

Marcus, Leah S. “Afterword: Confessions of a Reformed Uneditor.“ In *The Renaissance Text: Theory, Editing, Textuality*. Ed. Andrew Murphy. Manchester: Manchester University Press, 2000. pp. 211–216.

Marcus, Leah S. *Unediting the Renaissance: Shakespeare, Marlow, Milton*. London: Routledge, 1996.

Novak, Maximillian E. “The Defoe Canon: Attribution and De-attribution.” *Huntington Library Quarterly* 59.1 (1997): 83–104

Peng, Roger D., and Nicolas W. Hengartner. “Quantitative Analysis of Literary Styles.” *The American Statistician* 56.3 (2002): 175–185.
Project Gutenberg. URL: **http://promo.net/pg/**
Rogers, Pat. *The Text of Great Britain: Theme and Design in Defoe's ‘Tour.’* Cranbury, NJ, 1998.
Rothman, Irving N. “Defoe De-Attributions Scrutinized Under Hargevik Criteria: Applying Stylometrics to the Canon.” *Papers of the Bibliographic Society of America*, 94.3 (2000): 375–398.
Rudman, Joseph. “The State of Authorship Attribution Studies: Some Problems and Solutions.” *Computers and the Humanities*, 31 (1998), 351-365.
Rudman, Joseph. “Non-Traditional Authorship Attribution Studies in the *Historia Augusta*: Some Caveats.” *Literary and Linguistic Computing* 13.3 (1998): 151–157.
Slater, Eliot. *The Problem of “The Reign of King Edward III:” A Statistical Approach*. Cambridge: Cambridge University Press, 1988.
Stamatatos, E., et al. “Computer-Based Authorship Attribution Without Lexical Measures.” *Computers and the Humanities* 35 (2001): 193–214.
Text Encoding Initiative. URL: **http://www.tei-c.org**
Thorp, James. *Watching the Ps & Qs: Editorial Treatment of Accidentals*. Lawrence, Kansas: University of Kansas Printing Service, 1971.
Ward, Charles E. (Collector and Editor). *The Letters of John Dryden: With Letters Addressed to Him*. Durham, NC: Duke University Press, 1942.
Williams, David S. *Stylometric Authorship Studies in Flavius Josephus and Related Literature*. Lewistown, New York: The Edwin Mellen Press, 1992.

Collocations, Authorship Attribution, and Authorial Style

DAVID HOOVER
New York University
david.hoover@nyu.edu

Authorship attribution typically seeks a small number of textual characteristics that distinguish the texts of authors effectively from each other (see Morton, 1978 for a classic discussion). With small groups of texts, these features can be found by examining frequency lists manually, but statistical tests such as the t-test can also be used (see Binongo and Smith, 1999). For the purposes of authorship attribution, a few items occurring at consistent and consistently different frequencies in all of the known texts by all of the claimants may be sufficient for confident attribution.

Most multivariate authorship work focuses on frequent words, following the lead of Burrows (1987, 1988, 1989, 1992a, 1992b, 1994). Much persuasive recent work continues this tradition (Craig, 1999a, 1999b, 1999c, 1999d; Forsyth et al., 1999; Holmes et al., 2001a, 2001b; McKenna and Antonia, 2001; Tweedie et al., 1998). In two recent studies, however, I have shown that cluster analyses based on frequent words often fail to attribute known texts to their authors, and that analyses based on word sequences are sometimes more effective (Hoover, 2001, 2002). Continuing along these lines, I will test the accuracy of analyses based on collocations, while simultaneously examining the effects of using much larger numbers of items than are typically used. Large numbers of words, sequences, and collocations provide more information for potential stylistic analyses, assure that the results take into account a large proportion of the texts under consideration, and, as we will see, usually produce more accurate results. The results of my investigation also show that analyses based on collocations are often more accurate than those based on frequent words or sequences.

For this investigation I will define collocations simply as any two words that appear repeatedly within a certain span of words. Preliminary tests show that, perhaps contrary to intuition, meaningful collocations like house...yard, or car...highway, are not very effective for authorship attribution. They do not occur very frequently, and their occurrence depends too much on the content of the text. Many multivariate analyses have been based on function words alone, in the belief that such frequent and relatively insignificant words are most likely to reflect unconscious and regular authorial habits. This suggests the use of collocations of function words, but preliminary tests show that these are also not very effective. The most effective collocations are simply those that occur at the highest frequencies, with the exception of collocations of personal pronouns, which, like collocations of meaningful words, seem too much conditioned by content (especially the characters) of the texts. I omit personal pronouns and any items for which a single text

provides more than 80% of the occurrences (typically proper names).

To test the effectiveness of collocations in authorship, it seems best to begin with a corpus of texts by known authors, so that various spans, numbers of collocations, and statistical methods can be tested for effectiveness before trying the method on real authorship questions. I begin with a corpus of 10,000 words of pure narrative from fourteen third-person novels by six authors from about 1900, and, as a baseline, test the effectiveness of frequent words and sequences. For the restriction to narrative and to third- person, see Burrows (1987, 1992) and Hoover, (2001). The best results cluster the texts of five of the six authors. Although analysts usually select a small number of items (e.g., the 50 most frequent function words), much larger numbers of frequent words are often more effective. I test the 50, 100, 200, 300, 400, 500, 600, 700, and 800 most frequent items except where fewer items than 800 occur frequently enough to be included. (For this corpus, the best results for frequent words are based on the 300-800 most frequent.) When collocations are tested, various spans and linkages give various results, but several analyses correctly cluster the texts of all six authors, as Fig. 1 shows. A representative completely correct cluster analysis is shown in Fig. 2.

It seems useful to test the methods on another genre, as I did in previous work (Hoover, 2001), so my next corpus consists of the first 4,000 words of twenty-one contemporary literary critical articles by ten authors. Here, analyses based on frequent words and sequences each correctly cluster all of the texts once. Analyses based on collocations with spans of two, five, and ten words also succeed.

Analyses based on collocations seem to be quite effective in attributing texts to their authors in cases of known authorship, and can now be tested in an authorship simulation to see how well they work under conditions that more closely resemble true attribution problems. The simulation includes the fourteen narratives by six authors discussed above, adds four novels by two new authors, and then two “anonymous” novels, each known to be by one of the eight authors. Frequent sequences succeed for only six of the authors. Frequent words still fail to cluster Kipling’s texts correctly, but they do successfully cluster the four texts of the two new authors. They also consistently cluster one of the anonymous texts with Cather’s texts and the other with London’s. Analyses based on collocations with a span of four words are extremely effective and consistent: the 400, 500, 600, 700, and 800 most frequent correctly cluster all of the known texts, even when the graphs are strictly interpreted, as Fig. 3 shows. Like analyses based on frequent words, these also consistently cluster the anonymous texts with those of Cather and London. These identifications are correct. What makes these results even more impressive is the fact that four of the six added texts, including the two anonymous ones, are first-person rather than third-person narratives.

The results of my study confirm what many researchers have found: analyses based on the frequencies of frequent words are quite effective in attributing texts to their authors. Analyses based on frequent sequences of words are also often effective, and are more effective under certain conditions, as I have showed elsewhere (Hoover, 2002). Frequent collocations, however, are often more effective than either words or sequences, producing the only completely correct attributions in some cases and producing more consistently correct attributions in others. The frequencies of frequent collocations clearly reflect important aspects of authorial style. Analyses based on them constitute a promising method of authorship attribution and may also prove useful in stylistic studies.

REFERENCES

Binongo, J. N. G., Smith, M. W. A. (1999). “The application of principal Component analysis to stylometry.” *Literary and Linguistic Computing*, 14(4), pp. 445–65.

Burrows, J. F. (1987). *Computation into Criticism*. Oxford: Clarendon Press.

Burrows, J. F. and Hassall, A. J. (1988). “Anna Boleyn and the authenticity of Fielding's feminine narratives.” *Eighteenth Century Studies*, 21 (1988), pp. 427–453.

Burrows, J. F. (1989). “‘A Vision’ as a revision.” *Eighteenth Century Studies*, 22 (1989), pp. 551–65.

Burrows, J. F. (1992a). “Computers and the study of literature.” In Butler, C. S. (ed.), *Computers and Written Texts*. Oxford: Blackwell, pp. 167–204.

Burrows, J. F. (1992b). “Not unless you ask nicely: the interpretive nexus between analysis and information.” *Literary and Linguistic Computing*, 7(2), pp. 91–109.

Burrows, J. F. and Craig, D. H. (1994). “Lyrical drama and the ‘turbid mountebanks’: styles of dialogue in Romantic and Renaissance tragedy.” *Computers and the Humanities*, 28, pp.63–86.

Craig, H. (1999a). “Authorial attribution and computational stylistics: if you can tell authors apart, have you learned anything about them?” *Literary and Linguistic Computing*, 14(1), pp. 103–113.

Craig, H. (1999b). “Contrast and change in the idiolects of Ben Jonson characters.” *Computers and the Humanities*, 33, pp. 221–40.

Craig, H. (1999c). “Jonsonian chronology and the styles of a tale of a tub.” In Butler, Martin (ed.). *Re-Presenting Ben Jonson: Text, History*, Performance; Macmillan. St. Martin's, Houndmills, England, pp. 210–32.

Craig, H. (1999d). “The weight of numbers: common words and Jonson's dramatic style.” *Ben Jonson Journal: Literary Contexts in the Age of Elizabeth*, James and Charles, 6, pp. 243–59.
Forsyth, R. S., Holmes, D. I., and Tse, Emily K. (1999). “Cicero, Sigonio, and Burrows: investigating the authenticity of the Consolatio”, *Literary and Linguistic Computing*, 14(3), pp. 375–400.
Holmes, D. I., Gordon, L. J., and Wilson, C. (2001a). “A Widow and her Soldier: Stylometry and the American Civil War.” *Literary and Linguistic Computing*, 16(4), pp. 403–420. Holmes, D. I., Robertson, M., and Paez, R. (2001b). “Stephen Crane and the New-York Tribune: a case study in traditional and non-traditional authorship attribution.” *Computers and the Humanities*, 35(3), pp. 315–331.
Hoover, D. L. (2001). “Statistical stylistics and authorship attribution: an empirical investigation.” *Literary and Linguistic Computing*, 16(4), pp. 421–44.
Hoover, D. L. (2002). “New Directions in Statistical Stylistics and Authorship Attribution,” Association for Literary and Linguistic Computing and Association for Computers and the Humanities, Joint International Conference, Tübingen, Germany, July 24–28.
McKenna, C. W. F. and Antonia, A. (2001). “The Statistical Analysis of Style: Reflections on Form, Meaning, and Ideology in the ‘Nausicaa’ Episode of Ulysses.” *Literary and Linguistic Computing*, 16(4), pp. 353–373.
Morton, A. Q. (1978). *Literary Detection: How to Prove Authorship and Fraud in Literature and Documents.* New York: Scribner.
Tweedie, F. J., Holmes, D. I., and Corns, T. N. (1998). “The provenance of De Doctrina Christiana, attributed to John Milton: a statistical investigation.” *Literary and Linguistic Computing*, 13(2), pp. 77–87.

EMMA: Re-forming Composition with XML

NELSON HILTON
University of Georgia
nhilton@english.uga.edu

RON BALTHAZOR
University of Georgia
rlbaltha@english.uga.edu

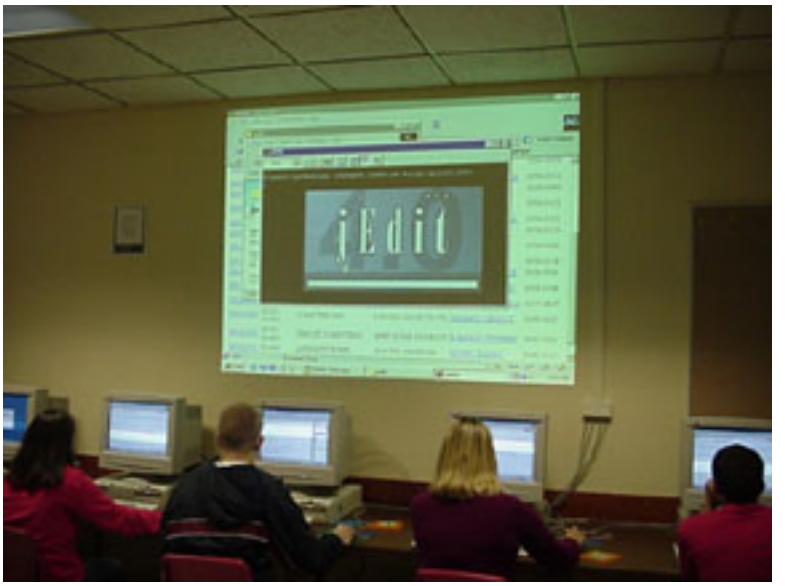
ALEXIS HART
University of Georgia
hart@english.park.uga.edu

ROBERT CUMMINGS
University of Georgia
rec@arches.uga.edu

ANGELA MITCHELL
University of Georgia
amitchel@english.uga.edu

CHRISTY DESMET
University of Georgia
cdesmet@english.uga.edu

Within the past ten years, the computers and writing community has embraced wholeheartedly such computer technologies such as MUDs/MOOs, synchronous online chat rooms, and asynchronous bulletin boards. But the field has yet to recognize the value of markup languages for the production and reception of text. Ideological emphasis on community-building, a preference for process over product, and post-structural fascination with fragmentation may have contributed to a certain blindness toward the potential of markup languages for composition pedagogy. Nevertheless, markup has the ability to transform writing as both product and process and therefore to re-form the field of composition theory and pedagogy.



IMAGINING AN ENGLISH MARKUP AND MANAGEMENT APPLICATION

This paper discusses the inception of the “EngComp Markup” project and the place of EMMA (English Markup and Management Application) generally within English Studies, the English Department as an institution at the University of Georgia, and the English curriculum. College students and teachers recognize already that writing today differs far from the undertaking it was prior to the information technology revolution; they enter the University assuming that research is done on the Internet and that text is processed on a computer. Documents are “engineered” as much as “Englished,” and, increasingly, composed of bits of multi-media content. Grammar—the organizing of grammata, or “letters”—demands a new kind of programmer. But while word processing is now the norm, English Departments and Writing Programs have thus far lacked tools to deal with digital documents across their life-cycle. For the most part, it is still “papers” that are due. Emerging technologies that utilize markup languages, however, will enable a fundamental shift in the very nature of the production, examination, and distribution of text. Moreover, markup serves to relate the heretofore incommensurate structures of narrative argument and database, with profound implications for our conceptions of the organization and articulation of knowledge. Thus, markup has the potential to rehabilitate university writing as a “product” by thoroughly transforming documents into text-in-process and by integrating writing production, review, and evaluation into a single structure.

EMMA’S GENESIS: BUILDING THE CLIENT

This presentation examines the development and implementation of EMMA over the past two years. EMMA employs a variety of technologies, but “her” core is the XML markup language. The editor used for the project is an open-source java editor called jEdit that we have modified for our purpose. The project then uses Cocoon (part of the Apache XML open-source development project), an XML publishing framework running on the Jakarta-Tomcat servlet engine. Cocoon is used to produce server-side XSL transformations of the XML documents. In the initial phase of the project, Cocoon produces a variety of HTML displays, but we continue to work toward using Cocoon as a complete document production and management environment that will include PDF production, a rich search capability, and multiple document concatenation. Currently we are using Filemaker Pro as the database for authentication and document tracking.

EMMA’S DEVELOPMENT: FROM SOFTWARE TO STUDENTS

Although the scope of the EngComp Markup project ranges beyond the First-year Writing Program and even beyond the English Department’s teaching mission, one central feature of that project is the development of EMMA for writing classes at all levels, but especially at the first-year level. EMMA proposes to revise the way in which students write, edit, and submit compositions for review, as well as the way in which instructors and peers can respond by using markup technology to enable web-based collection, modification, distribution, and archiving of student work. EMMA’s customized editing-software provides for instructor-specified DTDs to be used for particular assignments. It also provides a means of ensuring the correct identification and up-load of marked-up compositions back into the system archive, which manages peer review, instructor comments, portfolio creation, assessment, and access permissions.

XML markup has particular utility for the teaching of writing because it allows text to be identified and "carved up" in ways that are customized for content, structure, audience, and rhetorical purpose. More specifically, as a meta-markup language, XML allows document authors to encode data in discipline-specific terminology; to describe specific types of document (e.g., an academic essay); and to tailor a document type to a specific audience (e.g., English scholars or biologists). Furthermore, XML provides descriptively rich information content, stresses the separation of data content from data presentation, describes document content structure and semantic relationships rather than format, and makes information accessible and reusable.

This paper chronicles the intellectual development of EMMA as an idea, and particularly the evolution of the concept of a composition DTD and then multiple DTDs. In Spring semester of 2002, a group of faculty and instructors met to construct what we thought would be an inclusive DTD for First-year writing texts. This "mega-DTD" contained nineteen elements, which in the end proved unwieldy and produced a cluttered display. At the end of this phase of the project, the group came to three conclusions: first, from a design perspective, by necessity the project would follow a "spiral development model," in which development is incremental and cycled through stages of analysis, design, development, and testing; second, the writing program would benefit most from multiple DTDs, customized to fit individual instructors' assignments and pedagogy and, in time, organized into a staged and sequenced writing curriculum; and third, that the goals of a writing program demanded "loose" or California DTD construction. In the first demonstration of EMMA with a group of volunteer English majors, we began to troubleshoot technical problems and refined our sense of timing and focus. In this phase of EMMA's development, students could work with about five tags and two paragraphs per hour to hour-and-a-half session.

EMMA AT WORK IN THE WRITING CLASSROOM

During Fall semester of 2002, EMMA was implemented in ten writing classrooms, eight of them at the First-year level. Teachers with more or less involvement in the development of the client and more or less expertise with technology both incorporated EMMA into their pedagogy during two one-week stints in the First-year Composition Computer Labs. This paper discusses specific pedagogical uses of EMMA at both the macro- and microscopic level. Among the assignments, DTDs, and displays that will be demonstrated are: following argumentative threads through the drafting process; establishing and evaluating ethos; analyzing essay structure; and cultivating style with the "paramedic method."

From the first semester's pilot classes, we learned that EMMA and markup generally can be used to identify structures and information at a variety of levels. Some exercises, such as the "paramedic method," may be especially congenial to EMMA's capabilities because they build in the importance of visual knowledge to writing and revision. EMMA may also function to some extent as an artificial memory system for writers and readers of academic prose. Finally, we discovered that EMMA involves more, rather than less, interaction between teachers and students, and among students and peers; not only does the visual separation of writer from product in display provide the intellectual distance necessary for evaluating prose, but some basic pedagogical problems (e.g., inability to identify basic parts of speech) also make themselves manifest. In Spring 2002, the experiment will be repeated and, for selected classes, EMMA will be incorporated into the entire semester's curriculum instead of only two units.

OBSERVING EMMA: THE FIRST YEAR

During Fall 2002, the Director of First-year Composition at the University of Georgia and a graduate student using EMMA for her dissertation research conducted an ethnographic study of seven teachers and ten classes using EMMA. The study, which is being continued and refined in Spring 2003, involves observation of classes using EMMA and analysis of student products. The study focuses on usability, student attitudes, and the effectiveness of markup pedagogy on final products. Preliminary conclusions fall into the following categories:

- Cognition and Revision: EMMA allows students to see their essays in a "new light";
- Techne and Revelation: EMMA reveals the need for basic instruction, i.e., Parts of Speech;
- Technology and Community: EMMA encourages student-centered classroom pedagogies;
- Administration and Application: and in the future, EMMA may foster more consistent departmental teaching practices.

Initial student response is by nature mixed, but generally positive. In the pilot program, students focused on EMMA's utility for editing and revision, which may depend in part on the kinds of initial assignments developed. Students also used metaphors of vision to a large extent in describing the kinds of "insight" that they derived from using EMMA. The study will be repeated in Spring 2003 and a comparative study of essays from "EMMA" and "regular" composition classrooms will begin.

REFERENCES

Cocoon: **http://xml.apache.org/cocoon/**
jEdit: **http://www.jedit.org/**

Temporal Modelling

JOHANNA DRUCKER
Media Studies, University of Virginia
jrd8e@virginia.edu

BETHANY NOWVISKIE
SpecLab, University of Virginia
bethany@virginia.edu

Temporal Modelling is an interface designed specifically to meet the needs of humanities scholars wanting to interpret or analyse the subjective experience of temporality in historical documents or imaginative artifacts. Since its inception, Temporal Modelling has had two goals: 1) to provide a responsive interface for visualizing complex temporal relations in humanities data and 2) to experiment with an alternative approach to content modelling in humanities computing. Two years into the project, we think we can demonstrate progress on both fronts. We've designed a working prototype suited to humanities scholars that generates XML output concurrent with a user's graphical modelling. Our session describes the interwoven conceptual and technical development of this project (research, visual design development, conceptualization process, and prototype production), and also aims to make it known to a community who we hope will be among its primary users and first adopters.

Temporal Modelling demonstrates an alternative approach to content modelling for humanities computing. In the usual sequence of humanities computing events, a content model is developed and then used to create a richly marked data set. Visualizations or graphical displays follow as a result, not as a point of input. By contrast, we've created a tool that allows visualization to act as a procedure, not an outcome, of interpretation. Graphical sketching integrates interpretation into digitization concretely, creating a content model.

Temporal Modelling consists of two parts. The first we call a composition space or PlaySpace: an interface with a set of visual, graphical elements for making models of temporal relations. The mechanisms and processes of the composition space focus on: the positioning of temporal objects (such as events, intervals, and points in time) on the axis of a timeline; the labelling of those objects using text, color, size, and quality; the relation of objects to specific temporal granularities (the standards by which we mark hours, seasons, aeons); and, in complex interaction, the relation of objects to each other. These can be further modified by a process we call "inflection"—the assignment of attributes with either semantic or syntactic values. For example, a humanities scholar attempting to chart a sequence of events traced in family letters, in which many temporal events are referenced relationally, rather than by date, could use such a system in a preliminary stage of analysis, creating a visual scheme to represent events or references. In addition, however, Temporal Modelling contains tools of interpretation and analysis to characterize events through such inflections as anticipation, regret, foreshadowing or causality.

A user's interpretation is captured and formalized into a structured data scheme that develops concurrently, as the visualization proceeds. User gestures and image renderings are translated into an XML schema which can be exported, used to design a DTD, or transformed through use of XSLT or other manipulations. In the second part of our project, a complementary DisplaySpace will use the same graphical elements to display "published" models from the PlaySpace, as well as supporting display from imported structured data. DisplaySpace models can be manipulated for contrast and comparison in their "published" form or reloaded into the PlaySpace for further refinement. The composition space enables understanding through iterative visual construction in an editing environment that implies infinite visual breadth and depth.

In contrast, the display space channels energy into iterative visual reflection by providing a responsive, richly-layered surface in which subjectivity and inflection in temporal relations are not fashioned but may be reconfigured.

The objects, actions, and relations defined by our schemata and programming are not married inextricably with specific graphics and on-screen animations or display modes. Just as we have provided tools for captioning and coloring (and the ability to regularize custom-made systems with legends and labels), we have also made possible the upload and substitution of user-made standard vector (SVG) graphics for the generic notation systems we've devised. This is more than mere window-dressing. Our intense methodological emphasis on the importance of visual understanding allows the substitution of a single set of graphics (representing inflections for, say, mood or foreshadowing) to alter radically the statements made possible by Temporal Modelling's loose grammar. Users are invited to intervene in the interpretive processes enabled by our tool almost at its root level.

Temporal Modelling can be used to model date-stamped, or empirical data, but its strength is in its ability to structure the representation of subjective experience. Several of the design features of our project embody this conviction, most notably the "now slider". This element embeds an individual point of view (or several, depending on the project) into the graphical model and allows the interpretation to be played forward or backward ("progressed" or "regressed"). The model of events changes depending on the position of these individual now-sliders. An event once anticipated may give rise to an unforeseen set of outcomes and be transformed into an interval charged with regret or melancholy. Temporal Modelling attempts to embody the subjectivity of an interpretation, not merely depict a subjective approach.

Our initial conceptualization was informed by reading across a range of disciplines and fields. We were interested in a historical, trans-cultural inventory of ways time and temporality have been conceptualized. We grounded our fundamental distinction between time as an a priori category and temporality as a relational conception. Our readings were drawn from logic, religion, anthropology, and philosophy, as well other humanities and social sciences (see References). In addition, we made an inventory of visualizations for showing and analyzing data that have a temporal dimension (**http://www.iath.virginia.edu/time**). Because we are intent on creating both a composition space and a display space that can utilize the same set of visual elements, we wanted to consider conventions for presenting date-stamped information while concentrating our design on a system suited to humanities documents whose temporal references and/or relations do not conform to empirical models.

After this research survey, we created a basic conceptual schema for designing the Temporal Modelling project. This involved several steps 1) defining temporal primitives (entities, actions, and relations); 2) developing a graphical vocabulary for their presentation; 3) developing a labelling and annotation system that allows for customization by individual users. Our assertion is that representations based on empirical approaches assume an objective, homogeneous, continuous, and uni-directional notion of time. We wanted to design a system grounded in subjective, heterogeneous, dis-continuous, and multi-directional temporalities. To accomplish this, we conceived of a number of unique design features in addition to the now slider. These include "stretchy" timelines with variable scales and granularities, branching and alternative narratives of temporal events according to catastrophic (event-driven) and continuous (unfolding) models. Designs for ruptured and discontinuous time are also in the plans. The technical implementation of this design involved creating a tightly integrated relation between visualization and digitization of information. In addition, at each instance, the design of specific elements raises interesting intellectual problems at the intersection of philosophical and computational concerns—for instance, is "regret" a semantic attribute of an event, or a syntactic one, engaged in relational effects? Such issues are central to the intellectual and programming structure of this project.

Temporal Modelling is an experiment in using speculative methods— approaches to conceptualization and design of computational tools or projects with uncertain outcomes. The Speculative Computing Lab at the University of Virginia defines its mission as undertaking projects that have a risk of failure. These projects are not necessarily grounded in discipline-specific research and have a tool-based and interpretive aim rather than a collections or archive development goal. Speculative projects are motivated by the desire to foreground these interpretive practices, particularly the subjective practices that are central to traditional humanities. Temporal Modelling is an Intel-sponsored project of the Speculative Computing Lab in the Media Studies Program at the University of Virginia.

REFERENCES

Allen, J.F. "Time and Time Again: The Many Ways to Represent Time." *International Journal of Intelligent Systems*, vol. 6, no. 4, July 1991; pp. 341–355.

Burg, J., et al . "Using Constraint Logic Programming to Analyze the Chronology in A Rose for Emily." *Computers and the Humanities* 34 (4):377–392, December 2000.

Fraser, J.T. *Time, The Familiar Stranger*. Massachusetts UP, 1987.
Jensen, C. S., et al. “A Glossary of Temporal Database Concepts.” Proceedings of ACM SIGMOD International Conference on Management of Data 23, 1, March, 1994.
Jordan, P.W. *Determining the Temporal Ordering of Events in Discourse*. Unpublished masters thesis for Carnegie Mellon Computational Linguistics Program, 1994.
Price, H. “The View from Nowhen” in *Time’s Arrow and Archimedes’ Point*. Oxford UP: New York, 1996.
Reynolds, Teri. “Spacetime and Imagetext.” *Germanic Review* 73(2):161–74. Spring, 1998.
Schreiber, F.A. “Is Time a Real Time? An Overview of Time Ontology in Informatics” in *Real Time Computing*, 1992.
Steedman, M. “The Productions of Time.” (draft tutorial notes 2.0: University of Edinburgh. **ftp://ftp.cis.upenn.edu/pub/steedman/temporality/**).

Virtual Vaudeville: A Live Performance Simulation System

DAVID SALTZ
University of Georgia
saltz@arches.uga.edu

THE PROBLEM: REPRESENTING AND ARCHIVING LIVE PERFORMACE

Manuscripts, paintings, sculptures, films, and recordings are artifacts that can be preserved and archived for subsequent generations to appreciate and analyze. Live theatre, however, is ephemeral. Is it possible to archive a live performance? One can use film or videotape to document a present-day performance, and, with some creative interpretation and speculation, to recreate a performance from the past. But films and videotapes are incapable of conveying the experience of attending a live performance. A filmed performance offers only a single perspective on the action: the camera decides exactly where to look at each moment. Spectators at a live event, by contrast, act as their own camera operators, selecting their own point of focus—which may not even be on stage. Films omits a vital dimension of live performance: the viewer’s immersion in the world of the theater, and the crucial role that the community of spectators plays in constituting a performance event.

The underlying problem here extends beyond the theatrical performance. Precisely the same challenges arise with any kind of performative event, such as dance performances, rituals, political congresses, coronations, parades, festivals, battles, riots, etc.

One strategy to address this problem has been to build a physical reconstruction of an historic structure and to stage performances in it, as has been done, for example, with the Globe Theatre in London and with numerous structures in Colonial Williamsburg in Virginia. This solution requires an extraordinary, continuing investment of money and land, and so is feasible only in a very limited number of cases. Moreover, such physical reconstructions are available only to people at one geographic location and implement only one interpretation, and so they cannot be used to evaluate conflicting scholarly interpretations of the historical evidence. Perhaps the deepest problem with such historical constructions is that, while painstaking efforts may be undertaken to achieve historical accuracy in the physical environment, performers and perhaps even the support personnel, the audience itself—and so, ultimately, the context of reception—remains resolutely contemporary.

OUR SOLUTION: THE LIVE PERFORMACE SIMULATION SYSTEM

In January of 2002, I began work as Principle Investigator on a project designed to address this problem: “A Live Performance Simulation System: Virtual Vaudeville.” Our strategy is to recreate historical performances in a virtual reality environment. Virtual Vaudeville is, in effect, a single-user 3D computer game that allows users to enter a virtual theatre to watch a simulated performance. The objective is to reproduce a feeling of “liveness” in this environment: the sensation of being surrounded by human activity onstage, in the audience and backstage, and the ability to choose where to look at any given time (onstage or off) and to move within the environment. A vital concern is to find a way to bring the nuances of great stage performances into this

virtual environment. To this end, we are using optical motion and facial capture technology to capture real-world performances by professional, highly skilled actors, singers, dancers, acrobats and musicians.

This three-year project is supported by a $900,000 grant from the National Science Foundation, supplemented by an additional $110,000 from the State of Georgia. I am leading a team of researchers from seven universities, including the University of Georgia, the University of Pittsburgh, Georgia Tech and the Naval Postgraduate School in Monterey, that includes historians specializing in nineteenth century American theatre, music and culture, computer scientists specializing in high-performance 3D game design, and theatre practitioners.

Our long-term goal to develop a flexible set of techniques and technologies that scholars and theatre practitioners can use to simulate a wide range of performance traditions, from Classical Greece to Japanese Noh. Our short-term objective is to complete a fully-functional simulation of nineteenth century American vaudeville theatre.

VAUDEVILLE

American vaudeville an especially apt test case for Live Performance Simulation. Vaudeville was the most popular form of entertainment in the United States from the 1880s through the 1920s, functioning in its day much as television does today. Many vaudeville acts both reflected and helped to constitute the enthusiasms and anxieties of their time, especially those concerning the integration of new immigrant groups into mainstream American culture. Consequently a rich simulation of a vaudeville performance will be a useful resource, not just for those interested in theatre history, but for scholars and students of American history generally.

A vaudeville performance was divided into many short, self-contained segments. A typical vaudeville bill encompassed a wide variety of acts— contortionist performances, dance numbers, juggling acts, singing groups, comic monologues, blackface comedy, condenses versions of full-length plays—with particular acts in the lineup appealing differently to different groups in the audience. Consequently, simulating different acts of a vaudeville show and exploring the likely responses of different groups of spectators opens up for historical investigation a wide range of ethnic, gender, class, and racialized interactions during America's industrial age.

Our simulated performance takes place in B.F. Keith’s Union Square Theatre, a typical Vaudeville house seating approximately 2000 spectators, in the year 1895, fifteen years after the first Vaudeville theatre opened in New York. We are recreating four of the most popular and representative acts on the vaudeville circuit during that time: (1) the strongman Sandow the Magnificant; (2) the Irish singer Maggie Cline; (3) the comic “stage Jew” Frank Bush; and (4) the sketch comedy of the four Cohans, whose youngest member, George M. Cohan, went on to become one of the great stars of early twentieth century Broadway. As we approach the end of our first year of work on the project, have completed archival research into all of these acts and are creating the models and motion-capturing the performances. Our presentation will feature a demonstration of significant portions of the Sandow act that as of this writing are complete and fully-functional.

DESIGN

Virtual Vaudeville allows the user to switch between two very different ways of experiencing the simulated performances. In what we call “invisible camera” mode, viewers fly through the 3D space to observe the performance from any position in the theatre and zoom in as close to the performers as they please.

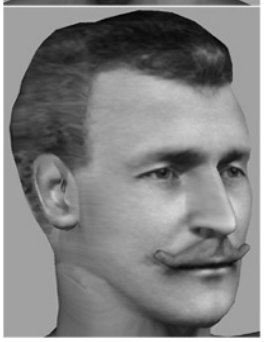
Alternatively, the viewer can adopt an embodied perspective, watching the performance through the eyes of a particular member of the audience. Virtual Vaudeville allows the viewer to select one of four spectators, each representing a different socio-economic group in 19th century America: (1) Mrs. Dorothy Shopper, a wealthy socialite attending the performance with her young daughter; (2) Mr. Luigi Calzilaio, an Italian immigrant fresh off the boat, attending the performance with his more Americanized brother; (3) Mr. Jake Spender, a young “sport” sitting next to a Chorus Girl (with him he may or may not strike up a relationship, depending on the viewer’s choices); and (4) Miss Lucy Teacher, an African American schoolteacher watching the performance with her boyfriend from the second balcony, where she is confined by the theatre’s segregation policy.

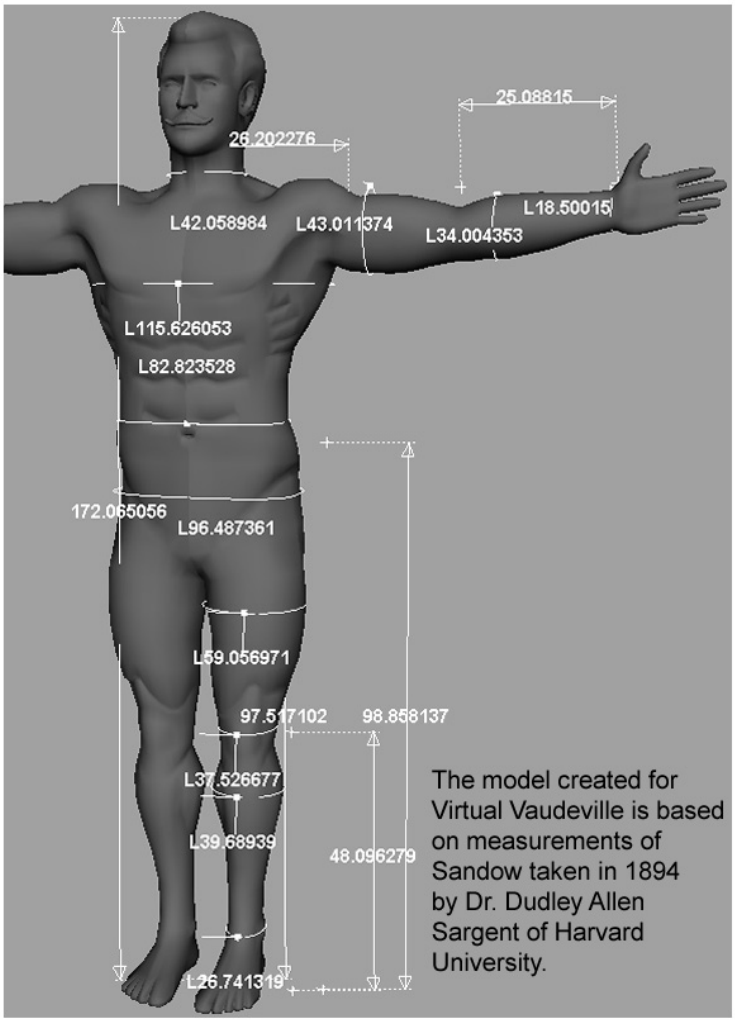
Viewers can switch between any of the avatars at any time, and can move the avatar’s head to focus on different areas of the stage or auditorium, and can trigger a limited set of avatar responses, for example applauding or laughing. Some of these responses are verbal, such as cracking a joke or heckling the performance. In these cases, the viewer selects only the generic response type, and the system produces a specific response appropriate what is happening onstage and off, taking into account the viewer’s previous interactions with other members of the virtual audience. Because the surrounding spectators respond interactively to the viewer's avatar, each viewer has a different experience of the performance event.

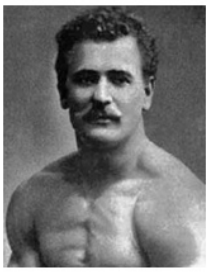
SIGNIFICANCE

Virtual Vaudeville offers scholars in all disciplines in the humanities a model for a new kind of "critical edition." A conventional published monograph can pick and choose details to examine, and so lacuna and even contradictions in the historical analysis are easy to overlook. The imperative of precisely recreating both on-stage and off-stage events will demand an unprecedented degree of scholarly thoroughness and rigor.

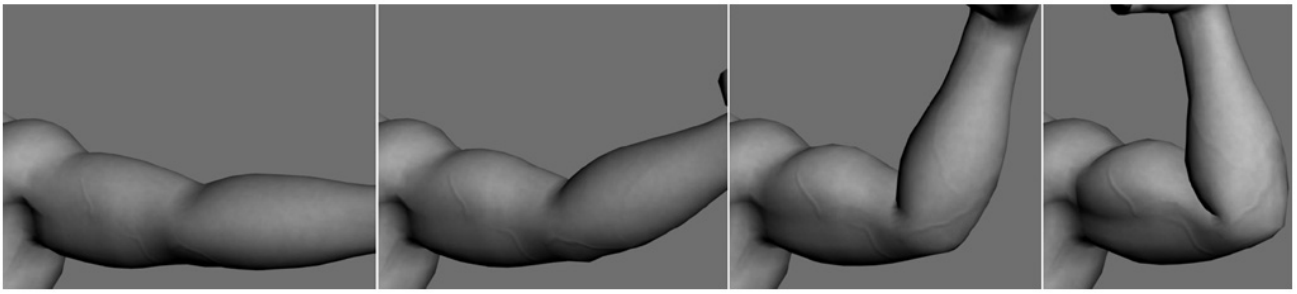
Virtual Vaudeville
Preliminary 3D Models of Sandow







Photograph of Sandow in 1995



Key to our project is the depth of the collaboration between technology, scholarship, pedagogy and art. This project is conceived to make a significant contribution to all four domains simultaneously, rather than merely using any one in the service of the others. The end result, we, hope, will represent an important advance in the design and implementation of virtual environments, building on recent successes in creating photo-realistic simulations of real 3D environments by introducing a large quantity of complex human performance data. It will constitute an invaluable work of applied scholarship, an unprecedented resource for visualizing past performances and testing hypotheses about historical performance practices. It will provide an unprecedented resource for students to engage with historical performance traditions as performance (and

not as literature or film). Finally, from an artistic perspective, the Virtual Vaudeville project will test the potential of virtual reality technology to provide truly nuanced and engaging theater experiences.

Tagging Time in PROLOG: from quick and dirty to TEI

JAN CHRISTOPH MEISTER
University of Hamburg
jan-c-meister@uni-hamburg.de

This paper presents work in progress from a current research project at Hamburg University that employs Humanities Computing methodology for developing and testing a new theory and model of "narrative time". Our premise is that narrative time should be defined in functional and not in essential or categorical terms: time is not an objective phenomenon, but a cognitive construct and can thus best be modeled in terms of a 'temporality effect'. This effect -- that is, the impression of temporal order in narrative, both on the level of fictional reality and narrative discourse – is to be explained and analyzed in terms of the distribution of empirical 'notions' (representations of objects) and 'temporal operators' throughout a representational medium, in our case: a narrative text.

Humanities Computing methodology plays a central role with regard to both the description (markup) and the subsequent combinatory analysis of relevant textual elements. However, adhering to a TEI compliant tagging approach proves unacceptably complicated. The paper therefore argues for a *quick and dirty* approach to time tagging based on feature structure tags that are defined in the form of PROLOG clauses.

THEORY

To date most theories of narrative—in particular those focusing on the domain of literary narratives—conceptualize of 'time' in terms of a dichotomy of *narrated time* vs. *time of narration* or, as Günter Müller's classic formulation goes, of *Erzählzeit* vs. *erzählte Zeit*. This is essentially an ontological distinction that attempts to set apart two 'worlds', each of with is seen to have its 'own time'. However, this distinction immediately becomes problematic when dealing with non-fictional representations of events which, irrespective of chronological proximity, are by definition situated on a singular objective time line. Our approach is therefore based not on the traditional narratological concept, but rather on the unitary model of time originally proposed by McTavern who distinguished between two *perspectives* onto time: namely, that of events in an objective *before-after* relationship (the so-called *B-series of time*), and that of events as occurring in the subjective cognitive order of *future-present-past* (*A-series*).

COMPUTER-BASED IMPLEMENTATION

As far as tools are concerned, the implementation of this theoretical model in a Humanities Computing orientated project has necessitated the development of two programs:

- TempusMarker -- a software tool providing automatic and semi-automatic markup routines for the tagging of temporal expressions in natural language texts. A prototype of TempusMarker has already been programmed.
- TempusParser -- an analytical tool that generates a version (or versions, as the case may be) of the base text in which all the sequences that form a complex narrative discourse are organized in strict chronological order. This (re)construction is the result of an algorithm driven process of analysis and recombination of textual segments during which the 'time stamp' of each segment as indicated by the temporal tags is interpreted.

orientated project has necessitated the development of two programs:

The computational implementation of McTavern's model for the purpose of concrete analyses of *A-series* governed textual discourse and its eventual reconstruction in terms of *B-series* ordered event sequences offers an interesting example for the difficulties faced by the computing Humanist who tries to tackle even modestly 'intelligent' hermeneutic problems. In the project to be presented here an added problem stems from its empirical orientation: rather than having experts (Literary Scholars) tag the texts we are using student groups in order to simulate as closely as possible the 'naïve' reader's processing habits. A demonstration of the actual tagging process, including the use of the TempusMarker prototype, will form part of the presentation.

Whereas the temporal value of the respective natural language expressions -- be they denotative or deictic -- is comparatively easy to establish either contextually, or from a dictionary, their explication in the form of standardized TEI markup (core tag set + additional tag sets for dates and time, ref. chapter 20.4. of

TEI guidelines) has proven rather unwieldy and often contra-intuitive to our readers. Defining more readily comprehensible *feature structure tags* would seem to be the alternative of choice; however, this raises the methodological question of how to design a sufficiently fine-grained feature structure that does not automatically become completely idiosyncratic to the particular research problem at hand, thus inadvertently restricting the uses of the tagged corpus at a later stage.

Against this background the paper advocates and will demonstrate a calculated *quick and dirty* approach to designing temporal feature structure tags. In particular, it will be shown how the PROLOG predicate structure can facilitate rapid prototyping of feature structure tags.

CONCLUSION

Expressing relatively complex hermeneutical problems and models in terms of Humanities Computing methodology and standards should be conceived of as a process of translation, rather than one of mere re-presentation. From a hermeneutical point of view semantic *tags* are not just descriptors, but rather predicates of a prepositional clause in which the tagged string itself is one argument, and its feature values the subsequent arguments. Capturing experimental temporal feature structure tags in the form of PROLOG predicates therefore holds two advantages: first, it offers a more intuitive approach to semantic tagging. Second, it facilitates automatic conversion of feature structures into composite TEI tags at a later stage, thus turning the *quick and dirty* into the beautifully intricate -- and fast at that.

NOTES

[1.] Available for download at **http://www.narratology.net/html/de005g3.html**. A detailed description of the tool (in German) is included.

[2.] A schematic representation of the TempusParser architecture and workflow is available at **http://www.narratology.net/html/de005graph.html**

REFERENCES

Habel, Christopher; Schilder, Frank: *From Temporal Expressions to Temporal Information: Semantic Tagging of News Messages.* **http://www.cs.brandeis.edu/~jamesp/arda/time/readings/schilder_habel.pdf** - 27.11.2002

Dammann, Günter; Meister, Jan Christoph: *The temporality effect: Design and computer-based application of a constituent model of narrative temporal order*. **http://www.narratology.net/html/de005.html** - 27.11.2002 (contains hyperlinks to the prototype of the TempusMarker tagging tool.)

John McTaggart EllisMcTaggart: *The Unreality of Time*. In: *Mind 17* (1908), p. 457-474.

Günther Müller: *Erzählzeit und erzählte Zeit*. In: *Festschrift für Paul Kluckhohn und Hermann Schneider*, Tübingen 1948, S. 195-212.

TEI Consortium: *TEI P4 - Guidelines for Electronic Text Encoding and Interchange. XML-compatible edition*. Edited by C M Sperberg-McQueen and Lou Burnard; 2001: (Chapter 6.4: Names and Dates, and Chapter 20.4: Names, Numbers, Dates, Abbreviations, and Addresses. **http://www.tei-c.org/Guidelines/index.html**

Deep Encoding

SUSAN BROWN
University of Guelph
sbrown@uoguelph.ca

WILLARD MCCARTY
King's College London
willard.mccarty@kcl.ac.uk

WENDELL PIEZ
Mulberry Technologies
wapiez@mulberrytech.com

PART I

It seems reasonable to assume that our experience of markup has to date been chiefly in its application to format. The publishing of texts for reading, in print and otherwise, demands it; the imperative to present documents in a variety of forms to suit a variety of media assume it; the examples used to illustrate approaches to markup demonstrate it again and again. Almost entirely the declarative approach to markup—which asserts that ontology rather than appearance should be encoded—is thus used to achieve a desired appearance by identifying the ontology. Although the ontology is what matters and is what we wish to communicate, appearance is how we communicate it to fellow humans. I declare this to be a paragraph because I want it to look like one so that the reader will see it for what it is.

For the purposes of argument let us then say that such formatting features, however important and however dense, characterize the surface of a text. But what about the textual depths? What happens when markup is used to identify textual entities whose existence the reader normally infers from the language of a text or on the basis of ambiguous signs, such as capitalized letters provide? Such entities are a matter for interpretation in a sense hardly applicable to a paragraph under usual circumstances but often to a capitalized word, especially in older texts, and always to a figure of speech. So, we might ask, what happens when markup is used as a language of interpretation as exegetes commonly understand the term?

In this session three papers will address markup as interpretative language by converging on the notion of "deep" encoding. Susan Brown, in "Delivering the Depths", will reflect on and analyze the interpretative dimensions of the *Orlando* Project, which uses markup as a compositional tool. Willard McCarty, in "Modelling the depths of a literary encoding", will discuss how his rethinking of a grammatical approach to personification in the *Analytical Onomasticon* Project creates and defines encoding depth. Wendell Piez, in "Whither Deep Markup?", examines the two projects in a theoretical context and draws conclusions and further questions from them.

PART II

DELIVERING THE DEPTHS: REPRESENTING THE ORLANDO PROJECT'S INTERPRETIVE MARKUP

Susan Brown

We begin with a brief review of the *Orlando* Project, a highly experimental effort to devise new ways of writing literary history by use of encoding (Brown 1998). We argue that this encoding is "deep" in the sense that we use SGML to tag not only structural but also intellectual properties of the text. In this paper we draw out the implications of deep encoding so defined for the scholarly writer.

The *Orlando* team devised four DTDs to reflect the complexity of the material to be written. These are used to compose "originally digital" text (Unsworth 2001) in the following forms: 1) discrete chronological items; 2) longer topic items; 3) biographical documents; 4) documents about the writing careers and publications of authors. The fourth and most complex of these, the "writing" DTD, is the focus of this paper.

The writing DTD has 144 tags, which, in our more than 1200 documents of this type, are used with

frequencies ranging from 30,567 times (the <BIBCIT> or bibliographic citation tag) and only once (the <DAY> tag). As these examples suggest, some tags resemble those of conventional markup, since we conform to the TEI where possible. However, the desire of the team to encode the intellectual priorities of the project and permit new ways of ordering and accessing literary historical documents resulted in numerous unique tags, including the following, listed here with their frequency of use to date:

PLASTLITERARYACTIVITY	141
PLITERARYSCHOOLS	121
PMANUSCRIPTHISTORY	363
PMATERIALCONDITIONS	654
PMODEOFPUBLICATION	877
PMOTIVES	594

These tag names, and the various attribute names and values associated with some of them, are far from transparent to the average person or even to someone schooled in the TEI. Nor are they transparent to the average literary historian or critic. “Mode of publication”, for instance, might be understood by some scholars to refer to aspects of the publication format (like folio or octavo) rather than or in addition to the aspects of publication (like private printing or subscription) that the *Orlando* Project has singled out for demarcation with this tag. This is a single example of how English studies has resisted attempts to standardize vocabulary and method; identical terms operate differently in different contexts. Indeed, precisely because the priorities and organizing principles of literary historical studies have often been left implicit, the *Orlando* team believes that markup offers an innovative method of making crucial aspects of a study evident for debate (Brown 2001b).

Each of the 65 unique writing DTD tags has thus been designated by project members as of crucial importance in the production of our particular view of women's literary history. The *Orlando* Project tagset thus interleaves familiar structural tags, such as <P>s with others, which we will call here “content tags”. The project's markup is thus deep in the sense that it is designed to structure the history intellectually, to be a “diagnostic and investigatory instrument” (Piez 2001, 197).

This tagging enterprise is experimental in a range of ways. The DTDs were designed to try to reflect the conceptions—quite abstract, although derived from specific examples and much familiarity with women's literary history—of scholars experienced in literary history but new to text markup. The DTDs were designed not only without the traditional process of document analysis—since there were no pre-existing texts, nor did the team wish to model the DTDs on print textuality—but largely without specific notions of how the content markup would translate for delivery. In other words, the DTDs embody the abstract principles informing the project as a whole, and the documents flowed from that. The DTDs were designed to permit complex scholarly prose, which meant not only devising specific content tags, but allowing for flexibility within the DTD, both in placement of tags and in their combination with each other. As a result, the writing DTD is significantly flatter than most other DTDs, even including our own biography DTD, when one compares their content tag hierarchies. This allows for considerably greater freedom in the structuring of documents, since exclusive content hierarchies are only minimally enforced. Authors of texts must situate their writing within one of four basic content tags:

1. <SUMMARY>;
2. <PRODUCTION>;
3. <TEXTUALFEATURES>;
4. <RECEPTION>.

Most of the other content tags are grouped according to one of the latter three categories. Thus, all the tags with the “p” prefix discussed above, belong conceptually to the “production” group. However, the DTD does not make use of those tags exclusive: each set of subtags is an inclusion of the <RECEPTION> and <TEXTUALFEATURES> tags the as the <PRODUCTION> one, so that, for instance, the <PMOTIVES> tag (with its attribute for MOTIVETYPE with attribute values of “Attributed” or “Self-identified”) is valid in the midst of a discussion of a particular text's <RECEPTION>. This addresses the problem of the messiness of actual critical prose, wherein one might be discussing the negative response of a critic who attributed certain motivation to the author, or where her own statement of motive might be embedded in an assessment of her own work. In effect, we incorporated overlapping hierarchies, because “the disorderly world of real data” is not well represented by exclusive hierarchies, and making data fit a conventional model can lead to “an impoverished view of the data.” (Durusau and O'Donnell).

The tagset is implemented in documents shaped as much by the markup language with which they are written as by the literary historical research that informs them. This has resulted in a rich representation of our data, notwithstanding differences in the application of some tags that result from the fact that they are, in fact,

"interpretive" (Butler 2000; Brown 2001a). In fact, the line between data and markup dissolves in the case of *Orlando* texts, since the argument of the text can be as much in the markup as in the critical prose—hence, again, it is deep. However, given the lack of consensus on method and vocabulary mentioned above, the delivery of this material poses considerable challenges if the markup is to allow users to 1) exploit it for their own purposes; and 2) examine the project's principles and priorities.

The danger of producing a massive alienation effect in a group of users often wary of new technologies cannot be overemphasized. Even with our extensive in-house training and documentation, it takes new student assistants at least a semester of part-time work (about 120 hours) to develop minimal proficiency with the tagset, and often twice that time to develop the confidence to write and tag original contributions. We cannot expect a similar time investment from our users.

The paper will demonstrate aspects of the *Orlando* Project delivery system that represent and educate users in its deep markup. These include:

- Chronology searches that make the presence of a limited number of central content tags known
- Organization of hyperlinking using content tags, to familiarize users with the structure and conceptual organization of the DTDs
- "Thematic" pages introducing the power and specificity of searches on the tagset
- A full text search engine that offers a panorama of the tags and their relationships, and allows searches to be constructed without knowledge of search syntax
- A full text search results screen that offers a user-friendly view of the content markup

users in its deep markup. These include:

REFERENCES

Brown, Susan, Sue Fisher, Patricia Clements, Katherine Binhammer, Terry Butler, Kathryn Carter, Isobel Grundy, and Susan Hockey. 1998. "SGML and the *Orlando* Project: Descriptive Markup for an Electronic History of Women's Writing." *Computers and the Humanities* 31: 271–85.

Brown, Susan, and Isobel Grundy; with Renee Elio, Patricia Clements, Sharon Balazs, Rebecca Cameron, Dave Gomboc, Allen Renear, Jeanne Wood. 2001a. "Intertextual Encoding in the Writing of Women's Literary History." ALLC/ACH 2001.

Brown, Susan, Isobel Grundy, et al. (2001b). Session of three *Orlando* Project papers: "The Hard and the Soft: Encoding Literary History," and "Risking E-Race-Sure/Erasure: Encoding Cultural Formations." Annual Digital Research in the Humanities Conference, School of African and Oriental Studies, London University, UK, 9 July 2001.

Butler, Terry, Sue Fisher, Susan Hockey, Greg Coulombe, Patricia Clements, Susan Brown, S, Isobel Grundy, Kate Carter, Kathryn Harvey, Jeanne Wood (2000). "Can a Team Tag Consistently? Experiences on the *Orlando* Project." *Markup Languages Theory and Practice* 2: 111–125.

Durusau, Patrick and Matthew Brook O'Donnell, "Overlapping Hierarchies/Concurrent Markup" **http://www.sbl-site2.org/Overlap/**).

Piez, Wendell. "Beyond the 'descriptive vs. procedural' distinction." Extreme Markup Languages 2001. Online at: **http://www.piez.org/wendell/papers/beyonddistinction.pdf**.

Unsworth, John. "Publishing Originally Digital Scholarship at the University of Virginia" ACH/ALLC 2001.

PART III
MODELLING THE DEPTHS OF A LITERARY ENCODING, WITH AN EXAMPLE FROM OVID

Willard McCarty

This paper explores some consequences of encoding a complex, intricate and lengthy poem for a densely occurring set of literary phenomena. In it I dwell on one of these phenomena, personification, in order to define and illustrate the idea of *deep encoding* and to reflect on its implications. I am not concerned with details of the metalanguage as such, except in one particular point. Nor am I much concerned with the tagged text or the scholarly product generated from it. Rather I focus on the epistemology of marking up a text: I ask, how do we make new knowledge rather than simply record what we already, somehow know, or think we know? In Wendell Piez's terms, in what sense may the markup of an ancient text be *proleptic* (Piez 2001)?

To establish the background of my research to date I briefly review material presented in prior ACH/ALLC conferences (Paris 1994; Kingston 1997; Virginia 1999).

The poem in question is the *Metamorphoses*, written in classical Latin hexameter by the Roman poet Publius Ovidius Naso at approximately the beginning of the Christian era. The *Met* is a problematically organized collection of mythological stories that became the primary means by which much of the

Greco-Roman heritage was handed down through the Middle Ages to the present day. The literary cohesiveness of the *Met*—how it works as a poem—is therefore an important problem, but it has proven a considerable challenge to critics. My approach to this problem, in the *Analytical Onomasticon* project (McCarty 2002), is to encode all devices of language that refer to persons (hence the Onomasticon, or "book of names"), and from the resulting tags to generate a set of indexes for constructing interrelations among stories. The *Onomasticon* is thus intended to aid further literary-critical study, especially along the lines of Wheeler 2000 and to a somewhat lesser extent Schmidt 1991.

In McCarty 1998 and 1999 I made first attempts to consider the epistemological consequences of markup by isolating treatment of personification (i.e. the creation of persons by rhetorical means, Paxson 1994). In this paper I take issue with these attempts, in which I applied the term "grammar" loosely to a set of criteria for encoding without giving any regard to implementation. I argue here, rather, that these criteria comprise only a first step. I adapt the Chomskian sense of a grammar as a computable set of rules by which the linguistic phenomena in question may be generated (Thorne 1972: 184–6). Hence, in these terms, for there to be a *grammar of personification*, individual instances must be computable from the primitive elements which can be said to be responsible for them. These primitive elements are described in some detail in McCarty 1998—for example, apostrophe, familial relationship and mental activity. How, then, might they be implemented such that instances of personification could be generated from a true grammar?

In the paper I give a few examples from the *Metamorphoses* to indicate the complexity that a grammatical perspective on Ovidian personification involves. When viewed in such a way, personification comprises not only the linguistic and rhetorical factors local to a candidate, but also five larger contexts that, according to my proto-grammar, affect the success of these factors: onomastic, narrative, poetic-mythological, ontological, personal. Furthermore, moving from a declarative phenomenology of personification to an actual grammar means that the rules by which success is computed themselves become part of the problem.

In the first instance the subject of this paper is not how to write these rules. Rather it is how best to conceptualize the kind of problem for encoding that personification exemplifies. How we compute a complex textual phenomenon, whether from the metadata or directly from the data, is an important aspect of this question, but I approach it indirectly with the help of two ideas: *depth* and *modelling*.

At some level markup must be simply declarative: thus, "X exists here". Let us define a *shallow* encoding as one in which only such declarative statements are possible or are admitted. One might say in such a case that the phenomena of interest are the primitives of the system— in a builder's terms, for example, the bricks, mortar, timbers, tiles, plaster and paint. Contrast the encoding of personification as described. It is an example of a deep (or at least deeper) kind because the object of study is not simply declared as a primitive but may be computed from declarable elements. Thus, if our builder were to take a particular interest in paint, it would cease to be a primitive in his system, rather would be something to be mixed from primary pigments and a carrier base.

Note that an encoded work may not be uniformly deep: for example, some objects of study in the *Onomasticon*, such as proper names, are simply declared as such, and within the work as conceived, it makes no sense whatever to compute them from more primitive elements. Hence an encoding may have depths but not be simply deep. Note also that a deep encoding is not necessarily a *thick* one, to borrow Wendell's term: the density of tagging is a separate matter.

Depth, then, is not a given but is created by the encoding scholar's focus in combination with available theory, the analytical tools at hand and the nature of the phenomena. It is expressed grammatically by a synthesis involving primitive elements and rules for their combination. It is a relative, not an absolute term. It implies no value-judgment, although a completely shallow encoding, however useful to scholarship, is hardly itself a scholarly work. And a *thin* encoding, however deep, would not give scholarship much scope.

The idea of depth has value in that it expresses a qualified imperative to extend what can be seen to a finer level of detail. The idea of a grammar has value because a grammar is independent of the text to which it is applied; it is exportable, with the result that either it is universalized or the phenomenon in question is resolved into two or more related kinds. Thus, for example, if depth is created in encoding personification within the *Metamorphoses*, the resulting grammar can be applied to other Ovidian works, others in classical Latin and in various languages across different historical periods. What is personification, exactly? Perhaps plumbing its depths through encoding will yield us much more revealing questions than that one.

Depth, I suggested, has a temporal dimension contingent upon developing ideas and interests in literary texts. Its method of expression, however, can aid that development directly by the way in which computational grammars are designed and published. If, that is, we conceive our grammars as devices for *modelling* phenomena of interest—i.e., manipulating representations of these phenomena—and if we publish these modelling tools, then our colleagues will be able to continue the work we begin, not simply agree or disagree with it. Such a modelling tool, e.g. for personification, would not simply allow for differing views of the trope to be expressed, rather more for them to be explored, particularly fruitful ones identified and so the phenomenon itself better to be understood.

In conclusion I return to the epistemological question with which I began: how may markup be used to get beyond the recording of old knowledge, with consequent reliance on sorting and formatting what we already know? How is new knowledge made through markup? Deep encoding, I argue, provides a powerful answer so long as we understand that depths are to be modeled and that successful modelling turns them into shallows while revealing further depths beneath.

This paper, then, constitutes a criticism of previous work and a proposal for a deepening of it. The existing proto-grammar is, as far as I can determine, the first of its kind for personification but is not in a form that can be tested. In the paper I sketch a plan for this work and show an example of how the markup might be done.

REFERENCES

Craik, F. and R. Lockhart. 1972. “Levels of processing: A framework for memory research“. *Journal of Verbal Learning & Verbal Behavior* 11: 671–684.

McCarty, Willard. 1998. “A provisional grammar of personification for the *Met*”. In “What is humanities computing? Toward a definition of the field”. **http://www.kcl.ac.uk/humanities/cch/wlm/essays/what/grammar.html** (23/11/02).

---. 1999. “Thinking with markup: the case of personification”. ACH/ALLC annual conference, University of Virginia, 9–13 June 1999.

--- (with John Bradley, Monica Matthews, Aara Suksi, Burton Wright). 2002. *An Analytical Onomasticon to the Metamorphoses of Ovid*. **http://www.kcl.ac.uk/humanities/cch/wlm/analyticalonomasticon/** (24/11/02).

Paxson, James J. 1994. *The Poetics of Personification*. Literature, Culture, Theory 6. Cambridge.

Piez, Wendell. 2001. “Beyond the ‘descriptive vs. procedural’ distinction”. *Extreme Markup Languages 2001*: 210ff. **http://www.piez.org/wendell/papers/beyonddistinction.pdf** (24/11/02).

Schmidt, Ernst. 1991. *Ovid's Poetische Menschenwelt: Die Metamorphosen als Metapher und Symphonie*. Heidelberg.

“Semantic encoding.” *Glossary*. Non-Roman Script Initiative, SIL International. **http://www.sil.org/nrsi/glossary.htm#semenc** (23/11/02).

Thorne, James Peter. 1972. “Models for Grammars”. In Teodor Shanin, ed., *The Rules of the Game: Cross-Disciplinary Essays on Models in Scholarly Thought*. London: Tavistock: 179-205.

Wheeler, Stephen M. 2000. *Narrative Dynamics in Ovid's Metamorphoses*. Classica Monacensia, Bd. 20. Tübingen: Gunter Narr Verlag. (See the helpful review by Alden Smith, *Bryn Mawr Classical Reviews* 2001.11.23 (**http://ccat.sas.upenn.edu/bmcr/2001/2001-11-23.html**), and Wheeler's response in *BMCR* 2001.12.18 (**http://ccat.sas.upenn.edu/bmcr/2001/2001-12-18.html**).

PART IV
WHITHER DEEP MARKUP?

Wendell Piez

The *Analytical Onomasticon* and *Orlando* projects are examples of electronic text research growing in interest as they tackle successive challenges—and for completely different reasons. In the two other papers offered in this session, the presenters have detailed why each of them can fairly lay claim to the rubric of “deep markup”, along with some explanation of what that may mean. In both cases, the claim is justified, because both of them—and again, in very different ways—go considerably beyond not only applications of markup that are merely presentational or utilitarian, hence “shallow” or “light”, but even beyond markup that is merely “thick”. They are coming not only to provide new kinds of access to information, but actually to suggest new arguments respecting new kinds of knowledge about their subjects. In other words, they are defining what in a markup system is their content.

Yet there is a distinct difference, as is evident in these two presentations, separating the subjects, designs and goals of the projects.

It is useful here to distinguish between a markup design that is *prospective* and one that is *retrospective*. Here the difference may be obscured by an important similarity: both markup schemes, as markup, are descriptive and declarative; that is, they are given not to direct processing specifically but to some kind of declarations of abstract type. Based in part on whether they look forward or back, and granting there to be an intriguing overlap between them and many mixed examples, I have elsewhere distinguished these two different forms of descriptive markup as “proleptic” and “metaleptic” (Piez 2001). *Proleptic* markup is descriptive markup that looks forward to future processing, indeed to future signification, as it proposes some kind of framework or ontology within which future ideas may be expressed (as rendered by certain automated processes). *Metaleptic* markup, while also declarative, is devoted ostensibly to recording,

recapturing or representing some work of signification—perhaps actually some extant artifact, textual or otherwise—that predates it or otherwise precedes it in authority.

The two very different sets of goals and requirements defined by proleptic and metaleptic encodings can easily conflict or become confused. Indeed, the kind a project adopts makes a difference in its ability to meet its goals. (For example, this distinction is a big reason why a tag set designed to support new writing will tend to be unsuitable for textual scholarship, though again, both may be descriptive.) And interestingly, though both *Orlando* and the *Onomasticon* are proposed to us as examples of deep markup, each serves as a fine example of an opposite extreme.

Orlando is proleptic, by design, in just about every respect. It may have a metaleptic aspect to the extent that its design tries to capture or formalize a pre-existing theory of criticism, but even in this aspect it is confessedly experimental. Likewise, it is based on TEI, which is largely metaleptic, but it tends to evade the more particularly descriptive elements in TEI for the more generic ones. (An example of *Orlando* markup will be examined to establish this.) In contrast the overt or ostensible markup of the Onomasticon is all metaleptic, for the sole purpose of describing something that has existed long before it: Ovid's poem. (An example of the Onomasticon will also be considered.)

How does depth figure into these considerations? The two projects look in opposite directions, forward and back. Hence it may seem puzzling how either or both proleptic and metaleptic strategies can, in practice, achieve some kind of critical mass from which “new kinds of knowledge about their subjects” (as I put it above) may be possible.

Willard provides a hint by suggesting that depth has something to do with the readiness of markup structures not merely to support ordinary operations of formatting or retrieval, but also to support some kind of higher-order operations or functions, which take the markup structures as “primitives” and which, at that higher level, can be used to discern abstract categories not directly expressed in the markup at all. The example he proposes is personification, which, he says, “is an example of a deep (or at least deeper) kind because the object of study is not simply declared as a primitive but may be computed from declarable elements.” Like shallow markup, deep markup consists of declarative statements. But the significance of these declarations is not merely in the simple predications they support, but rather in phenomena or assertions that can be discerned only from the appearance of these statements in combination. The *Onomasticon* example will be examined to see how this would work. The failure of its current syntax to support the kind of attribution proposed will be discussed, along with possible alternate syntaxes. The feasibility of what Willard plans will be argued.

In doing this, of course Willard is going far beyond merely a description of Ovid. It may be wondered, accordingly, if the markup scheme goes beyond its originally metaleptic project. I would suggest not: that although the markup is there to support processing (and indeed, very novel forms of processing), and so is forward-looking, nonetheless the processing envisioned is given, once again, for purposes of reflecting back on the text under study—or failing that, back on the markup protocol itself as a scholarly instrument. A truly proleptic markup application would turn these methods not to illuminating Ovid but to creating new meanings altogether. The rationale of the scheme, however, seems to be to reflect a theory of a text that can be tested against actual texts.

In this respect, the markup of the *Metamorphoses* (or potentially of any text that proves receptive to the proposed “grammar of personification”) will certainly “create new knowledge”, even verge, it seems, on prolepsis in providing a framework or grammar for an expression of some kind. Yet it shouldn't be supposed that metalepsis is just the saying over again of old things: quite the contrary; no less than prolepsis, metalepsis is an occasion of new signification, the creation of new knowledge. Indeed, while metalepsis looks back gesturally—predicating something about the world or of something in it—nonetheless it works to change the meaning of the thing it looks back at, by refiguring it and recontextualizing it. (This is characteristic of metalepsis in general as a trope, even irrespective of the way in which, in these cases, these new significations will be mediated or midwived by the operations of the machine.)

And here we return to *Orlando*. “Depth”, it seems, is something we can have in either metaleptic or proleptic modes, for it appears that in their own way, *Orlando*'s rich structures may make (quite beyond their simple applications for organization, search and retrieval) an analogous kind of depth to that of the Onomasticon. In other words, what if we accept Willard's definition, and wonder whether a proleptic effort like *Orlando* can be similarly deep? What kinds of primitives, in what kinds of combinations, can be discerned in *Orlando*'s markup? To what extent are these combinations reflective of various authors’ differing perspectives and interests? To whatever extent *Orlando*'s markup realizes its promise of access broken down into meaningful and useful categories of argument and commentary, how will these combinations and cross-sections be suggestive of new insights and perspectives?

It may take time to answer these questions, but the conditions are certainly right: the tag set is rich (TEI-based) and its extensions (particular kinds of classification of content) suggest several applications. *Orlando* plans next to deploy a pilot delivery system, which is due to exploit its markup structures in novel

ways, in user interfaces, navigation, and access: how will the project take advantage of the positive feedback loops thus created, as not only readers and users, but authors, get to see more readily the shape of larger designs, and hence the expressions, consequences, and uses of their tagging? It may be that a common practice in *Orlando* tagging emerges that characterizes and reflects a whole approach to literary history. These are very intriguing prospects.

REFERENCES

Bloom, Harold. 1982. *The Breaking of the Vessels*. The Wellek Library lectures at the University of California, Davis. Frank Lentricchia, Series Ed. Chicago: University of Chicago Press.

Caton, Paul. 2001. *Markup's Current Imbalance*. *Markup Languages: theory and practice* 3.1: 1–13. See the abstract and excerpt in *Cover Pages* **http://xml.coverpages.org/mltpTOC31.html#MLTP-31caton**.

Cournane, Mavis. 1997. *The Application of SGML/TEI to the Processing of Complex Multilingual Historical Texts*. Doctoral Dissertation, University College, Cork. Cork, Ireland.

Cover, Robin. "Conceptual Modelling and Markup Languages". *Cover Pages* **http://xml.coverpages.org/conceptualModelling.html**.

Sperberg-McQueen, C.M., and Lou Burnard, eds. 1997. "A Gentle Introduction to SGML". In *Guidelines for Electronic Text Encoding and Interchange*. 1994, repr. 1997. Chicago, Oxford: Text Encoding Initiative. pp. 13-36. Available online at **http://www.tei-c.org/Vault/GL/P3/SG.htm**

Hollander, John. 1981. *The Figure of Echo*. Berkeley, CA: University of California Press.

Quin, Liam. November 1996. "Suggestive Markup: Explicit Relationships in Descriptive and Prescriptive DTDs". *SGML'96 Conference Proceedings*. Graphic Communications Association.

Renear, Allen. 2000. "The Descriptive/Procedural Distinction is Flawed". Markup Languages: Theory and Practice 2.4: 411–20.

Goldfarb, Charles F. 1990. *The SGML Handbook*. Oxford: Clarendon Press. Annex A. Adapted from Charles F. Goldfarb, *A Generalized Approach to Document Markup*, in *SIGPLAN Notices*, June 1981.

Sperberg-McQueen, C.M., Claus Huitfeldt, and Allen Renear. 2000. "Meaning and Interpretation of Markup". *Markup Languages: Theory & Practice* 2.3: 215–234. See **http://www.w3.org/People/cmsmcq/2000/mim.html**

Creating a Virtual Center as an International Web-Based Interactive Infrastructure for Research and Teaching in the Language Sciences: A new Research and Library collaboration.

MARÍA BLUME
Cornell University
mb48@cornell.edu

ELAINE WESTBROOKS
Cornell University
elw25@cornell.edu

CLIFF CRAWFORD
Cornell University
cjc26@cornell.edu

JAMES GAIR
Cornell University
jwg2@cornell.edu

TINA OGDEN
The Ogden Consulting Group
tina@tinaogden.com

BARBARA LUST
Cornell University
bcl4@cornell.edu

We describe here the web-based Virtual Center for the Study of Language Acquisition (VCSLA) now under development at Cornell University as a close collaboration between the Cornell Language Acquisition Laboratory (CLAL), Cornell's Albert Mann Library and a set of national and international partners.

The purpose of this center is to foster and facilitate active and continuing interactive research on shared data involving many different languages. Taking advantage of the potentialities of the web, we bring together in a truly interactive way the expertise of researchers at CLAL and that of others at a number of scattered institutions, with the experience and capabilities of the library in information technology applied to storage of and access to shared data.

This first stage of development, financed by an NSF Planning Grant, has a number of crucial features, described here, that in assembly make it an innovative creation that may also serve, as both an ongoing enterprise in language acquisition research and a model for other fields as well.

1. While centered in CLAL, the VCSLA involves the active participation of national and international researchers at other institutions, incorporating a multi-directional flow of information and a perpetual linkage to the library.
2. It incorporates the Virtual Linguistic Lab (VLL), a new web-based interface for data transcription, analysis and access. This interface serves as a common but flexible framework for data entry and access in an interactive manner. This is well under development, and designed to be applicable to both research and teaching.
3. The enterprise embodied is cross-linguistic, including data and research on a widely spread set of languages that will be constantly augmented. VLL furnishes a detailed framework, assuring cross-language comparability, and facilitating both entry and access. It will include the considerable resources amassed in CLAL, and that provided by the participants physically

located elsewhere. Thus it will necessarily be under constant revision as information flows in both directions.

4. The presence of the Mann Library in VCSLA is a special and unique element of its design. Mann Library brings its considerable expertise in the areas of data preservation, data archiving, and metadata management into force on the goals of the enterprise. This effort is consistent with the Cornell Library's long standing active commitment to outreach activities that more firmly engage the library into faculty research and instruction in non-traditional ways, including several of its current digital and Virtual Library initiatives (See for example, "Reinventing the Humanities: Cornell Librarians and Faculty Members Create Electronic Collaborations", *Cornell Library READ!*, Winter 2002).
5. The cooperative effort with the library is crucial to the functioning of VCSLA. Among other things, it links the VCSLA to the Open Language Archives Community (OLAC), an international partnership of institutions and individuals who are creating a worldwide virtual library of language resources, by agreeing to use The OLAC Metadata Set (OLACMS), a set of metadata elements for describing language resources. This component thus extends the informational and accessibility resources of the VCSLA.The OLACMS Standard uses XML to represent metadata descriptions, and metadata librarians at Mann are engaged in making the Cornell Language Acquisition Lab and associated VCSLA a metadata provider as well as a service provider within OLAC. That is, the library has built the infrastructure that allows the CLAL to share its metadata with other OLAC participants and it also has provided the interface that allows CLAL to harvest the metadata from other OLAC participants.
6. There are in existence other forms of electronic data sharing for child language research, notably CHILDES (Carnegie Mellon). VCSLA differs from and is complementary to these in both scope and design, most notably in the incorporation of the VLL format for comparability of data entry and access, and in its essential interactive component. That is, it does not represent simply a databank that allows researchers to access and enter data, but a facility that allows interactive and cooperative communication and research by active scholars who are physically separated by location. It is also designed to be a useful element in the training of students, especially potential researchers in language acquisition Also, as an International Web-Based Interactive Infrastructure, it is designed to incorporate new participants and laboratories worldwide, including areas and countries where resources for training and research in the field are limited or nonexistent.

In our presentation we will sketch each of the above components of the VCSLA and its progress to date. We present concrete examples of its applications to both research and teaching in the language sciences.

Chinese Collections in Museums on the Web: Current Status, Problems, and Future

---

HSIN-LIANG CHEN
Graduate School of Library and Information Science, UT-Austin
chen@gslis.uexas.edu

INTRODUCTION

With the development of computing technology, many Chinese museums and museums with significant Chinese collections have digitized and provided images of their collections on Web sites. These online resources offer users around the world access to valuable treasures to learn about Chinese culture. However, there are obstacles that must be overcome to achieve the goals of promoting the Chinese heritage and educating new generations.

The purpose of this project is to study how museum practitioners use current image indexing practices and services to retrieve the images of the Chinese collections. Several issues, including image needs, information-seeking strategies, information queries, search functions, display formats and human-computer interaction are examined in this study.

This paper focuses specifically on the current practices of image management. The following questions are addressed:

- What kind of images do the museums index?

- How do the museums index their image collections?
- What kind of indexing tools do the museums use?

BACKGROUND OF THE PROBLEM

CURRENT IMAGE MANAGEMENT AT MUSEUMS AND ART LIBRARIES

Graham (1999) surveyed 60 art libraries in the U.K. The survey included the important issues of image collections, cataloging and indexing practices, content-based image retrieval (CBIR) systems, and the use of images. Graham's study reports on the current management of image collections and techniques for image and video retrieval in the U.K. Eakins and Graham (1999) study the current state of the art in CBIR systems within the U.K. and submit several suggestions to U.K. governmental agencies, users and managers of image collections, and CBIR software developers. The Visual Image User Study (VIUS) project at Penn State University is conducting an extensive and systematic assessment of its needs for digital image delivery (Pisciotta, et al, 2001). The VIUS project is working to develop digital picture libraries to serve new uses of digital images for teaching and research.

IMPORTANCE OF INFORMATION NEEDS AND INFORMATION-SEEKING BEHAVIOR TO SYSTEM DESIGN

Stephenson (1999) examined several cultural heritage image databases and identified key issues for future improvements. She pointed out that, in addition to technological challenges, the areas of audience, user behavior, and use should be addressed as well. According to Stephenson, museums should examine:

- For whom are museums building an image database?
- What image databases are museums building?
- What are those users' purposes in using the image database?
- What functionality do the users need to use the database?

These questions reflect a user-centered design philosophy which assists database designers understand their users thoroughly.

Defining the primary user group

One of the major goals of an online image database is to break the barriers of time differences, geographical locations, and limited physical access to materials. This brings up a critical question: who is the user? The online database is likely to serve a diverse group of users: local and remote users, experienced and naïve users, and existing and new users. Different user services are required for this diverse group of users. Museums should examine the institutional goals and purposes of their online image databases to determine the priority of their missions.

Understanding anticipated users

After determining their primary user group, the museums should study their users. The following questions should be considered:

- Goals of designing an online image database
- Differences between existing access to the image collection and the use of the new online image database
- Anticipated users' behavior in different locations
- Selection of surrogates for image indexing and retrieval
- Information architecture of the online image database

Supporting discovery and retrieval

To use the online image database successfully, the users should have knowledge of information retrieval in general, subject areas, and search systems. On the other hand, the museums should investigate the following factors:

- Indexing standards such as metadata
- Query analysis
- Effectiveness and efficiency of indexing tools and methods
- Multi-dimensional indexing and retrieval methods
- Users' capabilities
- User support and training

Supporting functionality

New functions may be created to facilitate users' search strategies. To achieve such goals, studies on interface design, human-computer interaction, and users' information-seeking behavior should be conducted. Different tools may be required by special users and environments.

Other challenges

In addition to the above key issues, the museums also face several challenges:

- lack of communication among museums;
- lack of indexing standards and tools; and

- lack of translation standards of Chinese into western languages.

Chinese museums and museums with significant Chinese collections should form a consortium to establish communication and to develop collaboration. Many western museums have begun those efforts. The Art Museum Image Consortium (AMICO) is one of those consortiums, but its image collections only have about 6,000 works from Asian cultures (AMICO, 2001). The lack of indexing standards and tools is the same challenge for all museums. Most museums either develop their own indexing standards and tools or do not have adequate professional personnel to manage their image collections (CLIR, 1999; Graham, 1999). Regarding translation standards, although American libraries started using Pinyin as the standard romanization scheme for Chinese characters on October 1, 2000 (RLG, 2000), many museums may not be aware of this change and may still use the Wade-Giles system. These challenges are important to the development of image collections.

METHODOLOGY

PARTICIPANTS

Six museums were selected for this study based on the size and diversity of their Chinese collections or their image management. The six participating museums were in the states of California, Illinois, Massachusetts, New York, Ohio, and Washington, D.C.

PROCEDURE

Pre-visit questionnaire

A set of self-administered questionnaires was used to collect librarians' views on cataloging/indexing practices, the functions of new image management systems, and the use of images. The questionnaires were distributed to librarians before an on-site visit and were collected between January and April 2002.

Follow-up phone interview

After all respondents answered the questionnaires, the investigator examined the questionnaires and conducted phone interviews with the respondents for unclear answers and in-depth information. The investigator identified several key people for observations and interviews when visiting museums.

On-site visit

Based on the knowledge gained from the questionnaires and phone interviews, on-site visits were conducted between June and August 2002. The investigator observed librarians and museum practitioners' image seeking behavior and also interviewed those people for further understanding of their search behavior. The investigator interviewed museum administrators to obtain their expectations for digital image management in the mission of the museum.

RESULTS AND DISCUSSION

According to the questionnaires, observations, and interviews, the investigator reports on the current status, problems, and future of image management:

Current Status: all six museums have been digitizing the Chinese collections. Photographic prints, photographic negatives, and transparencies/slides (35 mm) are the most popular formats. All the museums have used or purchased a computer-based image management system to organize the digital images and related information.

Problems: most museums did not have comprehensive records in the past, so they have spent substantial amounts of their budgets and people-power to establish basic records or re-enter data for the digital images. Their conventional cataloging and indexing practices are not suitable or transferable for the new image management system. Image management systems used by the six museums are not able to accommodate the features of the Chinese collections and their records. Most image management systems are not metadata/XML ready, which means that the expansion of the systems onto the Web may be limited. Each museum has its unique institutional structure, which hinders the workflow of image management and a lack of communication and collaboration exists among museums departments.

Future: the development of the indexing schema is critical to the management of digital images and to the museum practitioners and on-line users. Image management systems should be enhanced with the standards of metadata/XML, etc. for the Web-based environment.

Regarding the image-seeking behavior of museum practitioners, and the administration's expectations of digital images, the investigator will report the findings of these issues in the future.

REFERENCES

Art Museum Image Consortium (AMICO). URL: **http://www.amico.org**. (access September, 2001).

Council on Library and Information Resources. (1999). *Scholarship, instruction, and libraries at the turn of the century: Results from five task forces of learned societies and the Council on Library and Information Resources*. Washington, DC: The Council.

Eakins, J. P., & Graham, M. E. (1999). *Content-based image retrieval: A report to the JISC Technology Application Programme* [Online]: **http://www.unn.ac.uk/iidr/research/cbir/report.html**
Graham, M. E. (1999). *The description and indexing of images: Report of a survey of ARLIS members, 1998/99* [Online]. Available: **http://www.unn.ac.uk/iidr/ARLIS/**
Pisciotta, H., Brisson, R., Ferrin, E., Dooris, M., & Spink, A. (2001). *Penn State Visual Image User Study*. *D-Lib Magazine*, July/August. URL: **http://www.dlib.org/dlib/july01/pisciotta/07pisciotta.html/** (access September 2001).
Research Libraries Group (RLG). (2000). Library of Congress, other US libraries join international community on use of Pinyin. URL: **http://www.rlg.org/pr/pr20000-pinyin.html**. (access September, 2001).
Stephenson, C. (1999). "Recent developments in cultural heritage image databases: directions for user-centered design." *Library-Trends*, 48 (2), 410-437.

New Technologies, New Strategies for Integrating Information and Knowledge: Forced Migration Online

MARILYN DEEGAN
University of Oxford
marilyn.deegan@qeh.ox.ac.uk

HAROLD SHORT
King's College London
harold.short@kcl.ac.uk

The first ten years of the web have largely represented a triumph of interconnectedness over functionality: in the late nineteen eighties and early nineteen nineties information resources and teaching tools were being developed that were highly sophisticated and interactive. The web, for all its benefits of connectivity, actually resulted in a massive downturn in functionality, and we are only now able to recover some of those functions with newer developments. A paradigm example of this loss of functionality is in the Oxford English Dictionary: the first version of this, released in the late 1980s, with a DOS interface, represented a revolution in data access. Version 2, released in 1992 with a Windows interface added little in terms of functionality, but a great deal in terms of access. However, OED Online, released in 1999, gave connectivity and wider access at the cost of a huge loss of the functions that many users had come to rely upon—so much so that many users have never made the transition from CD. This is by no means the fault of the developers, but is a consequence of the platform that we are all now using.

New resources are now being developed for data access and retrieval that take full advantage of the benefits of interconnectedness, while giving us enhanced functionality and also allowing us to integrate complex technologies into an apparently seamless whole. This paper will discuss the development of an advanced Internet resource, Forced Migration Online (**http://www.forcedmigration.org**) that was launched in November 2002.

THE STUDY OF FORCED MIGRATION

Forced migration is defined by the International Association for the Study of Forced Migration as 'a general term that refers to the movements of refugees and internally displaced people (those displaced by conflicts) as well as people displaced by natural or environmental disasters, chemical or nuclear disasters, famine, or development projects'. Forced migration studies are essentially interdisciplinary, drawing from anthropology, history, politics, international law, sociology, psychology, and many other disciplines in the humanities and social sciences. The documentary base of the subject has grown rapidly over the last twenty years, and scholars and practitioners in the field rely for their information and studies upon a diverse body of work: conventional books and journals, but also largely 'grey' (unpublished or semi-published) literature. This grey literature can be difficult to get hold of, as it derives from so many different sources: government agencies, non-governmental organizations, academic sources, etc.

THE DEVELOPMENT OF FORCED MIGRATION ONLINE

The development of Forced Migration Online (FMO) began in 1997 at the Refugee Studies Centre (RSC) at the University of Oxford. The RSC has the world's largest collection of grey literature on forced migration

(some 15,000 items) and the Andrew W Mellon Foundation granted funding for a portion of this to be digitized. In 2000, the Mellon Foundation and the European Union gave further funding for the development of an integrated portal to be developed on forced migration. The project to develop this portal has been led by the RSC, but with technical and content partners from around the world: the FMO team coordinates participants in some 10 institutions and is working with many more than this to develop content further. FMO now contains 100,000 pages of fully searchable grey literature, 30,000 pages of full-text journal materials, a number of specially-commissioned research guides, a web catalogue with c. 700 entries, an organizations database with c. 800 records, and a prototype image database.

CREATING INFORMATION ARCHITECTURES FOR THE DEVELOPMENT OF FMO

As anyone engaged in the development of digital libraries and portals knows only too well, there is no one obvious tool or technology to implement such complex information resources, though many are currently in development. In FMO, there are a number of different technologies underlying the resource: a complexity which is well hidden from the user, for whom access is relatively simple. The full-text documents are presented using Olive Software's Active Paper Archive, which was originally developed for presentation of historic newspaper content on the web, and which has proved an excellent choice for the grey literature and journals. FMO is the first project that has used this product in this way, and the development was a joint research project between the FMO technical teams at the RSC and at the Centre for Computing in the Humanities at King's College London (CCH), and Olive Software. The structured information resources and catalogues are delivered using Esprit Soutron's xdirectory content management system, and various research guides and other documents are created and presented by means of XML/XSLT.

The core challenge is one of integration: integration of a wide variety of information types, drawn from geographically separated repositories capable of providing widely disparate levels of metadata; integration of materials in numerous languages in a variety of scripts; integration of the multiple technologies required to meet the differing information processing and delivery functions; integration of academic analysis and advice for practitioners, and of information and knowledge, to meet widely varying user requirements.

Delivering a coherent and integrated resource in a seamless way is a non-trivial technical challenge. It involves visual design, architectural design, development of DTDs and style sheets, and the implementation of leading edge (and therefore constantly evolving) products. Managing the input from so many people and places around the world represents another layer of challenge.

This paper assesses the problems of developing and integrating these complex technologies into a hybrid information environment, in particular looking a metadata, cataloguing, preservation, delivery and accessibility issues. It reports on the solutions and partial solutions developed so far, and assesses the extent to which the solutions fall short of the ideal. It also discusses a range of further challenges that the FMO team is now tackling: automatic metadata extraction from journal cross-searching tools for different products using advanced APIs; using focused crawlers as aids to cataloguing; automatic categorization of documents for the creation of regional and topical browse sets. The progress in meeting these challenges will be discussed in the context of the more established work in the project.

The paper also places the project in a wider context: of past and current work in the development and delivery of scholarly resources, including a number of projects at IATH in Virginia, in CCH at King's College London, and elsewhere; and of digital library research and development, including projects such as the 'hybrid library' projects funded by the UK government (including the Malibu project), the DSpace initiative of MIT and others, and the Mellon-funded FEDORA project based at Virginia and Cornell.

Data or Document? Migration of Descriptive Metadata for Medieval and Renaissance Manuscripts Between Data-Centric and Document-Centric Models: A Case Study

ELIZABETH J. SHAW
Aziza Technology Associates
ejshaw@azizatech.com

This paper will discuss a case study of the dual direction migration of metadata describing medieval and renaissance manuscript collections between the Digital Scriptorium database schema and an XML DTD developed to encode extant manuscript descriptions. This discussion highlights the challenges and tensions that exist in trying to merge data that is captured in different data models, roughly categorized as data-centric and document-centric. The observations garnered from this study may have broader implications for other efforts to meld distinct descriptive models for unitary searching and resource discovery.

CONTEXT

The author was hired by Digital Scriptorium to: 1)provide XML —> HTML XSLT transformations of an electronic version of the Huntington Library "Guide to Medieval and Renaissance Manuscripts" which has been encoded in the TEI Medieval Manuscripts Description Work Group (TEI MMSS) draft DTD , 2) migrate data from the existing Digital Scriptorium database to the the draft DTD and 3) migrate the TEI-MMSS encoded "Guide" to the Digital Scriptorium's existing database schema. In the context of this work, many interesting challenges have come to light.

The Digital Scriptorium database schema and the TEI MMSS draft DTD are closely allied since some of the same people worked on both data models. However, significant differences exist in nomenclature and structuring of the data, both because the TEI MMSS DTD is an extension of TEI and because the purpose of the TEI MMSS DTD is to allow for the capture of extant descriptive information. The database schema, created prior to the DTD is a more constrained data model and uses slightly different elements and nomenclature for its basis.

DESCRIPTIVE TRADITIONS

Although the library community has largely settled on a particular record-based model for descriptive information (the MARC record and AACR2) about printed works, serials and selected media, many communities of practice that have need of description to facilitate resource discovery and scholarship have no such common standard. In many cases, a narrative form of description has evolved in an attempt to capture the particular descriptive needs of the curators and users of the materials. Descriptive practice in disciplines such as manuscript collection or archives has varied greatly across nations, institutions and within individual institutions over time. This evolution has largely been dependent on the interests, energies and capabilities of individual curators.

Furthermore, descriptive practice varies greatly between communities of practice because of the particular characteristics of the disciplines that they support. Both the elements of description fundamental to supporting work and the nature of their expression may be unique. That these descriptive needs vary greatly has become more obvious as efforts to develop single metadata standards or standards that easily map across disciplines have had limited success.

Nonetheless, these communities often face similar problems as they attempt to migrate their formerly print based description to an electronic form that can be shared across institutions. Each community must revalidate what aspect of description is important to the work of the community, 2) define a common way of expressing information about those important things and 3) develop a model that accurately captures that information. The work must be done in an environment with no uniform extant practice and differing philosophies about the nature, scope and role of description in the discipline. This paper will focus on two approaches to capturing significant descriptive data within the same discipline and examine the challenges in merging the two.

DATA-CENTRIC VS. DOCUMENT CENTRIC DATA MODELS

The numerous schemes for descriptive metadata that have emerged since the advent of the web vary greatly in their scope, complexity and underlying assumptions about the nature of descriptive practice. Ronald Bourret's article XML and Databases and his associated materials provide an interesting framework from which to view these different approaches to capturing descriptive metadata. The emerging practices might be categorized into two distinct models:

- **Data-Centric models:** These record-like models capture discrete data elements in a structure that is well represented in a flat or relational model (ie. Dublin Core, MARC). Often these flat or relational representations atomize data at the most discrete levels at which one might manipulate it. Some require or enforce the normalization of the data, institute data typing and constrain the syntactical expression in which it is represented. The Digital Scriptorium database schema follows this record-like model, though without some of the rigor of syntax that is required in a MARC record.
- **Document-Centric models:** These models capture more narrative or discursive, often complex, hierarchical descriptions. Often, they allow considerable flexibility in what data elements are required, the form in which data is expressed, and rarely (especially prior to the advent of XML Schema) utilize data typing to constrain the expression of information.

These document-centric models may further be categorized as those that provide a representation of an idealized model of descriptive practice and those that represent extant descriptive practice. Those that attempt to capture an electronic edition of an extant description are often the least rigorous in their requirements for data elements and normalization because of the need to accommodate heterogenous practice. The tension between representing an idealized model of a particular document structure—either to facilitate machine processing of the data or to enforce practice in a community—and providing a model that captures existing representations has been discussed extensively by authors such as Piez and in efforts to develop data models such as the TEI dictionaries DTD .

CAPTURING EXTANT DESCRIPTION

Although document-like descriptions from various institutions and time frames seem similar in format they often contain distinctly different elements of description in distinct orders and with variant nomenclature. This has presented challenges to the developers of DTDs and schemas that attempt to capture descriptive metadata in these communities of practice. DTDs such as *Encoded Archival Description* (EAD) and the TEI MMSS draft DTD are explicitly designed to accommodate a variety of extant descriptive practice rather than constrain practice to the schema developers' prescribed notions of practice.

The extant documents that are encoded using these models often differ greatly from their record-like cousins. Narrative and complex, they utilize indirect reference, inference from context (the description of a manuscript written in French may never explicitly state the language in which it is written, assuming that the reader can discern from the quoted text imbedded in the description), inference from relationships to siblings or parent elements, and assumptions about relationships of the parts of the description that are not explicit. Although a human reader can infer these relationships, the lack of explicitness makes it more difficult to capture characteristics of the described object in machine processable ways. To capture this information explicitly requires modifying the extant description. Data-centric models tend to capture this information in explicit ways from the start.

MIGRATING DATA

Many practitioners and users of existing descriptive material find that even when description is reworked by knowledgeable humans in order to import it to a data-centric metadata representation, the records fail to capture significant information that is embedded in the narrative of these extant descriptions. In many cases, the original cataloger has specialized knowledge of either the collection being described or of the field of study. References to other significant or related work, connections between objects in the collections are often lost in their transformation to data-centric descriptions.

Tensions between models for capturing descriptive metadata are often amplified at the point at which migration occurs. Automated efforts to migrate encoded extant description have met with mixed success. Chris Prom's migration of EAD encoded descriptions to Dublin Core RDF records for importation into a cultural heritage OAI database points to some of the problems inherent in automated migration of extant data. Lundberg discusses the implications for information retrieval of data encoded in XML.

THE DIGITAL SCRIPTORIUM: CASE STUDY

This paper reports on the outcomes of the effort to automate the migration of data captured in two distinct models and the resulting implications for resource discovery. The paper describes:

- the differing data models,
- the process by which the migration occurred,
- the challenges faced both in migration from relational database to XML and in migration from XML to a relational database model,
- the development of a relational database model that attempts to capture both the XML encoded manuscript descriptions as well as the data in the existing database
- the trade-offs in these various representations and
- the implications for information display and resource discovery.

By highlighting points of ambiguity, strengths and weaknesses of the different approaches to metadata capture and challenges in merging the data into the same repository, it is hoped that this study will inform the work of others who face similar challenges in their efforts to provide descriptive metadata to their user communities.

REFERENCES

Ronald Bourret, XML and Databases. **http://www.rpbourret.com/xml/XMLAndDatabases.htm**

Digital Scriptorum. **http://sunsite.berkeley.edu/Scriptorium/**

Digital Scriptorium Database Schema. **http://sunsite.berkeley.edu/Scriptorium/datadic5.html**

Encoded Archival Description. **http://www.loc.gov/ead/**

Lundberg, Sigfrid, Excursions along the border between metadata for resource discovery and for resource description. **http://laurentius.lub.lu.se/search/presentation/laurentius.pdf**

Wendell Piez, "Beyond the 'descriptive vs. procedural' distinction." Extreme Markup Languages. (2001) Montreal, Canada. pp 197–214.

Christopher Prom, "Does EAD play Well with Other Metadata Standards?: Searching and Retrieving EAD using the OAI Protocols." *Journal of Archival Organization*. (2003) No. 1 Issue 3 (Forthcoming).

TEI dictionaries DTD. **http://www.tei-c.org/P4X/DI.htm**

TEI Medieval Manuscripts Description Work Group (TEI MMSS) draft DTD **http://www.stg.brown.edu/~tei-mmss/**

The Charles W. Cushman Collection: Enhancing Visual Resource Discovery Through Descriptive Metadata Based on Subjective Image Analysis

LINDA CANTARA
Indiana University Libraries
lcantara@indiana.edu

The Charles W. Cushman Collection[1] comprises nearly 18,000 Kodachrome slides, captured from 1938 to 1969 by Charles Weever Cushman, an accomplished amateur photographer, world traveler, and pioneer in the use of color photography. Cushman bequeathed the collection to his alma mater, Indiana University, on his death in 1972, yet its existence was virtually unknown until late 1999 when a university archivist rediscovered the suitcases in which the slides were stored, along with Cushman's notebooks containing identifying information and dates for each slide. The images visually document in color the vernacular history of people, places, and events that have previously been seen only or primarily in black and white. To provide online access to the collection, a project team at Indiana—including staff of the Digital Library Program, the University Archives, and the University Libraries— has digitized the slides and has designed a relational database to document administrative, technical, and descriptive metadata about the images. Transcriptions of Cushman's notebook annotations and corresponding documentation on slide mounts will be keyword searchable. In addition, we are creating descriptive metadata for each image by assigning terms for subject content, genre, physical characteristics, and geographic location. The terms are selected from standard controlled vocabularies, primarily the Library of Congress Thesauri for Graphic Materials (TGM I and TGM II), but also Library of Congress Authorities Files (LCAF) and the Getty Thesaurus of Geographic Names (TGN). When all the slides have been cataloged, the database will be mapped to an XML format—most likely Encoded Archival Description (EAD) but possibly Metadata Object Description Schema (MODS)—and will be searchable over the Web.

The use of controlled vocabularies in the descriptive metadata of a large image collection has the potential to optimize visual resource discovery, but many issues must be addressed and resolved to ensure the expense of assigning the terms truly results in improved image retrieval:

- The Library of Congress Thesaurus for Graphic Materials I: Subject Terms (TGM I) and Thesaurus for Graphic Materials II: Genre and Physical Characteristic Terms (TGM II) include topic headings developed by the Library of Congress Prints and Photographs Division over the past fifty years to catalog graphic materials. In adherence with ANSI/NISO Z39.19-1994,[2] terms are expressed in natural language word order following American English spelling conventions. Although new terms are added regularly based on proposals submitted by catalogers (Alexander), these thesauri frequently lack the specificity desirable for describing individual images. In addition, regardless of the domain knowledge and cataloging expertise of personnel performing image analysis to assign subject headings, inconsistencies are inevitable. Professional catalogers of traditional media readily acknowledge that subject analysis of printed materials may vary from cataloger to cataloger as well as from day to day. In this paper, I will discuss how we have attempted to approximate consistency of image analysis by multiple catalogers, the quality control procedures we have implemented to further ensure uniformity of subject term assignment, and the "work- arounds" we have devised to deal with encountered inadequacies in the selected controlled vocabularies.
- The Digital Library Federation, the Getty Grant Program, and the Andrew W. Mellon Foundation in cooperation with Rice University are currently sponsoring a Visual Resources Association (VRA) review and evaluation of existing data content standards and current practice in order to compile a guide to good practice for describing cultural objects and images. The forthcoming Cataloguing Cultural Objects: A Guide to Describing Cultural Objects and their Images—the CCO Guide—will provide recommendations and examples for using data value standards tools and building authority records.[3] Used in concert with XML data structure and communication standards, the CCO Guide will facilitate consistent creation of descriptive metadata for visual resources, enabling interactive and complex search options. Paradoxically, Open Archive Initiative Metadata Harvester (OAI-PMH) service providers require data providers map rich metadata to unqualified Dublin Core, while content aggregators like the Research Libraries Group (RLG) Cultural Materials service must currently treat all controlled vocabulary terms as keywords.[4] Although neither OAI service providers nor RLG preclude the use of multiple or complex metadata formats, content creators nevertheless face a quandary: why commit extensive and expensive personnel hours to create increasingly rich metadata when interoperable and federated access is possible when implementing simpler, less costly, metadata standards?[5]
- The foremost concern when designing and implementing an image retrieval system should be the searching needs of the user community (Sundt). Art historians, for example, may be trained in the use of one or more controlled vocabularies specific to art images,[6] but users of vernacular photography collections may have limited experience with any controlled vocabulary. Since lack of familiarity with the underlying vocabulary tools may actually diminish rather than enhance user interaction, the search interface should ideally facilitate direct mining of the encoded metadata, directing the user to broader, narrower, related and cross-referenced subject terms. This paper will conclude with a review of our usability test findings in respect to a prototype interface that integrates TGM I and II terms into the search facilities for the Cushman Collection.

NOTES

[1.] This project is funded by an IMLS National Leadership Grant.

[2.] ANSI/NISO Z39.19-1984 is the American National Standards Institute/National Information Standards Organization Guidelines for the Construction, Format and Management of Monolingual Thesauri. NISO recently announced plans to revise these guidelines: see **http://www.niso.org/committees/MT-info.html**.

[3.] See **http://www.vraweb.org/projects.html** and "Project Proposal for Guide to Good Practice: Cataloging Standards for Describing Cultural Objects and Images," Digital Library Federation, January 12, 2001, at **http://www.diglib.org/standards/vrawork.htm**.

[4.] See the Research Libraries Group (RLG) Cultural Materials site at **http://culturalmaterials.rlg.org/cmiprod/workspace.jsp**. In this regard, the RLG states, "Ultimately, we hope to add powerful 'assisted searching' tools, to help users navigate across different vocabularies and subject schemes, but in the meantime we are simply using whatever subject terms are provided in the source

data." (See **http://www.rlg.org/culturalres/descguide.html#content**).

[5.] I am, of course, playing the devil's advocate here.

[6.] For example, Getty Art & Architecture Thesaurus (AAT) at **http://www.getty.edu/research/tools/vocabulary/aat/**, ICONCLASS at **http://www.iconclass.nl/**, and Categories for the Description of Works of Art at **http://www.getty.edu/research/institute/standards/cdwa/index.html**.

REFERENCES

Alexander, Arden and Tracy Meehleib. "The Thesaurus for Graphic Materials: Its History, Use, and Future." Cataloging & Classification Quarterly 31:3/4 (2001), 189–211.

Encoded Archival Description <**http://www.loc.gov/ead/**>.

Getty Thesaurus of Geographic Names (TGN) <**http://www.getty.edu/research/tools/vocabulary/tgn/**>.

Guidelines for Implementing DC in XML <**http://dublincore.org/documents/dc-xml-guidelines/**>.

Library of Congress Authorities <**http://authorities.loc.gov/**>.

Library of Congress Thesauri of Graphic Materials: TGM I (Subjects) <**http://www.loc.gov/rr/print/tgm1/**> and TGM II (Genre and Physical Characteristics Terms) <**http://www.loc.gov/rr/print/tgm2/**>.

Metadata Object Description Schema (MODS) <**http://www.loc.gov/standards/mods/**>.

Open Archives Initiative FAQ <**http://www.openarchives.org/documents/FAQ.html**>.

Sundt, Christine L. "The Image User and the Search for Images." Introduction to Art Image Access: Issues, Tools, Standards, Strategies. Murtha Baca, ed. Los Angeles: Getty Research Institute, 2002. 67–85.

Preservation of the New Media Arts

MEGAN WINGET
UNC - Chapel Hill
winget@email.unc.edu

It has become a commonplace to declare that we are living through one of the great epochs of human discovery. "The digital revolution." The Internet as a "fundamental and extensive force of change." With the advent of the Internet as a mode of communication and information discovery, we are leading lives, and facing challenges undreamed of in previous generations except as the stuff of science fiction. But what if it's actually true? What if we are living through a period as significant as the Renaissance or the Enlightenment? What if our current circumstances are comparable to the Depression? Or the Post–war era? How about the counter–culture revolution of the 60s? Each of these eras, great and small, are marked, not only by their social upheavals, but also by their great strides in artistic representation. The Renaissance had Michelangelo and Leonardo; the Enlightenment had David and Ingres. In the modern era, we have Frank Lloyd Wright, Piet Mondrian, Mark Rothko, Andy Warhol, Georgia O'Keeffe, le Corbusier, Picasso, Matisse, and the list goes on and on. Great artists, whose work helps us make sense of the past. Of course, we also have meaningful contemporary art—but there's a problem. Whereas we can still look at Michelangelo's paintings and sculptures, and in many cases, can look at preliminary studies and drawings—we're running the risk of loosing contemporary art as soon as ten years after its creation.

The new media art community has recently come to the realization that the objects and ideas of their most compelling thinkers and artists run the risk of disappearing forever, because these objects are often in digital or other variable formats, which tend to rapidly become, at best, inaccessible; and at worst, irretrievably lost. Unless there is a systematic and persistent exploration of preservation and archival procedures for these significant cultural objects, they will, in all likelihood, not be accessible for future generations of artists, scholars, or the general public. One of the new media community's most important advances towards this goal of systematic investigation is the development of the "Archiving the Avant-Garde" project, which brings together new media venues, curators, and artists to devise possible solutions to this very difficult problem of the disappearing objects of the avant-garde.

There are three distinct challenges that the new media art community has to confront. They must first assess the nature and goals of new media art in general. It's very difficult to preserve a nebulous "something," especially if that "something" is a complex series of digital objects with variable and intricate relationships between parts and the whole, which also happens to comment and feed off of a culture in which the conservator is currently living. Unless the preserving agency has a convincing vision of what is intrinsically important to maintain, the resulting objects run the risk of being inauthentic and/or presenting an inaccurate

indication of the artist's intentions.

It would also be prudent for this community to consider the current state of art conservation and preservation in related fields. It's important to recognize the challenges and philosophical conflicts that art conservators of the last half century have faced. No matter how cutting edge and exciting these new media artifacts are, they are not being created in a vacuum, and conservators have faced similar (although not digital) problems for hundreds of years. I will specifically review methods of traditional art conservation, and how those professionals are dealing with less stable forms like conceptual and performance art.

Finally, the community should recognize and reflect on the history of access and preservation from an archival point of view. The project is, after all, called "Archiving the Avant-Garde," and there are some common tropes and methods within the archival community, particularly the ideas of diplomatics and intrinsic value, for example, that the new media art community might find helpful.

Information technology professionals and humanist scholars have a responsibility to keep the past alive. The information scientists provide the infrastructure and methodologies, and the humanists provide the interpretive structure and deep understanding from which future generations will gain insight into our collective consciousness. Unless we begin working together, building systems that grow naturally out of the scholarly tradition, rather than expecting the scholarly tradition to twist itself into knots trying to fit within the current Information Infrastructure, we will never be able to transcend the prosaic world of subject headings and metadata elements. Information artists and the new media art organizations are accomplishing, in the work itself, this very feat of bringing together traditional forms of expression with new technologies. We should take a lesson from them. We need to find a way to save and preserve the richness of this experience for future generations.

Linguistic Issues in the Text-Encoding of Sanskrit

PETER SCHARF
Brown University
Scharf@brown.edu

MALCOLM HYMAN
Harvard University
mhyman@fas.harvard.edu

VENU GOVINDRAJA
SUNY Buffalo
govind@cedar.buffalo.edu

RALPH BUNKER
Maharishi University of Management
rbunker@lisco.com

This panel will study the principles upon which text-encoding schemes for Sanskrit are based and the uses to which they are suited. The issues under investigation range from character encoding to more complex markup such as the indication of word boundaries. The choices made in the design of a text-encoding scheme have important ramifications for text-processing functions such as data entry, efficiency of encoding, linguistic processing (e.g., morphological and phonological analysis), and rendering. Speakers will present solutions that utilize new technologies such as XML and OpenType. Although the discussion in this panel will focus on Sanskrit, many issues are relevant also to other languages and writing systems.

Character-encoding schemes may range from the purely sound-based to the purely graphic. In a sound-based scheme, the basic unit of analysis is the phone (speech sound). A sound-based scheme may either take the phone as an atomic unit or decompose the phone into a bundle of phonetic features. Taking the phone as the atomic unit, a purely phonemic system includes only the distinctive sounds (phonemes) of the language in its inventory. Most systems which take the phone as the atomic unit, however, extend their inventory to include at least some contextually conditioned sounds, and most sound-based systems mix phonemic and phonetic principles. In graphic schemes, the basic unit is the graph (written shape). Graphic schemes may involve three principles. They may take the character as an atomic unit and encode only

graphemes; they may decompose the character into a set of strokes and encode partial glyphs; or they may encode complex glyphs (ligatures).

Sanskrit, the primary culture-bearing language of India, is written in the Devanagari script. This script includes both syllabic and alphabetic features. Consonant graphs imply an inherent short /a/ vowel, unless another vowel is explicitly indicated or the absence of a vowel is indicated by the virama sign. Consonant sequences are written as ligatures; traditional Sanskrit orthography requires glyphs for more than a thousand such sequences. We will survey and categorize current encodings for Sanskrit, including general standards-based schemes for Indic languages, and specialized schemes used by Indologists for Sanskrit both in Devanagari and in Roman transliteration. Moreover, we will examine the principles upon which the various schemes are based, the applications to which they are best suited, and their potentials and shortcomings.

After clarifying the general issues, Hyman and Scharf will discuss their sound-based encoding scheme. This scheme has been used for representation of a digital library of Sanskrit texts and as the internal encoding for an automatic Sanskrit morphological analyzer. Tools are available that facilitate transliteration and re-encoding into a number of standard formats, including Unicode.

The word-boundary problem has limited the utility of Sanskrit digitized text. Word boundaries are often obscured phonologically in Sanskrit by the replacement of consecutive vowels with a single sound or graphically by the rendering of consecutive consonants with a single ligature. Word indices, concordances, grammatical and lexical analysis, however, require access to word-boundary information. Earlier work in text encoding has addressed this problem in ad hoc ways, typically by using Roman transcription and by undoing sandhi. The former sacrifices flexibility in rendering and the latter destroys prosodic information. An ideal system would provide all the information necessary for both flexible rendering and linguistic analysis. Work currently under way at Brown to create a computational implementation of a lexically based production grammar and parser promises to allow automated processing that can correctly divide digitized Sanskrit text into its component words.

Govindaraju is compiling Devanagari character databases and digitized Hindi lexica and tagged text corpora to serve as an accurate and comprehensive benchmark to test algorithms used in the field of Devanagari OCR research. He is also developing tools for truthing scanned documents. He will discuss techniques for word and line separation, character segmentation, lexically driven and lexicon-free techniques for word recognition, and linguistic tools for post processing.

Bunker surveys the support that OpenType fonts provide for ligatures. Because Unicode separates rendering issues from character-encoding, ligature selection need not be encoded in the text but can be left to the font. Bunker's research aims at developing a database of all ligatures found in printed Devanagari texts and software that allows a user to build a customized OpenType font automatically. Any application that supports OpenType fonts will then be able to render Sanskrit text with the user's choice of ligatures. This solution leverages the emerging OpenType standard to allow high-quality Devanagari typography without cumbersome data-entry conventions or specialized software.

After the four fifteen-minute presentations, the panel will openly discuss cross-linguistic issues, the potential for mutual application of techniques developed for various scripts and languages, and the development of intelligent techniques to advance Sanskrit text-processing.

The Screen or the Window: A Critical Proposal for Reading Computer Representations

MICHELE WHITE
Wellesley College
mwhite@wellesley.edu

The structuring qualities of the computer screen play an important part in how we understand the Internet and other computer-facilitated representations but they are rarely addressed in the critical literature. In this presentation, I consider how the screen and mediation remain largely invisible because the Internet is described as a material setting that users inhabit. I provide a brief study of popular and academic narratives and focus on renderings of the computer and Internet as a window or entrance that appear in telepresence art works, webcams, and advertisements for computer screens. Close visual and textual analysis and critical considerations of photography's referential aspects are employed. These methodologies provide a way to highlight the computer screen and encourage an alternative understanding of the computer and Internet.

The tendency to engage with the computer and Internet as physical and animate has significant

ramifications. Our conceptualization of the medium determines the questions that we can ask and whether the representational aspects of stereotypes can be perceived and critiqued. Judith Mayne has argued that feminist film theorists such as Laura Mulvey have attributed "the polarity of gender, of masculinity versus femininity, to the very structures of pleasure and identification in the classical cinema" (48). Computers and Internet settings employ similar devices. However, the forms of identification produced by computers and Internet settings have an even more consequential effect because users spend significant amounts of time engaging with computers; computers and networks also appear in film, television, and print advertising; dream or trance-like experiences are often part of the engagement; the user's identification with characters and other representations can be intense; and there is an idea that people are alive and their bodies are accessible through the Internet.

Considering the screen rather than the delivered representations is difficult because of the rhetoric about physical and populated Internet settings. For instance, Esther Dyson makes it appear like people live on the Internet rather than use it when she states that "the Net includes all the people, cultures, and communities that live in it" (2). Eduardo Kac uses his telepresence artwork, *Teleporting an Unknown State* (1994/96, 1998, and 2001), to render the "Internet as a life-supporting system" (Kac) and suggests that there is "Birth, growth, and death on the Internet" (Kac a). Such narratives about people inhabiting the setting and live interfaces are connected to early descriptions of the computer and Internet. William Gibson suggests that physical settings and screens will combine in the first sentence to *Neuromancer* when he states that "the sky above the port was the color of television, tuned to a dead channel" (3). Gibson starts *Neuromancer*, which had a significant effect on computer culture, with the screen but he displaces its constructed aspects with live programs, cyber "space," and populated interfaces.

Internet sites continue to provide "welcome" messages, invitations to enter, and depictions of spatial progressions. For instance, the webcam operator Gwen encourages the spectator to "enter the life of a college student" and Cindy coaxes the spectator to "ENTER ... with an open mind." A variety or multi-purpose sites are referred to as "portals" and Gretchen Barbatsis, Michael Fegan, and Kenneth Hansen indicate that spectators "engage the computer screen as a gateway to another place." Window-like effects are employed because they seem to provide views "onto" other terrain, articulate an inside and outside, and suggest that computers and the Internet deliver a continuous spatial landscape. Thomas J. Campanella suggests that "Webcams, the Web's windows on the world, knit the Net to the physical spaces we inhabit. The accompanying illustration by Jack Desrocher supports this conception by depicting the computer monitor as a curtained window that is incorporated into the home setting. Such renderings make representations seem to be part of the lived space of the spectator.

The conception that the computer and Internet are a window or portal, which provides an entrance into other places, is related to societal conceptions of photography. Photographic images are often talked about as if they provided direct access to the thing depicted—the referent—rather than being representations. This cultural conception is conveyed by Susan Sontag when she states that photography is "not only an image (as a painting is an image) an interpretation of the real; it is also a trace, something directly stenciled off the real" (154). Internet sites often use data, texts, and graphics instead of digitized photos and digital imaging technologies to make representations seem real. However, society's usual conception of photography can still provide another example of a technology and cultural form in which the framed aspects of images are often ignored. Histories of photography and other referential media-like film and television-indicate that the physical and animate aspects attributed to the computer and Internet are not unique and that these renderings should be associated with other produced forms. Such photography theorists as Rosalind Krauss, Martha Rosler and John Tagg, also offer some significant literature to employ in engaging with the produced and framed aspects of computer and Internet renderings.

A re-employment of the photographic vocabulary of cropping and rectangular formats can highlight the relationship between individual screen elements, software windows, and monitors. Noting and articulating the repetitive aspects of computer and Internet representations can also indicate that it is cultural conventions rather than people that are conveyed. For instance, Tagg argues that realism "works by the controlled and limited recall of a reservoir of similar 'texts,' by a constant repetition, a constant cross-echoing" (99). Such critical considerations of photography may be more productive in highlighting the produced aspects of computer and Internet settings than in helping society to engage with photography differently. Computers and Internet settings rarely employ traces of material objects in the ways this concept operates in photography. The redeployment of photography theory can also begin to render a new vocabulary for describing and conceptualizing these settings. We should continue to consider other possibilities for speaking, writing, designing, and visually rendering the Internet since these acts produce the setting.

REFERENCES

Barbatsis, Gretchen, Michael Fegan, and Kenneth Hansen. "The Performance of Cyberspace: An Exploration into Computer-Mediated Reality." *Journal of Computer Mediated Communication* 5. 1 (September

1999). 22 Mar. 2002 <**http://www.ascusc.org/jcmc/vol5/issue1/barbatsis.html**>.
Campanella, Thomas J. "Be There Now." salon.com. 7 Aug. 1997. 30 June 2002 <**http://www.salon.com/aug97/21st/cam970807.html**>.
Cindy "Intro." 13 Jan. 2001 <**http://www.summer-web.com**>.
Dyson, Esther. *Introduction. Release 2.0: A Design for Living in the Digital Age*. New York: Broadway Books, 1997.
Gibson, William. *Neuromancer*. New York: Ace Books, 1984.
Gwen. "Gwencam." 19 Oct. 2000 <**http://gwen.webica.com/cam/**>.
Kac, Eduardo. "Teleporting an Unknown State." 26 Jan. 2003 <**http://www.ekac.org/teleptrvl.html**>.
---a. "Teleporting an Unknown State." 26 Jan. 2003 <**http://www.ekac.org/teleporting.htm**>.
Mayne, Judith. "Feminist Film Theory and Criticism." *Multiple Voices in Feminist Film Criticism*. Ed. Diane Carson, Linda Dittmar, and Janice R. Welsch. Minneapolis: U of Minnesota P, 1994.
Sontag, Susan. "The Image World." *On Photography*. New York: Penguin Books, 1977.
Tagg, John. "A Means of Surveillance: The Photograph as Evidence in the Law." *The Burden of Representation: Essays on Photographies and Histories*. Minneapolis: U of Minnesota P, 1993.

Visual or Verbal: Two Approaches to Creating an Immersive Virtual Environment

EUNICE JOHNSTON
North Dakota State University
eunice.johnston@ndsu.nodak.edu

(Co-authored: Eunice Johnston (eunice.johnston@ndsu.nodak.edu), Jeffrey T. Clark, Brian M. Slator, Gary K. Clambey, Aaron Bergstrom, Shawn Fisher, Justin Hawley, James E. Landrum III, David Martinson, J. Liessman Vantine)

One benefit of the development of Immersive Virtual Environments (IVEs) for education is that students in the humanities may now encounter and learn about places and cultures that which are difficult to visit or may no longer exist. As increasingly sophisticated graphical interfaces have become possible, many IVEs have decided to use this technology. Yet some might question whether the highest fidelity interface is always necessary or even desirable for the study of the humanities. This project seeks to answer that question by developing two parallel versions of an IVE that represents the Like-a-Fishhook Village and Fort Berthold area in western North Dakota as they existed in 1858, when people lived there, and in 1954, when it was being excavated. Both versions will permit users to explore the environment freely, but one version creates a complete, three-dimensional visual representation while the other is constructed mostly of text with a few carefully chosen images and sounds.

To understand the story Like-a-Fishhook Village, the last great earth-lodge village in the Great Plains, and Fort Berthold, the nearby fur trading post, is to understand the history of west. The Mandans, who had befriended Lewis and Clark and the Corps of Discovery in 1804–04, and their close neighbors, the Hidatsas and Arikaras, had lived in sedentary earth-lodge communities along the Missouri River for centuries, hunting, raising corn and other crops, and trading extensively with neighboring tribes. Despite a basic similarity of economic and social life, these peoples differed remarkably in language and customs. The unification of these three tribes at Like-a-Fishhook village tells of the significant impact that the coming of European American people had upon Native American cultures. Although they had suffered several epidemics previously, a smallpox epidemic in 1837 reduced the Mandan population to fewer than 200 individuals. The Hidatsa and Arikara populations, although not as severely struck, shrank as well. Ethnohistorical evidence suggests that the first permanent residents of Like-a-Fishhook Village were Hidatsa who arrived in 1845, and that they were joined shortly thereafter by a smaller group of Mandans; the two tribes felt it was beneficial to combine their numbers in order to resist the attacks by Sioux tribes in the area. Although living together, each tribe maintained its own language and culture. P. Chouteau, Jr. and Company, an offshoot of the American Fur Company, quickly established a trading post north of the village, and European Americans were part of the community from then on. Many changes occurred in the next forty years. The fur traders were followed by missionaries, United States government representatives, and soldiers, all people who wanted the natives to change their ways. Fur trading resulted in increased competition for natural resources, which caused more friction than usual among the tribes, and the Arikara, hoping for support in the battles against the Sioux, joined the other two tribes at the village in early 1860s. In the mid 1880s, the United States government

forced the residents to abandon their village life and to take up individual allotments scattered throughout the reservation.

Archeological salvage excavations were carried out at the site in 1950–52 and in 1954 by the State Historical Society of North Dakota, under contracts with the National Park Service and by the River Basin Surveys of the Smithsonian. However, the rising waters from the Garrison Dam and Reservoir project ultimately inundated the site, which now rests about a mile offshore under the waters of Lake Sakakawea (Garrison Reservoir). Although many individuals were involved in that work, the final report was written by G. Hubert Smith and published by the U.S. Department of the Interior (Smith, 1972).

A multidisciplinary team is creating two parallel IVEs based on the same content: both recreate the village site as it appeared in 1954, when it was being excavated, and also in 1858, when it was occupied, and visitors will be able to "time travel" from one period to the other. The 3-D, graphically rich version is modeled on other IVEs that have been created at NDSU for students to learn to do science by conducting simulated experiments and solving authentic (albeit virtual) problems. Science-based systems of this sort have demonstrated statistically significant impact on student learning in controlled studies (McClean et al., 2001). The great strength of this version is the visual stimuli including an environment that visually depicts all space and creates the sense of moving through that space. In addition, novice users will probably find it easier to learn how to interact with this IVE. The largely text-based version uses the enCore Xpress MOO interface, and while it does allow the inclusion of some visual and audio components, it requires students to use written language to interact with the virtual environment. The MOO be better able to create multiple perspectives of the multiple cultures represented in the site. In addition, students will, after a process that involves review by scholars and others, be able to enrich the MOO with their own writing, to become producers as well as consumers of content. The MOO version is platform independent and requires less bandwidth than the 3-D graphically rich version. Both types of interfaces have the ability to engage users in the environment, but the question is whether one or the other is more effective for certain types of learning.

The 3-D graphically rich version of the Like-a-Fishhook IVE will be used by archeology students to learn to think like archeologists. However, both versions will be used by a multidisciplinary, writing-intensive humanities class, which will include students from Fort Berthold Community College, descendents of the village inhabitants. Students will experience the site via the IVEs: they will observe, analyze, reflect-and write-about the area, the cultural changes, the history, and the virtual environments themselves. Students from all disciplines will meet together (at least in virtual space) and discuss the conventions of their disciplines. Then they will explore the IVEs together with a set of questions that require exploration and investigation in order to find the answers. They will also write parallel assignments that are appropriate for their disciplines. For example, all students will keep journals to record their visits to the IVEs, journals that stress purposeful observation and inquiry, but each will concentrate on recording the type of information suitable for his or her discipline as they travel through the IVEs. The journals will be used as the basis for other writing assignments. For example, all students might write about an artifact found during the excavation, but the anthropology student might write an analysis of its cultural significance, the creative writing student a poem or a piece of creative non-fiction about it, and the public history student a museum script for it. Longer papers will require additional research using conventional sources as well as more innovative ones like the Digital Archive Network for Anthropology (DANA) being developed at NDSU and which includes 3-D scans of actual objects found at the Like-a-Fishhook site. The creative writer might write a piece of historical fiction, the anthropology student a scholarly article on some aspect of the cultures depicted in the IVEs, and the public history student a brochure for the site such as the ones that are produced for visitors to actual historic sites. Students from all disciplines might collaborate on the making of a short documentary on some aspect of the area or a multi-vocal web essay that combines the expertise and perspective of several disciplines. After a process of review, student writing can be incorporated into the MOO version to enrich it for future visitors.

The value of each IVE as it relates to the learning in the humanities will be assessed in several ways. The journals kept by the students during their visits to the IVEs will be assessed to determine (1) the level of engagement in each version of the IVE, (2) any differences they note about the different experiences in the IVEs, (3) the type of information they notice in each IVE, and (4) how well they understand the time periods, the cultures, and the places they are observing. In addition, students will complete surveys that ask them to respond to questions that solicit similar information. These results will provide useful information about the strengths and weaknesses of both types of interfaces.

REFERENCES

Clark, Jeffrey T., Brian M. Slator, Aaron Bergstrom, Francis Larson, Richard Frovarp, James E. Landrum III, William Perrizo. (2001b). "Preservation and Access of Cultural Heritage Objects Through a Digital Archive Network for Anthropology." Proceedings of the 7th International Conference on Virtual Systems and Multimedia (VSMM-2001). Berkeley, CA, Oct. 25–27.

Haynes, Cynthia, and Jan Rune Holmevik. eds. (1998) *High Wired: On the Design, Use, and Theory of Educational MOOs.* University of Michigan.
McClean, Phillip, Bernard Saini-Eidukat, Donald P. Schwert, Brian M. Slator, Alan White (2001). "Virtual Worlds in Large Enrollment Biology and Geology Classes Significantly Improve Authentic Learning." In *Selected Papers from the 12th International Confernce on Collge Teaching and Learning* (ICCTL-01), Jack A. Chambers, Editor). Jacksonville, FL: Center for the Advancement of Teaching and Learning. April 17–21, pp. 111–118.
Slator, Brian M., Jeffrey T. Clark, James Landrum III, Aaron Bergstrom, Justin Hawley, Eunice Johnston, and Shawn Fisher (2001). "Teaching with Immersive Virtual Archaeology." Proceedings of the 7th International Conference on Virtual Systems and Multimedia (VSMM-2001). Berkeley, CA, Oct. 25–27.
Smith, G. Hubert. 1972. "Like-a-Fishhook Village and Fort Berthold, Garrison Reservoir, North Dakota." *Anthropological Papers* 2. National Park Service. U.S. Department of the Interior. Washington, D.C.

Computational Approaches to Linguistic Variation

JOHN PAOLILLO
Indiana University
paolillo@indiana.edu

JOHN NERBONNE
Rijksuniversiteit Groningen
nerbonne@let.rug.nl

WILLIAM KRETZSCHMAR
University of Georgia
kretzsch@arches.uga.edu

JEAN-CLAUDE THILL
SUNY Buffalo

PART I

Linguistics meets the other humanities, especially cultural history, most deeply and extensively in studies of linguistic variation. Starting with research in the nineteenth century, when Gillieron demonstrated that French linguistic geography (dialectology) found patterns similar to those in folk architecture and legal traditions, and continuing with Kurath's demonstration of a link between American vocabulary and colonial patterns of migration and settlement, the study of linguistic variation has continued to be a focal point for interdisciplinary studies in the humanities.

Advances in information technology have revolutionized linguistics in general and the study of linguistic variation in particular in several ways. First, we are able to manipulate much larger archives of data than were previously possible. Whether in text, audio or multimedia form, the empirical base of data available to linguistic scrutiny is far larger today than ever before. Thanks to the internet and the wolrd-wide web, remote resources are becoming available and are being incorporated into analytical procedures. Second, our analyses can achieve greater sophistication, whether through simulation, computationally assisted analysis, or complex statistical modelling. These linguistic analysis techniques include especially algorithms for sequence comparision ("Levenshtein distance" or "edit distance"), morphological analysis (lemmatization or "stemming"), basic syntactic categorization ("part-of-speech tagging"), the analysis of textual affinity (the small neighborhoods within which words appear), thematic affinity (the larger neighborhoods within which words appear), and full syntactic analysis ("parsing", or the assignment of syntactic structure to sentences and phrases). Taggers, lemmatizers and parsers are now common tools for all sorts of linguistic analysis, and statistical models are possible in analyses that were entirely categorical in previous decades.

These technical advances have opened new perspectives on the analysis of linguistic variation in modern linguistic analyses. Social and geographic patterns of language variation are revealed in carefully planned corpus studies, facilitated by the use of the language analysis techniques, and visualized using further

computational techniques. At the same time, these developments enhance our ability to test theoretical models using empirical data, and sooner or later point out their flaws and shortcomings, challenging our prior conceptions of the processes of linguistic variation by exposing the weaknesses in our theoretical models.

The purpose of this workshop is to bring researchers together who wish to harness computational power as a source of improvement in methods for studying language variation. To this end we bring together three perspectives on the impacts of computational methods on studying variation in language. These three contributions address the use of geographic information systems as means to explore linguistic variation and its geographic dependence, the use of linguistic processing to refine the analysis of variation, and the role of statistical models in large-scale analyses of language variation. These three perspectives provide three important reference points for the application of new computational techniques to the study of language variation. Together, they provide a view of the role of computational techniques in understanding this central topic in the humanities.

PART II
VOCABULARY AND PRONUNCIATION IN LINGUISTIC VARIATION

John Nerbonne

Linguistics, archaeology and cultural history share a fascination for the means with which people express identity, in particular, the associations with the area in which they live, with their social class and gender, perhaps with their ethnicity and profession. Computational studies of linguistic variation enable us to see the degree to which various linguistic devices contribute to this expression of identity. The first task of these computational studies is inevitably to bring some order to the plethora of data relevant to the study of cultural expression. In the case of linguistics, this is speech and writing, and the relevant data is available in the form of dialect atlases, compendia of varying linguistic forms collected in comparable ways from speakers of varying geographic and social origins.

Given a large amount of dialect data, there is a good chance that one will encounter “noise”, i.e., inaccuracy, nongeographic variability, and incompatibility both in the choice of information recorded and in the level of detail at which it is recorded. In addition, dialectologists have been aware since Kloeke and Bloomfield that, even abstracting away from the noise, dialect varieties inevitably contain genuine linguistic features with counterindications, exceptions, and gaps. There are furthermore many linguistic features to explore, and many ways of combining them. Finally, it may be the case that it is difficult or even impossible to validate results—there may be no consensus among dialectologists about which aspects of the geographic distribution of linguistic variation are most significant. The LAMSAS data set, available at **http://us.english.uga.edu/lamsas/**, is one such large, “noisy” and rebarbative set. The challenge is to identify how linguistic similarity is expressed in such data.

We treat both lexical and phonetic differences in this talk, and we also examine the relative contributions of pronunciation and lexis to dialect differentiation. In order to rise above the atomistic level of the individual sounds or lexical items, we employ aggregate measures of distance, the (non-)identity of lexical items on the one hand (essentially the same measure proposed by Seguy, 1971, and elaborated on by Goebl, 1984); and a string similarity measure which we apply to phonetic transcriptions on the other. Because the measures yield numeric characterizations of lexical/phonetic distance, it may be aggregated over many pairs of similar concepts. In order to overcome the problem that there is little expert consensus, we propose a numerical characterization of the fundamental dialectological postulate, namely that of “local coherence”: nearby language variants tend to be similar to one another. Such a principle requires clarification as to which language variants are to be included, and it is admitted not generally true (e.g., town Frisian is geographically not coherent). But we show nonetheless that the principle can be put to beneficial use in exploring dialect data in an investigative phase of research. We illustrate the results when applied to the entire LAMSAS data set, and use it to help choose which infrequent data to omit from analysis and also to evaluate the two modest modifications we propose to Seguy and Goebl’s work.

The results of the lexical analysis confirm neither of the best known dialect divisions for the LAMSAS area, i.e., neither Kurath’s nor Carver’s. Both Kurath and Carver relied on lexical analysis, where Carver’s (1987) analysis sees the North-South division as dominating dialect differences on the Atlantic coast, while Kurath’s saw a significant “Midlands” area corresponding to southern Pennsylvania and extending into West Virginia and the inland South. In our analyses, Kurath's “Midlands Area” is split into North and South, and the penultimate aggregation is unstable, confirming sometimes Kurath and sometimes Carver. Further details confirm Kurath rather more than Carver, e.g., in recognizing a coastal South region. Phonetic analysis confirms the primary significance of the North-South split.

Finally, we are in a position for the first time to evaluate the usual assumption of variationists that

lexical and phonetic variation usually "coincide fairly well" (Kurath and McDavid, 1961) in their association with extralinguistic variables such as geography (where modern variationists would tend to add social class, gender and age). In fact the two levels of linguistic structure do tend to correlate to a highly significant degree ($r = 0.65$) in the LAMSAS data, and likewise therefore associate with the same geographical areas, but the lexical data is much less consistent. To achieve the same consistency of measurement, we need to examine ten pairs of lexicalizations for every single pair of pronunciations. We express our associations, and ultimately, our personal identify whenever we use language, but the expression is ten times as recognizable in speech as it is in writing.

REFERENCES

John Nerbonne with Wilbert Heeringa and Peter Kleiweg "Edit Distance and Dialect Proximity" In: David Sankoff and Joseph Kruskal (eds.) *Time Warps, String Edits and Macromolecules: The Theory and Practice of Sequence Comparison* 1999, pp.v–xv.

John Nerbonne and Peter Kleiweg "Lexical Distance in LAMSAS." Submitted to: John Nerbonne and William Kretzschmar (eds.) *Computational Methods in Dialectometry*: Special issue of *Computers and the Humanities*, scheduled to appear in 2003.

PART III
ZOOMING IN ON LONGITUDINAL VARIATION

John C. Paolillo

Studies of language change traditionally rely on two kinds of methods. The first and oldest involves systematic comparison of evidence from different historical periods of a language. Typically, the evidence is drawn from written texts, which are usually written by different authors at widely separated periods. Such studies face the challenge of arguing that the texts used are comparable enough to justify their comparison, and to license the inference from observed variation to change. But numerous contextual factors, such as text genres, author identities, regional dialect differences, etc. confound such comparisons (Herring, et al. 2001). The second type of study looks at synchronic variation in a speech community and compares the age distribution of linguistic features among members of the community to draw inferences about earlier and later historical periods. Such studies must assume that language variation is relatively stable within any given individual, or there would be no reason to believe that an individual's current linguistic behavior accurately reflects an earlier time period.

Consequently there are large gaps in our understanding of language change. To what extent is intra-individual variation implicated in language change? And to what extent does intra-individual change accompany historical change? Some recent studies of language variation and change partially address these issues by combining community studies with longitudinal data collection (e.g., Yoneda 1993), others develop large-scale statistical models of multiple authors over time (e.g., Burrows 1992, Finegan and Biber 1989), and even others employ corpora of spoken data from radio (Van de Velde et al. 1998) and cinema (Elliott 2000) controlling for effects of medium and genre. Yet, the field still lacks studies comparing continuous longitudinal variation to individual variation.

Two recent developments, both linked to changes in computer technology, facilitate a new kind of study that can begin to answer these questions. The first is the ability of modern computers to store and analyze large amounts of language data. Corpus linguistics is emerging as the preferred paradigm for studying language variation and change: not only can the computer assist in locating data items of interest in large compendia of computerized data, but the ability to handle large amounts of data helps us control for more contextual factors and arrive at more reliable statistical analyses. The second development is the widespread adoption of the Internet as a communications medium. People with a broad range of personal and professional backgrounds use the Internet for primary communication in a wide range of purposes. Much of the material is archived so that continuous records lasting more than ten years are now becoming available. Through careful selection of archived materials, it is possible to construct corpora of continuous Internet communications that control for contextual factors such as genre as well as individual usage. Furthermore, the Internet is believed by some to be an active force for language change within English, by promoting certain forms of abbreviation and ellipsis, increased informality, decreased politeness, and the like (Baron 1984, 2000; Crystal 2001). Currently available archives of Internet communication hold the key to understanding whether these things are in fact true. In addition, the ability to view continuous records of language use offers an unprecedented opportunity to observe the relation between individual variation and language change.

In this paper, I examine variation among 21 different linguistic variables over the complete 11-year history of MsgGroup, and early ARPANET/Internet email discussion group established in 1975 to discuss the protocols of electronic mail. The linguistic variables were selected from the 67 features in Biber (1988), so

that candidate data items could be readily identified automatically, and so that the variation in the electronic mail corpus could be readily compared with a comprehensive study of English variation. The variables were studied both in aggregate, using factor analysis, and individually, using variationist models. Messages from each of the 329 participants in MsgGroup were tracked over the entire 11-year span, so that individual longitudinal trends could be observed and compared to the overall trends.

What emerges from this study is a very complex picture of language variation which points not to language change, but rather individual adaptations based on immediate and local communicative circumstances. In the factor analysis, three fairly robust factors emerge: a factor of general elaboration, one of lexical and syntactic complexity, and one of person reference and epistemic certainty. None of these factors correlates strongly with historical time, and the variance on all factors remains high throughout the 11-year observed span. In other words, the main shared patterns of variation in the corpus do not turn out to be patterns of change over time. Individually, however, some of the linguistic variables, such as first and second person pronouns, are significantly correlated with historical time, and when patterns of individual participants are analyzed, yet other significant trends are observed. In many cases, such as the sharply decreased use of contractions over time by one senior female system administrator, these trends run counter to the trends of most other group members. Sometimes these trends are traceable to specific events, such as a particularly active and divisive debate that occurred in on MsgGroup in 1979 concerning the finger protocol. Individuals appear to adapt their use of linguistic variables in response to changes in status and social affinity that arise from these events.

Zooming in on longitudinal language variation thus reveals a complex situation in which individual variation and historical change do not necessarily recapitulate one another, as suggested by prior models. Instead, individual patterns of variation appear to respond to complex local and political concerns resulting in broad intra-group variance when viewed continuously over time. These findings thus challenge us to propose theories of language change that successfully relate such complex continuous variation to long-term historical trends.

REFERENCES

Baron, Naomi. 1984. “Computer-mediated communication as a force in language change”. *Visible Language* 18: 118–141.

Baron, Naomi. 2000. *Alphabet to email: How Written English Evolved and Where it’s Heading*. London: Routledge.

Biber, Douglas. 1988. *Variation Across Speech and Writing*. Cambridge: Cambridge University Press.

Biber, Douglas and Edward Finegan. 1989. “Drift and the evolution of English style: a history of three genres”. *Language* 65: 487–517.

Burrows, J. F. 1992. “Computers and the study of literature”. In C. S. Butler, ed., *Computers and Written Texts*, 167–204. Oxford, UK: Blackwell.

Crystal, David. 2001. *Language and the Internet*. Cambridge: Cambridge University Press.

Elliott, Nancy. 2000. *A Sociolinguistic Study of Rhoticity in American Film Speech from the 1930s to the 1970s*. Doctoral dissertation, Indiana University.

Herring, Susan, Lene Schosler, and Pieter von Rainen. 2000. *Textual Parameters in Older Languages*. Amsterdam: Benjamins.

Van de Velde, Hans, M. Gerritsen & Roeland van Hout. 1996. “The Devoicing of Fricatives in Standard Dutch. A real-time study based on radio recordings”. *Language Variation and Change* 8: 149–175

Yoneda, Masato. 1993. “Survey of standardization in Tsuruoka City, Japan: Comparison of results from three surveys at twenty-year intervals”. Paper presented at *Methods VIII: International Conference on Dialectology*. University of Victoria, British Columbia.

PART IV
SELF-ORGANIZING MAPS AS AN APPROACH TO GIS ANALYSIS OF LINGUISTIC DATA

William A. Kretzschmar, Jr., and Jean-Claude Thill

The nearest-neighbors method of Density Estimation and Complete Spatial Randomness methods in general will be best applied in models that consider the status of individual linguistic features (e.g. Kretzschmar 1996, Kretzschmar and Lee 1993). Self Organizing Maps (SOM) uses grouping algorithms in order to model dialects, not just features. The notion of “dialect” for each method, however, is not equivalent to the traditional NeoGrammarian or Bloomfieldian sense of the term, but instead derives from the mathematical procedures used to build groups. In this paper we discuss the application of the technique to data from the Linguistic Atlas of the Middle and South Atlantic States (LAMSAS. The programming that supports the SOM

project has been carried out with MapObjects for the layered displays, and C++ for statistics, and Visual Basic for integration of displays and functions; the procedure is fully automated, subject to user-selected parameters. Besides the implementation of the SOM algorithm, the program is useful as a general-purpose tool to recover information about speakers and to measure groups of speakers against the overall set of speakers with univariate statistics.

The model for SOM is the neural network of the human brain. This is perhaps an overly romantic characterization, which comes down to reduction in the complexity of input data through reduction of the number of dimensions in which it occurs, until we have a two- dimensional lattice, or feature map. The basic idea as applied to LAMSAS is that a set of input nodes corresponding to each of the 1162 LAMSAS speakers, for a set of N linguistic targets, will be processed statistically to form output nodes which may then be compared for their relative similarity (our procedure follows the SOM algorithm as elaborated in Roussinov and Chen 1998). This "similarity" is computed as the distance between nodes in N dimensions, corresponding to the N linguistic targets. Processing takes place in N iterations, one for each of the N targets. The user must specify how large a matrix of groups of speakers the statistic is to create (say, a 5 x 5 matrix of 25 groups), and how many groups of groups will be established within the matrix (say, three relatively similar groupings out of the 25 groups in the matrix.

For example, in a typical SOM run (say, on 19 different linguistic features, consisting of the three common variants of the gully item, the nine common variants of the heavy rain item, and the seven common variants of the thunderstorm item), the algorithm selected and displayed, as instructed with a parameter setting, three groupings of speakers identified by color within a matrix of 25 smaller groups of speakers. Within this matrix, the positions of the groups are not in abstract geographical space; the speakers within each of the 25 groups need not be located anywhere near each other. The SOM program also produces a graphical representation of the "distance" between each of the 25 nodes (a "Umatrix" display), in which the darkness of the colored diamond between each node (the dots) indicates the degree of relation between adjoining nodes. Each node of the matrix is convertible to a layered GIS display which shows which individual LAMSAS respondents belong to the node. Some of the nodes contain large numbers of speakers, some few speakers. A display from one of the nodes just a single speaker, (a frequent occurrence in this rain/storm/gully matrix), or may reveal a geographical cluster of speakers. By observing the groupings that the SOM algorithm creates, and by inspecting the statistical outputs which underlie the groupings, we hope to learn more about the general behavior of groups within the region.

Our research is now complete enough to suggest how useful this approach will be for the linguistic goals of LAMSAS. We will compare results from many SOM runs on different combinations of features, and then discuss the nature of the estimates produced by the SOM algorithm. We can then compare SOM results to those derived from another method for creation of diealect models, the Levenstein Distance algorithm as now being executed by Nerbonne, Heeringa, and Kleiweg.

REFERENCES

Kretzschmar, William A., Jr. 1996. "Quantitative Areal Analysis of Dialect Features". *Language Variation and Change* 8.13–39.

---, and Jay Lee. 1993. "Spatial Analysis of Linguistic Data with GIS Functions". *International Journal of Geographical Information Systems* 7: 541–60.

Roussinov, Dmitri, and Hsinchun Chen. 1998. "A Scaling Self-Organizing Map Algorithm for Textual Classification". *Artificial Intelligence* 15:81–112.

PAD: Preservation, Archiving, and Dissemination of Electronic Literature

DAVID DURAND
Brown University
David_Durand@brown.edu

MARJORIE COVERLEY LUESEBRINK
ELO, Irvine Community College
luesebr1@ix.netcom.com

NICK MONTFORT
U. Penn
nickm@nickm.com

JESSICA PRESSMAN
ELO
jesspres@ucla.edu

SCOTT RETTBERG
ELO, Richard Stockton College
rettberg@eliterature.org

The Electronic Literature Organization is a non-profit organization with the mission of promoting and facilitating the reading, publishing, and writing of electronic literature. ELO is the only group of its kind, providing the infrastructure and information for the dispersed community of writers and readers working in electronic media. In its three years of existence, ELO has presented numerous readings, produced the first major hypermedia awards (2001), held the first international symposium on the topic of electronic literature (2002), and created a free online database for housing and exploring electronic literary works (The Directory).

The State of the Arts Symposium, held at UCLA in April of 2002, gathered people interested in electronic literature from diverse areas of interest and expertise into discussion of the primary issues and concerns facing this cultural form. The overwhelming area of concern proved to be the threat of obsolescence due to rapidly "advancing" technical platforms. Out of this event grew the impetus and initiative for ELO's PAD (Preservation, Archiving, and Dissemination) Project.

ELO's drive to fulfill our mission statement has prompted us to initiate a preemptive strike on the impending obsolescence of electronic literature by the inevitably short life span of its programmable media. PAD is a multi-leveled and multi-leveled initiative aimed at approaching this problem from technical, academic, legal, and communal perspectives. PAD is currently in its planning phase, operating through five committees: Operations, Archiving and Display, Academic Dissemination, Technological and Software Development, and Copyright and Open Source. PAD is voluntarily managed by scholars and experts from such ranging fields positions as experts in literature, computer science, and digital archiving to creative writers, businessmen, and a foundation representative. It is obvious from our committee members that the issue of preserving, archiving, and disseminating digital art and information is an urgent topic in all fields, one that unites disciplines and departments. Digital archiving is a subject under exploration and experimentation by numerous groups from various disciplines, many of which will be represented at the ACH conference. ELO's panel will include five PAD committee members, each of whom represent different PAD committees and will therein present various aspects of the problem of digital obsolescence, the ramifications and potential solutions as discerned and understood by PAD.

This academic panel will address the need for pre-emptive measures that preserve and archive endangered works of electronic literature as well as inform about PAD's current operations and future plans. The fact that electronic literary works are constantly at risk of fading into technological obsolescence will inevitably obstruct the ability to teach and study works already in existence as well as impede upon future artistic creation. As all conference attendants know, the ability to do scholarly or pedagogical work in the

humanities is entirely dependent upon the archiving capabilities and technological support available. The fact that important works of contemporary literature will soon be inaccessible to scholars, editors, librarians, or students is a major impediment to our cultural and technological future. This panel is therefore a vitally important one to present at ACH.

Due to the diverse reaches of this project, ELO is submitting proposals for two related sessions to this year's ACH conference. The first is the session above, a traditional academic discussion of ELO's PAD project, its objectives and implications. The second, presented in a separate proposal, is a creative session of short readings of electronic literature. This combination of academic and creative panels aims to present electronic literature to an audience with diverse and divergent exposure to this literary and technological art form. In so doing, these sessions will jointly present the logical, intellectual, and aesthetic reasons, for supporting and encouraging the preservation of electronic literature. PAD's work depends upon and inevitably affects all people working, either centrally or tangentially, in computing, the humanities, and moreover, humanities computing. We look forward to the opportunity to share our work, brainstorm for further ideas, and generate interest and support for this project. We hope that ACH will invite us to participate in this year's conference.

PANELISTS AND A BRIEF DESCRIPTION OF THEIR PRESENTATION

David Durand will discuss technical aspects preserving e-texts by means of format conversion, to marked-up format, and the concomitant need for the development of appropriate re-presentation software. This strategy contrasts in several distinct ways with a system emulation strategy (as discussed) by Nick Montfort. As a running example, he will discuss his experience with both conversion and emulation in rescuing data from Brown University's FRESS hypertext system from the early 70s.

Marjorie Coverley Luesebrink will examine a selection of endangered and threatened works of electronic literature. Many of the early examples of hypertext literature exist only on platforms that cannot be executed by contemporary computers. As part of the Preservation, Archiving, and Dissemination activities, the committee is compiling a comprehensive list of e-literature works that can no longer be accessed in their original form. The near-obsolescent works often represent key eras in the development of e-literature and its supporting technologies. "The 'O' Word" will look at the coding practices, literary strategies, and placement in the evolving history of the field that might argue for preservation of these and other pieces.

Nick Montfort will discuss how the experience of electronic literature works as functioning, interactive computer programs is often essential to their appreciation. He will describe how re-implementation, the development of new interpreters, and emulation has been used to keep works such as Weizenbaum's 1965–66 Eliza/Doctor and several decades of interactive fiction accessible and fully functional. He will also consider how this approach can be used today as part of a strategy for preserving e-lit works of many different sorts, works which will be read and studied for many different reasons.

Jessica Pressman will provide an explanation of the position of ELO within the academic and cultural community of electronic literature, ELO's resources and its reasons for accepting the responsibility of the PAD project.

Scott Rettberg will outline some of the challenges particular to the preservation of electronic literature, identify some of the archiving methodologies that the ELO and other institutions and organizations are currently applying, and introduce the ELO's Preservation, Archiving and Dissemination initiative. Rettberg will summarize the initial findings of PAD's working group, and its general approach to the challenge of lessening the impact technological obsolescence on the distribution and study of electronic literature.

Present and Future Directions in Developing Online Resources for Renaissance Studies

WILLIAM R. BOWEN
Renaissance Society of America / University of Toronto
william.bowen@utoronto.ca

RAYMOND G. SIEMENS
Malaspina University College
siemensr@mala.bc.ca

STEPHANIE F. THOMAS
Sheffield Hallam University
petals@btinternet.com

CHRIS R. ROAST
Sheffield Hallam University
petals@btinternet.com

INNES E. RITCHIE
Sheffield Hallam University
petals@btinternet.com

PART I
ITER: BUILDING AN EFFECTIVE KNOWLEDGE BASE

William R. Bowen

The relationship of a card catalogue to a collection of print media is insufficient as a model for Iter, a not for profit partnership devoted to enhancing the teaching and study of the European Middle Ages and Renaissance through the development of online resources. Indeed, as it becomes increasingly common to distribute digital collections over the web, it is becoming very clear that Iter can provide more than a sophisticated body of inter-related databases which include pointers to digital collections by enabling researchers to interact with the digital documents themselves. This will, of course, raise new challenges of collaboration, access, knowledge management, standards, and delivery, all of which require answers reflecting the needs of Iter's community of scholars.

This paper will focus on the strategies currently being entertained by Iter in building an effective knowledge base for study of European culture form 400 to 1700. It will begin by briefly outlining the current status of its databases which seek to describe the gamut of print and digital media used for formal scholarly communication (e.g. articles, essays, books, and reviews), the more ephemeral scholarly communications (e.g. calls for papers, awards, grants, research opportunities), the relevant academic societies and research institutions, and, finally, the scholars themselves. Some attention will be given to the inter-relationships between these databases and the ways in which Iter will provide powerful searching tools and alerting services.

The presentation will continue with the more challenging issue of the connection between the databases and the objects being described. From the perspective of the user of Iter's resources, this part of the paper will look both at current models of commercial and non-commercial knowledge bases, and selected literature (e.g. on the W3C Semantic Web), in order to elucidate current expectations and how they might be met effectively. In this way, the paper has a natural connection to other presentations in this session which focus on the online presentation of documents.

REFERENCES

Besser, Howard. "The Next Stage: Moving from Isolated Digital Collections to Interoperable Digital

Libraries." *First Monday* 7.6 (June 3, 2002).
Bowen, William R. "Iter. Where Does the Path Lead?" *Early Modern Literary Studies* 5.3 / Special Issue 4 (January, 2000): 2.1–26.

PART II
ALGORITHMIC APPROACHES TO AN ELECTRONIC SCHOLARLY EDITION OF EARLY MODERN MATERIALS

Raymond G. Siemens

For more than a decade, two dominant perspectives on the electronic scholarly edition have prevailed. One, the "dynamic text," consists of an electronic text and advanced textual analysis software; it presents, in essence, a text that indexes and concords itself, allowing the reader to interact with it in a dynamic fashion (see Lancashire, 1989). The other, often referred to as the "hypertextual edition," exploits the ability of encoded hypertextual organisation to facilitate a reader's interaction with the apparatus (textual, critical, and otherwise) that traditionally accompanies scholarly editions, and with relevant external textual and graphical resources, critical materials, and so forth (Faulhaber, 1991, 134 ff.). Advances in computing, and computing's further advance into disciplines that include textual editing, over the past decade have made it clear that electronic scholarly editions can incorporate dynamic interaction with the text and its related materials and, at the same time, also reap the benefits of the fixed hypertextual links that typify the standard relation of materials we find in most editions of this sort.

Indeed, contemporary scholarly consensus is that the level of dynamic interaction in an electronic edition itself—if facilitated via text analysis in the style of the dynamic text—can replace much of the interaction that one typically has with a text and its accompanying materials via explicit hypertextual links in a hypertextual edition (surveyed in Siemens, 2001, 2002). That said, at the moment, there is no extant exemplary implementation of this new "dynamic edition," an edition that transfers the principles of interaction allowed by a dynamic text to the realm of the full edition, comprising that text and all its extra-textual materials (textual apparatus and commentary, and beyond).

Toward documenting the elements of a dynamic edition, and toward the ultimate goal of creating an exemplary instance of such an edition, my paper will provide a typological survey of a number of contemporary electronic scholarly editions; in doing so, it will explicitly raise and address the several key concerns associated with applying algorithmic approaches associated with text analysis to such an edition. Discussion will include issues relating to the creation and use of digital resources in the humanities, specifically digitization and representation of text, encoding techniques, interface, and metadata. Discussion will centre on, and examples will be drawn from, among others, an electronic edition of the Devonshire MS (British Library Additional MS 17, 492) that is currently in progress.

REFERENCES

Faulhaber, Charles B. "Textual Criticism in the 21st Century." *Romance Philology* 45 (1991): 123–148.
Lancashire, D. Ian. "Working with Texts." Paper delivered at the IBM Academic Computing Conference, Anaheim, 23 June 1989. Noted in Faulhaber (128, 135).
Siemens, R.G. "Unediting and Non-Editions." In *The Theory (and Politics) of Editing*. *Anglia* 119.3 (2001): 423–455. Reprint, with new introduction, of "Shakespearean Apparatus? Explicit Textual Structures and the Implicit Navigation of Accumulated Knowledge." *Text: An Interdisciplinary Annual of Textual Studies* 14. Ann Arbor: U Michigan P, 2002. 209–240.

PART III
THE EXPLORATION AND DEVELOPMENT OF TOOLS FOR ACTIVE READING AND ELECTRONIC TEXTS

Stephanie F. Thomas, Chris R. Roast, Innes E. Ritche

The Active Reading project, based at Sheffield Hallam University, is concerned primarily with creating an electronic scholarly edition of a Renaissance work—namely Shakespeare's King Lear, which makes available the textual variants of published editions of that work. This serves to enhance scholarly activity by enabling insight into the variety of editorial practices that have contributed to understanding of the work. It is important that, within the study of literature and critical analysis, reading should not be viewed as a passive activity but one in which a reader's active interpretation is central and can be supported by access to textual variants.

To examine several paper-based editions of a work for textual variants, is to undertake a time-consuming and somewhat confusing task—faced with reams of paper, and having to "keep place" within that material in order to make comparisons. A modern solution to this task would naturally be to gather the various editions together and produce one electronic edition, where points of variation amalgamate the individual paper editions and make them interactively and visually available to the reader. We refer to this as a definitive edition; limiting the loss of structure (both textually and physically) that a conflated text/ edition might produce.

A prototype system to support active reading has been developed, which consists of an edition of the various available texts of a poem by Sir Thomas Wyatt[1]. The combined edition is encoded in XML and various methods for displaying the texts and their variants are implemented using a combination of XSL and JavaScript. The prototype has served to illustrate the mechanisms by which the definitive edition could function, and has demonstrated the feasibility of this approach.

The project focuses upon the reasoning for choices and selection in the encoding of the text—why one method is chosen over another, how this affects the appearance of the edition, and the decisions open to the end user. Hence a definitive edition determines the choices the user/ reader can make in the exploration of what is in effect the readers own edition of the text. The overall objective is to enhance how users engage with the text(s) through interactivity, in a manner that does not hamper their creativity and thus enables scholarly development and understanding. Evaluation of the current prototype edition has been undertaken and is ongoing amongst the target user group—undergraduate students of English Studies. Preliminary results are providing valuable insights into understanding of the needs of the user/reader, and criteria by which effective active reading can be judged.

This paper aims to illustrate the advantage of enabling understanding of the editing process, through the use of interactive technologies and text encoding tools. The resulting edition and the research generated will offer new ways of comparing textual variants, and of reading and understanding these texts —for use in research, in teaching, as a learning tool, and as a template for the creation of future electronic editions. The interdisciplinary nature of the project itself offers a challenge, attempting to merge the worlds of Renaissance Literature, Human-Computer Interaction (HCI), and Computing. Employing HCI methods in designing and developing a new edition is a vital tool for understanding the active reading requirements, and for generating an edition that is both simple and effective to use.

NOTES

[1.] **http://homepages.shu.ac.uk/~sfthomas/activeR/theyflee.html** [NB: can be viewed with Internet Explorer only, page also uses frames.]

Identifying Multiword Tokens Using POS Tagging and Bigram Statistics

MARK AREHART
University of Michigan
marehart@umich.edu

I describe and evaluate three methods for automatically identifying in English text a frequently occurring type of multiword token, the lexicalized noun compound. The methods combine symbolic part-of-speech information with different measures of collocational strength, namely minimal frequencies of occurrence in a corpus, log likelihood of association, and a combination of these two. The results of testing the methods on two software manuals of approximately 170,000 and 210,000 words suggest that although raw frequency is the best single measure overall, the combined strategy is useful to the extent that one favors precision over recall. I also discuss the limitations of the corpora and the test and suggest additional applications of the methods.

A noun compound, like stone wall or stock market, is a series of nouns that function syntactically as a single noun, inheriting the features of the head (final) noun. In some languages, noun compounds are not separated by whitespace and are thus trivially identifiable as words. For instance, the English compound departure time is Abfahrtzeit in German (Abfahrt 'departure' + Zeit 'time') and lähtöaika in Finnish (lähtö + aika). A lexicalized compound is one that has acquired a conventional or specialized meaning. The compound garbage man, for instance, could refer to a man made of garbage (cf. snow man) or a man who delivers

garbage (cf. milk man), but has a more salient lexicalized meaning. In some cases, lexicalization is reflected in orthography, as in the single words fireman and policeman, and occasionally one finds both multiword and single-word versions, such as air mail and airmail (both attested in the Brown corpus). It would be useful to be able to treat compounds like garbage man as single terms on par with their orthographically unitary counterparts for purposes such as document indexing and classification.

I approach the task of identifying such compounds as a secondary tokenization step. Tokenization, often unfairly regarded as an uninteresting bit of text preprocessing, requires one to make nontrivial decisions about what constitute minimal "word-like units" for further analysis (Grefenstette & Tapanainen 1994). In addition to garden-variety words, tokens include punctuation, which is important for identifying clause and sentence boundaries, and multiword units. Karttunen et al. (1996) divide these multiword units into several categories: adverbial expressions like "all of a sudden," prepositions such as "in spite of," date and time expressions, proper names, "and other units." Typically, a basic tokenizer first segments the text into simple units, then one or more "multiword staplers" group tokens together again (Karttunen et al. 1996). What is unique about the approach presented here is the combination of part-of-speech information, used to identify noun compounds, and collocational measures. Although "highly collocated" and "lexicalized" are not the same thing, I suggest that the former can serve as one useful indicator of the latter.

The procedure works as follows. The text is processed by a basic tokenizer, tagged for part-of-speech, and then the noun sequences are extracted. The goal is then to identify the subset of these compounds that might qualify as lexicalized terms by measuring the collocational strength of the component nouns. The simplest way to identify possible collocations is to extract all those that occur at or above a certain frequency cutoff. One might hypothesize, for example, that if a certain noun compound occurs five times in a text, then it has a lexicalized meaning. Collocational strength can also be measured by compiling all of the bigrams found in the text and comparing the rate of co-occurrence of the elements with that expected by chance. Although there are several possible measures, I use the log likelihood statistic, which has been shown to be preferable to alternatives such as chi-square and mutual information (Dunning 1993). To generate collocations of more than two words, a separate bigram merging process is performed on the corpus. If, for example "NASDAQ composite" and "composite index" are both significant collocations, then the trigram "NASDAQ composite index" will be extracted as well if it occurs in the corpus. In practice, this method can generate terms that are quite long, such as "American Stock Exchange Market Value Index," an example extracted from a portion of the Wall Street Journal corpus. It is also possible to combine these methods, by extracting compounds that are above certain frequency and likelihood threshholds. Although the two measures generally correlate (that is, frequently occurring compounds tend to have larger likelihood scores), they are sensitive in different ways to corpus size.

To evaluate the procedure, I extracted noun compounds from two software manuals and compared each list to the compounds found in each manual's index, which would be expected to contain the significant terms. A baseline procedure using all the noun compounds averaged 0.26 precision and 0.77 recall on the texts. In other words, about a quarter of the compounds occurring in the texts were found in the indexes. Recall is less than 1.0 because some ideas or topics that do not occur in the text as compounds are reformulated as such for the index. The 0.77 score thus serves as an upper bound on recall. Assigning equal weight to precision and recall, the best performing strategy was to use a minimum frequency of 4 for the first text and 3 for the second, with an average of 0.474 precision (82.3% higher than the baseline) and 0.527 recall (31.6% lower than the upper bound). Adding a conservative log likelihood score threshold increased precision by an average of 13.6% but lowered recall by an average of 24.1%. Unless one substantially discounts recall, the likelihood score was not as useful as anticipated.

The results indicate that measures of collocational strength are useful in separating lexicalized noun compounds, which are profitably viewed as multiword tokens, from nonlexicalized ones that are best analyzed as token sequences. Such methods can facilitate the automatic indexing and classification of documents for textual analysis and search and retrieval applications. Two important limitations on these results are the restriction to a particular kind of technical text and the nature of the test itself. It is by no means self-evident that an index should contain all and only the lexicalized noun compounds of a text. Such a test is at best indirect, and the project should thus be viewed as preliminary work indicating the feasibility of the approach. Future work will test the generalizability and robustness of the methods by identifying other tests, such as using a glossary rather than an index, and applying the methods to corpora of different sizes and genres.

REFERENCES

Dunning, Ted. 1993. "Accurate methods for the statistics of surprise and coincidence." *Computational Linguistics* 19(1): 61-74.

Grefenstette, Gregory, and Pasi Tapanainen. 1994. *What is a word, what is a sentence? Problems of tokenization*. Third International Conference on Computational Lexicography. Budapest: 79–87.

Karttunen, Lauri, J-P. Chanod, Gregory Grefenstette, and A. Schiller. 1996. “Regular expressions for language engineering”. *Natural Language Engineering* 2(4): 305–238.

Cluster Analysis of the Newcastle Electronic Corpus of Tyneside English: A comparison of Methods

HERMANN MOISL
University of Newcastle
hermann.moisl@ncl.ac.uk

VAL JONES
University of Twente
jones@cs.utwente.nl

The Newcastle Electronic Corpus of Tyneside English (NECTE) project is based on two separate corpora of recorded speech, one of them collected in the late 1960s as part of the Tyneside Linguistic Survey (TLS), and the other in 1994 by the Phonological Variation and Change in Contemporary Spoken English (PVC) project. Its aim is to combine the TLS and the PVC collections into a single corpus and to make it available to the research community in a variety of formats: digitized sound, phonetic transcription, and standard orthographic transcription, all aligned and available on the Web.

We are currently developing a methodology to study NECTE from a sociolinguistic point of view, and have begun by looking at the one formulated by the TLS, which was radical at the time and remains so today: in contrast to the then-universal and even now dominant theory-driven approach, where social and linguistic factors are selected by the analyst on the basis of a predefined model, the TLS proposed a fundamentally empirical approach in which salient factors are extracted from the data itself and then serve as the basis for model construction. To this end, an electronic corpus was created from a subset of the data, and various cluster analysis algorithms were applied to it in order to derive social and linguistic classifications of the sample. Stability of classifications across different clustering methods was already a known theoretical problem. The clustering techniques available at the time, and still widely used today, are sensitive to factors such as vector distance measure, clustering algorithm, and the order in which data items are presented —different combinations of these factors typically yield different analyses of the same dataset. These effects were observed in the TLS classifications. In an experiment on artificial data sets Jones (1979) demonstrated that certain combinations of clustering algorithms are capable of imposing erroneous structure on data which was inherently unclassifiable (by design). The types of structurings derived were consistent with theory and observation; for example Ward’s method tended to ‘discover’ spherical clusters irrespective of the natural structure of the data and classifications were shown to be sensitive to input order of datapoints. These properties of clustering techniques raise at least two issues relating to validation of classifications:

- objectivity—to what extent does a given analysis represent the actual structure of the data, and to what extent is it an artefact of the clustering method?
- selection—upon what criteria does one choose among alternative analyses?

We seek an approach to classification which improves on the methods available at the time when the first analyses of the TLS data were conducted, that is, techniques as insensitive as possible to variation in the sort of parameters identified above. In this paper we consider as a candidate a method that had not been invented when the TLS was active —the self-organizing map (SOM). The discussion is in three main parts. The first part outlines the TLS methodology, the second describes self-organizing maps, and the third compares the consistency of the analytical methods used by the TLS with that of the SOM relative to the TLS phonetic data.

TLS METHODOLOGY

The TLS aimed to model the overall linguistic variability of an urban community, that of Tyneside in north-east England, and more specifically

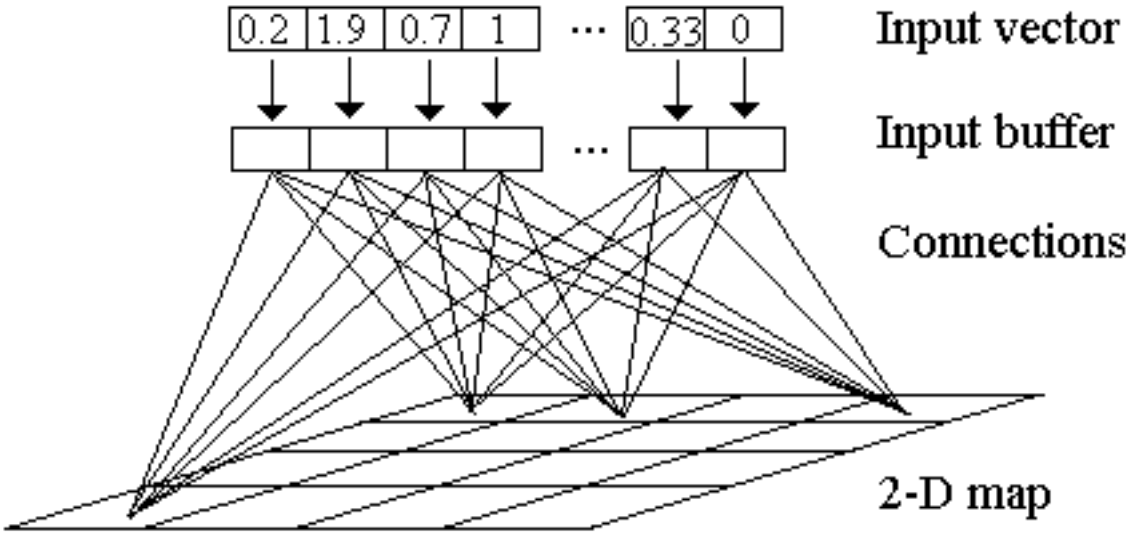
- to identify and exhaustively characterise the varieties of speech which co-occur in that area, and
- to determine the distribution of both the speech varieties and their constituent elements across the relevant social subgroups

To achieve this aim, a methodology was proposed which differed fundamentally from the one

standard at the time, and which was taken to have important theoretical and practical consequences for the discipline of sociolinguistics. In place of the theory-driven approach which characterized the work of Labov and Trudgill, among others, in which a sociolinguistic model was defined and then used to select a relatively small number of social and linguistic variables for analysis, the TLS was conceived as a set of methods whereby the salient features of linguistic and social diversity were to be empirically determined by, rather than presupposed in, the model. This methodology was designed to generate multiple candidate hypotheses about the data by applying high-dimensional multivariate analysis methods to it in different ways, from which those most useful to the aims of the TLS could be selected.

SELF-ORGANIZING MAPS

The self-organizing map, also known as the Kohonen net after its inventor, is a k-dimensional surface of processing units, where k is usually 2, together with a buffer into which input vectors are loaded. Associated with each unit is a set of connections from the input buffer such that, for a buffer of length n, there are n connections per unit, and each connection can take on a real-number value or 'strength' in some range, typically -1..1 or 0..1 (for clarity, only sample connections are shown in Figure 1):



Because it is an artificial neural network, the SOM is not explicitly configured or 'programmed' to behave in some desired way like a conventional computer, but rather learns its behaviour by exposure to input data using a learning algorithm. A full explanation of SOM learning would take us too far afield; details can be found in most textbooks on artificial neural networks. In outline, though, learning takes place by repeated presentation of vectors drawn randomly from a set V, and adjustment of the connections at each presentation. The SOM is initialized by assigning random values in some range (ie, 0..1) to the connections. When the first vector vi is presented, it activates the processing units to varying degrees, depending on the differences in connection strengths between the input buffer and each unit; the most highly activated unit uj is selected, and the connections are adjusted so that, next time vi is presented, uj will be even more highly activated than before, thus associating vi ever more strongly with a specific location on the map. As learning proceeds over—usually—many thousands of presentations, each of the vectors in V is associated with a specific unit in the map. After learning is complete, the entire set V is presented once again. Each vector activates the unit with which it has learned to become associated, and the result is a pattern of activations on the map surface. That pattern is significant: the distances among activated units represent the similarity relations in the input vector space.

COMPARATIVE STUDY

This comparative study confines itself to the phonetic-level representation of the TLS corpus. In order to apply cluster analysis to this data, the TLS had to represent it numerically. The method was as follows. For each of the 52 informants whose phonetic-level transcriptions had been digitized, the number of token occurrences of each of the 542 state types S defined in the transcription protocol was counted, where a 'state' is a discrete phonetic segment type. Each informant's phonetic profile was thus represented as a 542-element integer-valued vector V, in which any element Vi contained the number of token occurrences of state Si. The set of informant vectors was stored in a 52 x 542 matrix which, after normalization, served as input to the various clustering algorithms used in the analysis. The present study replicates the TLS data representation

and cluster analyses, and then compares the performance of a SOM on the same data, using a variety of settings for initialization of connections, sequence of input vector presentations, and map dimension.

CONCLUSIONS

Preliminary results from a relatively small subset of the 52 TLS informants indicate that the SOM performs as well as the other clustering algorithms in terms of its ability to identify and represent clusters, but that it is far less affected by variation in processing parameters. Results for the full TLS dataset will be available if and when this paper is presented.

REFERENCES

Everitt B. (1993) *Cluster Analysis*, 3rd ed. E. Arnold, London. 170 p.

Hair J., Anderson R., Tatham R., Black C. (1995) *Multivariate Data Analysis*, 4th ed. Prentice Hall, Englewood Cliff NJ. 751 p.

Jones-Sargent V. (1983) *Tyne Bytes: a computerized sociolinguistic study of Tyneside English*. P. Lang, Frankfurt am Main and New York. 368 p.

Kohonen T. (1995) *Self-Organizing Maps*. 2nd ed. Springer, Berlin and New York. 312 p.

Pellowe J. (1972) "A dynamic modelling of linguistic variation: the urban (Tyneside) linguistic survey." *Lingua* 30, pp.1–30.

Pellowe J., Jones V. (1978) "On intonational variety in Tyneside speech." In Trudgill P. (ed.) *Sociolinguistic Patterns of British English*. Arnold, London, pp.101-121.

Sargent V (nee Jones) (1979). "Cycles and the equal society." *Classification Society Bulletin*, 4:3, 31–45.

Strang B. (1968) "The Tyneside Linguistic Survey." *Zeitschrift für Mundartforschung*, Neue Folge 4, pp. 788–794.

Computational Generation of Limericks

GREG LESSARD
Queen's University
lessardg@qsilver.queensu.ca

FOREWORD

```
A program called VINCI from Queen's
Employs computational means
To make metre and rhyme
In right to left time
So limericks appear on your screens.

The people attending this session
We hope will have gained the impression
That these powerful tools,
With appropriate rules,
Can give rise to poetic expression!
```

BACKGROUND

There is evidence that the generation of at least some classes of humour is a rule-governed activity, formalisable in terms of algorithms, and implementable computationally in order to generate actual humorous utterances. In earlier work, using the VINCI natural language generation environment, we have shown this to be the case for several subclasses of verbal humour, including Tom Swifties and several classes of riddles. (See Lessard & Levison, 1992, 1993, 1995, 1997). Others have done similar work (see for example Binsted & Ritchie, 1997). This does not mean that we suppose the specific algorithms used in generation to exist as internal mental representations in speakers, but only that the productions observed in humans admit of formalisation.

It must be admitted, however, that puns and riddles form instances of what is often called verbal humour. This is typically non-narrative, based on relatively simple paradigmatic lexical relations, and usually evaluated in terms of cleverness rather than funniness. (See Lessard, Levison & Venour, 2002, for a

discussion of the distinction.) Instances of verbal humour are in many respects inherently simpler than more narrative structures like jokes. Limericks, on the other hand, presuppose at least a primitive narrative structure. As such, they provide a useful proving ground for more advanced work on both language generation (since it is necessary to take account of textual coherence) and humour generation (since narrative structure is involved).

LIMERICKS

Limericks may be characterised generally as five-line verses with the aabba rhyme scheme, one of a range of metrical foot structures, and some attempt at humour or at least cleverness. (For discussion, see Bibby, 1978). A typical example will illustrate the model:

```
There was an old man from Peru
Who dreamt he was eating his shoe
He awoke with a fright
In the midst of the night
And found it was perfectly true.
```

Taken as a subset of more general poetic language, limericks confront us with a number of practical problems, but also with a number of theoretical challenges when they are compared with prose-based natural language generation. For example, traditional approaches to text planning and generation tend to be driven by conceptual and syntactic rather than metrical structure. This is unsurprising when one considers that the object of most systems of the sort is to represent some encyclopedic or data-base-like description of some state of affairs in a discursive format. In such a context, the primary factors to be considered include the model of the user (see for example Paris, 1993), the overall structure of the text plan (see for example McKeown, 1985), paragraph-level phenomena (see for example Mann, 2002), and problems of reference (see for example Dale, 1992).

As a consequence, planning tends to be top-down (from universe of discourse, to dialogue model, to context, to overall text, to paragraph, to sentence) and many of the phenomena to be dealt with (anaphora, for example) can be examined in a left-to-right sequence. This has meant that conceptual tools such as phrase structure grammars or their variants have provided sufficient power to deal with the problems raised. Even phenomena such as freer word order may be captured in such systems by means such as the splitting apart of linear precedence and immediate dominance as in formalisms like Generalized Phrase Structure Grammar (Gazdar et alii, 1985).

The production of poetic text, on the other hand, requires that content be filtered by metrical and phonological factors such as the number of syllables, stress, and rhyme. A particularly interesting consequence of this fact is that generation must take account of the ends of lines in the production of the beginning and middle. For example, in the case of the limerick quoted above, we may represent the syllable structure as follows, where s represents an unstressed and S a stressed syllable:

```
There   was    an   old man    from Pe  ru
Who     dreamt he   was ea     ting his shoe
                He   a   woke   with a   fright
                In   the midst  of   the night
And     found  it   was per    fect ly  true
s       S      s    s   S      s    s   S
```

It can be seen that the first, second and fifth line contain an iamb followed by two anapests, while the third and fourth lines contain two anapests. In addition, rhyming lines (Peru-shoe-true) must end on the same stressed syllable nucleus and coda (to use the terminology of metrical phonology (see for example Hogg, 1987). Note that these are not necessarily lexical items: in the first, second and fifth lines, the sequences “his shoe” and “-ly true” rhyme with “Peru”.

GENERATION

A full analysis of the semantic, syntactic, lexical and phonological relations found in the limerick would take more space than is available here. However, from the generative perspective, it is possible to imagine three mechanisms to satisfy this constellation of constraints:

1. exhaustive generation: in essence, if we define a set of templates, we can generate large numbers of instances of potential limericks and ask a human to select those few judged to be of good quality. (This is the monkeys typing Shakespeare model, or Borges’ library.)
2. initial generation followed by tweaking: we may also begin by producing a sequence of lines and then selectively edit these to approach the target of an acceptable limerick, by means of

the addition, removal or shifting of items. This may be how human poets work: it is how we produced the limericks at the head of this abstract. (Valéry is said to have claimed that the first line of a poem was a gift of the gods and that the remainder was the product of the poet's craft and efforts.) Since the edit operations described above are well-known in the computational context, it is not impossible to imagine their implementation.

3. right-to-left generation: given the direction of the constraints in limericks, a possible computational model involves the selection of the right-most elements of a sequence of lines as a starting point and then within each line, the constraint-satisfaction generation of the intervening elements required to flesh out the entire limerick.

In fact, these right-to-left formal constraints must interact with top-down constraints which govern the semantic coherence of the text. In other words, in order for the limerick to be coherent, the narrative events which it includes must be consistent with the subject. These thematic constraints may be fairly loose, as in the Old Man from Peru poem, where there is no particular logic which attaches old men from that part of South America and the ingestion of footwear (although more generally, the poem respects the constraint that old men are humans and that humans eat food which is a subset of tangible objects, and that footwear is also a subset of tangible objects). On the other hand, they may be quite tight, as in the VINCI limericks, which capture a set of characteristics of the software (its origin, the architecture of the program). Compare as well Bibby's attempts to produce limericks appropriate to each of the Cambridge colleges.

The limerick thus represents the semantic expansion from one or more initial lexical choices, by means of the selection of traits or actions which are at least consistent with these. This presupposes a rich representation of the characteristics of each lexical item, including both high-level semantic traits such as that between humans, eating and objects, but also more encyclopedic 'information, such as the fact that VINCI is a program produced at Queen's. (See Peeters, 2000 for a discussion of some of the problems found at this 'lexicon-encyclopedia interface'). Note also that a sufficiently rich set of traits will provide the seed for variant limericks based on the same starting point. Consider for example the following example by Edward Lear:

```
There was an Old Man of Peru,
Who watched his wife making a stew;
But once by mistake,
In a stove she did bake,
That unfortunate Man of Peru.
```

At the formal level, problems arise both between lines (rhyme) and within lines (metrical structure). Consider the latter in the context of the initial old man from Peru poem. Let us assume that we have determined to use a metrical structure based on an iamb and two anapests and that we have selected the the place-name Peru. At this point, we have used up two of the eight available syllables. We also know that the next earliest syllable must be unaccented. The solution found here is "from". At this point, we know that we have a prepositional phrase "from Peru" which requires (among other choices) a noun phrase with the appropriate accent structure. This constraint is satisfied here by "an old man"). However, in order to verify that this noun phrase is appropriate, we must have knowledge of its metrical structure. For example, if the target were an iamb, we would need to seek, among other possibilities, a DET N structure. This application of constraints cascades right to left until the overall metrical target is met, within the semantic constraints already specified.

IMPLEMENTATION

In the conference presentation, we will illustrate the implementation of the approach described above using the VINCI natural language generation environment (see **http://www.cs.queensu.ca/CompLing** for details). Briefly, VINCI allows for the initial specification (PRESELECTION) of a constellation of lexical items which may be constrained by semantics, morphological and syntactic characteristics, and phonology. Preselected items, as well as all their characteristics, are available to subsequent steps of the generation process. Among other things, this allows for multiple layers of preselection, in which a first item determines a set of rhyme constraints, as well as a set of semantically appropriate elements.

Within each line of the limerick, and between lines, the VINCI mechanism of GUARDED SYNTAX allows the control of subsequent steps of the overall generation to be conditioned by the framework existing at that point. (A simple example of guarded syntax: if a noun phrase node carries the attributes 'first person' or 'second person', then its children may only be instantiated by a pronoun, whereas if the parent node carries the attribute 'third person', then the children may be pronouns, full noun phrases, or proper names.) Applied to the production of a limerick, and assuming right-to-left generation, if the parent node of a tree contains a prepositional phrase which sums to an anapest, then the next left-most chunk of syntax must inherit this information and construct an appropriate item from a library of possible patterns.

One consequence of this model is that syntax is reduced from a single overarching tree to the sum of a number of possible micro-trees. Such an approach is comparable to the model of Tree Adjoining Grammars, in which insertion and development occurs at the level of lexical-based subtrees (see Joshi & Schabes, 1997). It is also somewhat similar to the Labelled Deductive System approach used in parsing by Kempson et alii, 2001. Another consequence is that by precluding tweaking and editing, the system itself is forced to deal with the interplay of linguistic levels and constraints, without human assistance, thus providing an acid test for both the model and the implementation. Of course, this leaves aside the question of evaluation, itself a thorny issue, since the criteria used are many, varied and probably fuzzy as well. In related work, we are examining the ability of humans to learn how to produce limericks, as well as their ability to evaluate them.

REFERENCES

Bibby, Harold Cyril. (1978) *The Art of the Limerick*. Hamden, Conn. : Archon Books.

Binsted, Kim; Ritchie, Graeme. (1997) "Computational rules for punning riddles." *Humor*, 10 (1), pp.25–76.

Dale, Robert. (1992) *Generating referring expressions: constructing descriptions in a domain of objects and processes.* Cambridge, Mass.: MIT Press.

Kempson, Ruth; Meyer-Viol, Wilfried; Gabbay, Dov (2001) *Dynamic Syntax: the Flow of Language Understanding.* London: Blackwell.

Gazdar, G.; Klein, E.; Pullum, G.; Sag, I. (1985) *Generalized Phrase Structure Grammar*. Cambridge, MA: Harvard University Press.

Hogg, Richard. (1987) *Metrical phonology: a coursebook.* Cambridge: Cambridge University Press.

Joshi, A.K.; Schabes, Y. (1997) "Tree-Adjoining Grammars", in *Handbook of Formal Languages*, G. Rozenberg and A. Salomaa (eds.), Vol. 3, Berlin: Springer, pp. 69–124.

Lessard, Greg; Levison, Michael; Venour, Chris. (2002) "Cleverness versus funniness." April Fool's Day Workshop on Computational Humour, Trento.

Lessard, Greg; Levison, Michael. (1997) "Rule-governed wordplay and creativity." In *Mind II: Computational Models of Creative Cognition*, Dublic City University. **http://www.compapp.dcu.ie/~tonyv/MIND/greg.html**

Lessard, Greg; Levison, Michael. (1995) "Linguistic and Cognitive Underpinnings of Verbal Humour." International Cognitive Linguistics Association Conference, Albuquerque.

Lessard, Greg; Levison, Michael. (1993) "Computational Modelling of Riddling Strategies." ACH/ALLC Joint Annual Conference, Georgetown University, Washington, DC. (Extended abstract in conference proceedings, pp. 120–122.)

Lessard, Greg; Levison, Michael. (1992) "Computational Modelling of Linguistic Humour: Tom Swifties." ALLC/ACH Joint Annual Conference, Christ Church, Oxford. (Extended abstract in conference proceedings, pp. 175–178.)

Mann, W. (2002) RST Web Site. **http://www.sil.org/~mannb/rst/**

McKeown, Kathleen. (1985) *Text generation : using discourse strategies and focus constraints to generate natural language text.* Cambridge: Cambridge University Press.

Paris, Cécile. (1993) *User modelling in text generation.* Francis Pinter Publishers: London.

Peeters, Bert (2000) *The Lexicon-encyclopedia interface.* New York: Elsevier.

Constraint, Practice, and Interpretation

BETHANY NOWVISKIE
SpecLab, University of Virginia
bethany@virginia.edu

ANDREA LAUE
SpecLab, University of Virginia
akl3s@virginia.edu

STEPHEN RAMSAY
University of Georgia
sramsay@arches.uga.edu

GEOFFREY ROCKWELL
McMaster University
grockwel@mcmaster.ca

PART I

Game theorist Martin Shubik holds that the common, defining element in all games—serious and light-hearted alike—is their unwavering emphasis on the "role of the rules" in shaping interaction and practice. These three talks adopt a ludic perspective on the rules that govern the methods and goals of humanities computing—those sometimes unnoticed constraints influencing our theories, constructions, and interpretations. What are the rules of this particular academic game? And how might we leverage them to reveal new insights about our work, to enable new activity?

The first talk, "Llullian Method," is an overview of a medieval system for generating knowledge and provoking interpretation by mechanical means. Not only did Ramon Llull's Great Art employ a constraints-based logic akin to modern computation and linguistic or semantic analysis, but it also incorporated the interpretive faculties of its human user into its own algorithmic system in a manner currently advocated by theorists of hermeneutic approaches to information design. Bethany Nowviskie closes with a proposal for some experimental applications of Llull's visual and mechanical methodology to humanities computing.

The second presentation, "Rules for Reading," offers a discussion of the relationship of classical and contemporary narratology to structural and formal approaches to interpretation. The constraints and grammars long applied to the analysis of narrative are given new meaning through a shift from a deterministic to an interpretive role. Andrea Laue presents an experiment in the temporal markup of Conan Doyle's Sherlock Holmes stories by readers made sensitive to the role of time as a structural element in mystery narratives. She then demonstrates that visualizations of the reading process can illuminate interplay between narrative and interpretation.

Finally, Stephen Ramsay and Geoffrey Rockwell present what they characterize as a performance piece on the relationship between literary writing and computer programming. "Programming as Writing as Programming" suggests that the mastery and use of computer languages is not a pre-critical skill, but rather an interpretive—even artistic—act. By a process of "ludic inversion," literary writing is analyzed as an attempt to constrain or program readerly interpretation. Knuth's call for literate programming is taken up in the context of humanities computing and—like Nowviskie's combinatoric mechanisms and Laue's narratological visualizations—is offered as an example of the methodological application of rules and constraints to the interpretive work of the humanities.

PART II
LLULLIAN METHOD AND INTERPRETATION IN HUMANITIES COMPUTING

Bethany Nowviskie

When, according to one legend, the leaves of the lentiscus plants on the highest mountain of Majorca became marked with strange alphabets, Raimundus Lullus understood his divine charge: the creation of a Great—and symbolic—Art. Llull's Ars Magna, developed in the latter decades of the thirteenth century, has been recognized as a precursor both to computer science (in its emphasis on a mechanical calculus) and the philosophy of language (in its use of symbols and semantic fields).

Ramon Llull's colorful and quixotic life (from the episode in which he rode a horse into a cathedral to his near-suicidal martyrdom in North Africa) has served to detract attention from the seriousness of his project. The associations of his Great Art with alchemy, Kabbalism, and clashes between Muslims and Christians (in fact, the Ars Magna was designed to convert the infidel not through faith or grace but by means of irrefutable logic) sometimes make it difficult for us to extract Llull's methodology and the machines he designed to support it from an entangling network of rhetoric and myth.

Who was Llull, and what was his Great Art? More importantly, how should we understand the Llullian method in the context of our own work in humanities computing? This overview of Llull's generative and logical systems emphasizes their relevance to questions of interpretative practice, visual thinking, and algorithmic or constraints-based methodology in humanities computing today.

Llullian procedure was played out within a body of rules, designed to be as universal and eternal as possible and meant to be followed strictly, as they encoded a logic of interpretation. But the Ars Magna was not merely an abstract methodology; instead, Llull designed a logic of questioning and answering supported by mechanical systems—the first on record. Llull's Art constituted a text-based mechanism and a procedure for using and interpreting the results of textual manipulations.

Llull's primary device was system of interlocking, independently-mobile wheels, on which, in compartments or camerae, were inscribed letters symbolic of theological concepts. These represented the smallest principles of faith. For example, a small wheel might contain the scholastic transcendentals of unum (the one), verum (the truth), and bonum (the good)—an embodiment of the idea that nothing can exist without exhibiting unity, truth, and goodness. A more complex wheel might contain letters symbolizing the sixteen divine attributes: goodness (B), greatness (C), eternity (D), and so forth. Connecting lines drawn in specified patterns among the letters of such a wheel yield dozens of simple combinations, which Llull interpreted as revelatory (and yet logical) truths about the Divinity. We therefore see that the goodness of God is great (BC) and eternal (BD), or that the greatness of God is good (CB). The permutations, especially in more articulated wheels, some of which encode the objects of knowledge (angels, man, heaven, the imagination, stones, flames, plants) and rhetoric (what, whence, why, when, whether, where, how?), become exceedingly complex and allow for scientific as well as theological questioning: Where does the flame go when a candle is extinguished? Why does rue strengthen the eyes while onion weakens them?

Llull demonstrated the use of his Art for posing and examining difficult philosophical problems that had been taken up in other medieval contexts, such as: Can a fallen angel repent? Could God damn Peter and save Judas? Will the unborn child of a martyr be saved through a baptism of blood? In the books accompanying his charts and diagrams, Llull sometimes offered full arguments and commentaries on such questions, sometimes outlined the combinatorial process by which the questions could be answered using his wheels, and sometimes simply showed that such sophisticated questioning could be generated by means of the Ars Magna. In each of these cases, Llull's emphasis is as much on the interpretive interface between the machine and its user as on the embedded and generative logic of the wheels.

The ostensible purpose of the Great Art was to justify the Christian faith through mechanical, logical, and algorithmic analysis. This work, like the machines that supported it, was self-testing in the sense that the execution of iterative combinatorial motions was only carried out until contradictions emerged. At that point, the wheels themselves could be examined and reconfigured. In this way, Llull's Art was both a generative and autopoietic mechanism. Llull was interested not only in proving, through his logical visual grammar, the validity of old beliefs, but also in watching new truths and refined systems emerge.

Few among us would be comfortable suggesting that there are eternal truths in humanities computing or the disciplines to which it is applied, but a workable method for placing and analyzing old ideas in new combinations is always welcome. To this degree—and most especially because Llull's algorithms and machines all bend toward hermeneutic ends—it is interesting to consider their possible connection to problems of interpretation in humanities computing.

Are there currently established rules for producing and interpreting the results of humanities

computing practice? Is it desirable to formalize the unwritten rules and obscured components and principles of our hermeneutic work? How might thinking formally (and perhaps visually) about our methodology help to clarify what we do to our students, our colleagues, and ourselves? The example of Llull could prompt us to go a step further: can we imagine combinatorial, generative, algorithmic, and autopoietic systems that could enliven our thinking about the objects, methods, and aims of humanities computing? A method which once felt charged with the grandeur of God might now—in the aftermath of generative aesthetics, OuLiPian composition, ars combinatoria, deformance, and 'pataphysical speculations—seem to be infused with an attractive and experimental ludic spirit. Are there new uses for Llull's old machines?

REFERENCES

Bexte, Peter. "Ars Combinatoria: Zum Ursprung der Denkmaschine." in Klaus Peter Denker (ed). *Weltbilder / Bildwelten: Computergestütze Visionen.* Hamburg, 1995.

Chalmers, Matthew. "Hermeneutics and Information Representation." Draft. 2001. PDF. Available: **http://www.dcs.gla.ac.uk/~matthew/papers/hermeneutics.pdf**.

Eco, Umberto. *The Search for the Perfect Language*. Cambridge: Blackwell, 1995.

Gardiner, Martin. *Logic Machines and Diagrams*. Chicago: Chicago UP, 1982.

Künzel, Werner. *Die Ars Generalis Ultima des Raymundus Lullus: Studien zu einem geheimen Ursprung der Computertheorie*. Berlin, 1991.

Sales, Ton. "Llull as Computer Scientist; or, Why Llull Was One of Us." Universitat Politecnica de Catalunya, Barcelona. Available: **http://www.geocities.com/llull_brazil/compsale.html**.

Snodgrass, Adrian, and Richard Coyne. "Is Designing Hermeneutical?" *Architectural Theory Review* 1.1 (1997): 65–97.

Yates, Frances Amelia. *Lull & Bruno: Collected Essays*. London: Routledge, 1982.

Zweig, Janet. "Ars Combinatoria: Mystical Systems, Procedural Art, and the Computer." *Art Journal* 56.3 (1997): 20–29.

PART III
RULES FOR READING

Andrea Laue

In a century marked by literary narratives declaring and sometimes celebrating multiplicity and indeterminacy, several critical traditions have sought a single, stable structure to define narrative form. Structuralist narratologists developed systematic descriptions of narrative form, and computer-aided research in the humanities, with its empirically-based studies of the linguistic skeleton of narratives, extended this tradition. Researchers in both fields hypothesize that there exists a formal, rule-based structure that constitutes the minimum condition of narrative, a schema that can be extracted, studied and replicated. Although most narratologists no longer subscribe to this theory, much computer-aided research continues the search for a Platonic ideal. The purpose of this paper is twofold: to suggest new methods and goals for computer-aided studies of narrative and to demonstrate that visualizations of the process of interpretation might help us better understand narrative.

Contemporary narratologists propose a division between classical and postclassical narratology[1]. Current computer-aided study of narrative is grounded in classical narratology, a tradition informed by structuralism and interested in the elaboration of a universal structure of narrative. In building a description of the archetypal narrative, classical narratologists ask two basic questions: 1) what is the most basic unit of narrative; 2) what set of rules constructs a unified structure from those units? In an answer to this question classical narratologists think they will uncover the skeleton that characterizes all instances of what readers identify as narrative. These structuralist notions still obtain in most applications of computing technologies to the study of narrative. Such research often separates form and meaning; much research proposes narrative units that can be identified and quantified, rendering a static—even statistical—accounting of narrative form[2,3]. Classical narratological models of narrative are attractive—discrete units and formal rules are easily computable—but perhaps dangerous for our discipline.

Narratologists suspicious of structuralist notions of narrative form have proposed various new paradigms for the investigation of narrative[1,4]. Represented by a wide variety of approaches—feminist poetics to artificial intelligence—postclassical narratology in its most general sense involves an investigation of the methods by which readers invest strings of events with narrative structure. Drawing on cognitive models of reading, postclassical narratology posits narrative as a sequence of cues that promotes some structuring activity. The structuring activity may be rule-based, but the rules reflect preferences rather than unambiguous and universal forms[5,6]. In its shift from form to process, postclassical narratology questions common distinctions between reading and interpretation, arguing that the former is not prior to and free of the latter.

This questioning of the boundaries between reading and interpretation is crucial to postclassical narratology and to textual studies. Postclassical narratologists criticize structuralist models because the end is always already known—intermediate events are only interpreted in light of a known ending. The "retrospective" form that results ignores a crucial dynamic of reading—the projection an ending that makes sense of intermediate events, an ending that is constantly re-imagined as the intermediate events resist previous interpretations[7]. Thus reading might be understood as a time-based process of constant re-interpretation in light of an anticipated ending, and literary narrative as a textual collection of cues that trigger this process. Two decades ago Stanley Fish argued that we misuse the word "read," that we pretend that there exists a text that can be perceived prior to interpretation and that reading involves some unstructured experience of that text[8]. In this sense a text is only what we, as readers, make of it. Jerome McGann extends this argument in his theories of reading as deformance[9]. In both cases, the text is as much in the reader as it is in the artifact, and what we think of as structure emerges from the interaction of the two. This interaction is evidenced by the emergent text.

I begin my investigations of reading as interpretation with a study of clues in Arthur Conan Doyle's Sherlock Holmes stories. Although ostensibly about something—a crime—Doyle's fiction is literally about the (re)construction of something—the acts preceeding and constituting the crime. In its temporal inversion between the fabula and the sjuzhet, its incorporation of character-bound narrators and narratees, and its literalization of causal logic through the use of clues, Doyle's stories seem to perform narrative. Detective fiction has been central to structuralist studies of narrative, and many theorists of detective fiction argue that clues are the critical component of the genre both on a formal and on a socio-historical level[10,11]. I begin my study of the emergent text by looking at how interpreters structure a narrative around clues. I give readers a copy of Doyle's "The Red-Headed League" and ask them to mark points in the text at which they identify something as a clue. They use two colors to mark the text, one to indicate something that seems to be a clue when first encountered and another to indicate something that seems to be a clue only after reading further in the text. So readers mark the text the instant they identify some bit of information as a clue to the mystery, and they also have the option of flipping back and finding a detail that seems significant only after reaching some later point in the story. This layering asks crucial questions about the dynamics of reading detective fiction: How does anticipation of a particular type of ending contribute to the identification of textual cues and to the activation of preference rules that govern constructions of narrative?

Schematic visual representations of narrative structures are common in articles and books on narratological theory, whether classical or postclassical, and these graphs are most often intended as a visible skeleton of the story itself. Although several postclassical narratologists construct and utilize visualizations of narrative, none attempts to graph narrative as a process, to capture the discovery of textual cues and the enactment of preference rules. In my attempts to do just that, I graph all markings made by readers, not privileging the most recent—most "correct"—structuring. Following this method, I hope that I better capture the dynamic of reading, the constant processing and reprocessing that results from an anticipation of closure, the emergent text.

REFERENCES

Herman, David. 1999. "Introduction." Pp. 1–30 in *Narratologies: New Perspectives on Narrative Analysis*, edited by David Herman. Columbus, OH: Ohio State University Press.

Meister, Jan Christoph. 1995. "Consensus ex machina? consensus qua machina!" *Literary and Linguistic Computing* 10(4):263–70.

Snelgrove, Teresa. 1990. "A Method for the analysis of the structure of narrative texts." *Literary and Linguistic Computing* 5(3):221–5.

Ronen, Ruth. 1990. "Paradigm Shift in Plot Models: An Outline of the History of Narratology." *Poetics Today* 11(4):817–42.

Herman, David. 2002. *Story Logic: Problems and Possibilities of Narrative*. Lincoln, NE: University of Nebraska Press.

Jahn, Manfred. 1999. "'Speak, friend, and enter': Garden Paths, Artificial Intelligence, and Cognitive Narratology." Pp. 167–94 in *Narratologies: New Perspectives on Narrative Analysis*, edited by David Herman. Columbus, OH: Ohio State University Press.

Brooks, Peter. 1984. *Reading for the Plot*. Cambridge: Harvard University Press.

Fish, Stanley. 1980. *Is There a Text in this Class?* Cambridge: Harvard University Press.

McGann, Jerome. 2001. *Radiant Textuality*. New York: Palgrave.

Moretti, Franco. 2000. "The Slaughterhouse of Literature." *Modern Language Quarterly* 61(1):207–27.

Ginzburg, Carlo. 1984. "Clues: Morelli, Freud, and Sherlock Holmes." *The Sign of Three*. Edited by Umberto Eco and Thomas Sebeok. Bloomington: University of Indiana Press.

PART IV
PROGRAMMING AS WRITING AS PROGRAMMING

Stephen Ramsay and Geoffrey Rockwell

Discussions of programming in the context of the arts and humanities are almost always constrained to practical matters—the technical dilation of the development of coded instructions for processing by a computer. As such, programming would seem to fall into that category of pre-critical skills necessary for the pursuit of some higher, more avowedly humanistic endeavor. What hasn't been addressed, are the ways in which programming itself might be conceived as not merely an interpretive, but a “literary” practice. Nor has much attention been paid to a suggestive, if radical converse: that writing might be viewed as a programming practice. In this paper, therefore, we take an extreme view and think of programming as writing and writing as programming.

We believe the distinction between human-readable and machine-readable discourses—the usual axis upon which the difference between programming and writing is said to rest—obscures the crucial human-readable aspects of code. Code is written to be read and reread by the humans who create and maintain it. Programmers speak of “elegance” in programming style, and deride the unreadable “kludge.” Programming is sometimes intended to amuse and is often resplendent with jokes buried in comments, variable names, and control structures. Moreover, code is nearly always accompanied by some larger exegetical framework (the documentation) intended to illuminate its ways and means. For all these reasons, it is difficult to say how programming differs from the process of writing under the constraints of form, genre, and audience expectation.

Writing, by a process of ludic inversion, may likewise be viewed as a form of programming. Creators of written artifacts seek to encourage the emergence of certain readerly patterns while discouraging the emergence of others. The rhetorician's art, therefore, consists in elaborating the manner in which texts may be manipulated for the purposes of constraining the range of alternative textualities available to the reader or hearer. This implies that texts are forever threatened with the exigencies of unintended effects—alternative formations that are the result of the text's position within a universe of discourse that is by nature beyond the control of any authorial agent. The traditional terms of writerly engagement—genre, association, metaphor, theme—are also, and perhaps even originally, the tools of the rhetorician seeking to gain some programmatic control over the complex valences of language and meaning. The author, to put it forcefully (if at the same time metaphorically), tries to program the reader.

In this paper we provide a number of illustrations of the ways in which programming is a literate practice and writing a programming practice. This illustration leads up to Knuth's call for “literate programming,” which we will recast as a project particularly suited to humanities computing. Our argument takes the following general form:

1. In computing, programming is taught in text. This manifests itself in traditional practices from the “Hello, World!” trope to Perl Poems. We will survey a number of textual practices in programming and argue that these are not just playful oddments, but reflections of the centrality of reading to the practice of programming.
2. As important as the execution of code is to computing, it is the documentation of code that allows it to be maintained and used. Coding also happens in conjunction with other forms of writing from flow charts, specifications, to comments. Programming, like text, is not an isolated practice; it is an act that happens in a community and which must be maintained to have the desired effect. Therefore, programming is a form of dialogue with the machine, users, and other programmers. The textual practices that at first seem peripheral to programming are actually central to its professional practice.
3. Reading is the human parallel to machinic execution. When we read, we play with the text—performing an operation that uses the text as a script. When we write, we imagine and try to control the reception of a text. As an example one can look at the role of rhetorical punctuation and technique in scripting oral performances of written works.
4. The history of programming is connected to the history of formal languages and an attempt to write things that can be reliably interpreted. This goes back to Plato’s Phaedrus and concerns about the interpretation of texts that are detached from the oral dialogical support of an academy. In the philosophical search for a universal language that cannot be misinterpreted, philosophers and mathematicians developed the logic from which programming evolved. Likewise, one of the most fascinating debates today about code revolves around its ontological status as both text and machine. This is not just a theoretical debate. When people began to publish books that contained encryption algorithms, one side claimed that that portion of the book qualified as a machine (and was therefore subject to patent and export

restrictions). The more libertarian side wanted to claim that it was merely text (and therefore subject to laws allowing freedom of speech).

Literary programming (or writing code) is a practice with a unique place in humanities computing. It is the purest form of the hybrid work (in a tradition from encoded texts to multimedia works) that is characteristic of our hybrid discipline.

REFERENCES

Topics on Computer Programming. Cat's Eye Technologies. <**http://www.catseye.mb.ca/esoteric/**>

Estelle. "Tampering with the Text to Increase Awareness of Poetry's Art. (Theory and Practice with a Hispanic Perspective) *LLC* 11 (1996): 155–62.

Donald Ervin. *Literate Programming*. Stanford: Center for the Study of Language and Information, 1992.

Wall, Tom Christiansen, and Randal Schwartz. *Programming Perl* 3rd. ed. Sebastopol: O'Reilly, 2000.

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

Applications of the Open Archives Initiative

MARTIN HALBERT
Emory University
mhalber@LearnLink.Emory.Edu

JOANNE KACZMAREK
University of Illinois
jkaczmar@uiuc.edu

DAVID SEAMAN
Digital Library Federation
dseaman@clir.org

STEPHEN SCHWARTZ
University of California
shs@library.ucla.edu

The application of computing solutions is regularly applied to the matter of assisting scholars with their research and analysis of findings. Computing technologies have become increasingly important in libraries and archives as these institutions strive to make the many unique resources they hold more easily accessible to their users. As information institutions begin to exploit digital technologies an often under appreciated distinction is the one between digital content creation and digital library services. Aggressive programs and initiatives to digitize primary source content and archival finding aids have resulted in immeasurable piles of sometimes interesting, sometimes trivial, heterogeneous, unorganized data, leaving the user in a state of "information overload". Digitization efforts are typically organized at a local level and scholars elsewhere are often unaware that information resources of possible interest have been digitized. What's been lacking are the infrastructure technologies and protocols that support the implementation and development of digital library services which organize and make more useful the array of diverse and widely distributed collections of digital content and descriptive metadata that have been and are continuing to be created at a rapid pace.

This panel session will discuss experiences so far with the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH), an emerging digital library protocol designed to enable the implementation of new and innovative online services and to enhance and facilitate interoperability and the ways in which scholars can discover and access unique information resources wherever they reside. Each participant will provide a brief overview of their involvement with OAI followed with a discussion of the infrastructures and techniques used by various projects to explore the scalability and extensibility of the protocol.

OAI-PMH BRIEF BACKGROUND

The Open Archives Initiative (OAI) **http://www.openarchives.org** was launched with a meeting of digital librarians and computer scientists specializing in archiving, metadata, and interoperability in October 1999. The shared objective of this group was to pave the way for universal public archiving of the scientific and

scholarly research literature on the Web. What was needed were conventions that would allow any paper deposited in any public archive to be found from anyone's desktop, worldwide. This was essentially a model for a virtual public library. The group recognized there may be many different archive initiatives likely to emerge from this vision, each presenting the possibility of different conceptual, organizational and technical foundations. The issue of maintaining the archives as "open" would be crucial for these initiatives to successfully become part of the scholarly communication system and as such, technical issues concerning interoperability would be crucial. From these original discussions it was determined that using a harvesting approach to aggregating metadata describing the resources held by archives would be simplest to implement and more scalable as more archives become available. From this emerged the OAI Protocol for Metadata Harvesting (OAI-PMH), **http://www.openarchives.org/OAI/openarchivesprotocol.html**, first released in January 2001. Since then it has emerged as a practical foundation for interoperability among digital library efforts. OAI-PMH includes two distinct yet simple components. At one end, data providers use the OAI-PMH to expose metadata in a variety of forms. At the other end, service providers use it to harvest the metadata from the data providers and subsequently process it, adding value in the form of services like domain specific portals to specific user communities.

With an interest in exploring the sustainability and extensibility of the protocol, the Andrew W. Mellon Foundation announced the funding of seven service provider (harvesting) projects in July 2001. The Digital Library Federation has generously provided organizational support for these OAI activities. Several of the projects focus on aggregating materials of possible interest to scholars seeking cultural materials. One of these projects is the Cultural Heritage Repository at the University of Illinois, **http://www.oai.grainger.uiuc.edu**. Emory University, home to two of the seven original Mellon-funded projects, also focuses on cultural materials, **http://MetaScholar.org**. The UCLA Digital Library Program is host for the Sheet Music Consortium project, **http://digidev.library.ucla.edu/oaisheetmusic/**. These projects provide the OAI-PMH backdrop for this panel session, outlining details of the technology from the service provider perspective and ultimately end-user utility of their portals.

DIGITAL LIBRARY FEDERATION

The Digital Library Federation (DLF) is a consortium of libraries and related agencies that are pioneering in the use of electronic-information technologies to extend their collections and services. Through its members, the DLF provides leadership for libraries broadly by identifying standards and "best practices" for digital collections and network access, coordinating leading-edge research-and-development in libraries' use of electronic-information technology, and helping start projects and services that libraries need but cannot develop individually. One such supported project is the Open Archives Initiative. The DLF support takes two forms. Financially the DLF provides some organizational support for OAI. The DLF is also responsible for organizing the Mellon-funded OAI projects that are evaluating the OAI technical framework.

EMORY UNIVERSITY

Emory University Library, in collaboration with SOLINET/ASERL, has developed two scholarly portal services based on metadata harvested from a variety of archives and digital library projects. The MetaArchive.Org and AmericanSouth.Org portals represent the first national projects to deploy the ARC software developed at Old Dominion University and the Open Digital Library (ODL) software developed at the Virginia Tech Digital Library Research Laboratory. These two portals, operating under the umbrella name of the MetaScholar Initiative **http://MetaScholar.Org**, are designed to engage scholars, librarians, and archivists in the process of making a large body of metadata useful in scholarly research.

Twenty institutions (primarily located in the geographic South) have initially contributed metadata to the central union database underlying both systems. ASERL participants established OAI-PMH data provider systems to dynamically serve out updated metadata from their local digital archives. Many libraries in the MetaArchive project are smaller four-year institutions which will make use of metadata conversion services provided by Emory in order to make their local finding aids and archival databases available for harvesting. Initially, metadata was collected representing primary research materials from a set of subject domains including the culture and history of the American South, papers of major political figures, and religious institutional records.

An interdisciplinary team including faculty from many research institutions around the country was tasked with evaluating the American South portal. This Scholarly Design Team was chaired by Dr. Charles Reagan Wilson of the University of Mississippi, editor of the Encyclopedia of Southern Culture. The team was also charged with exploring several new approaches to scholarly communication.

SHEET MUSIC CONSORTIUM

Johns Hopkins University, Indiana University and the UCLA Digital Library Program have joined forces to create a virtual catalog of sheet music in the United States called the Sheet Music Consortium. This project is using the OAI-PMH to harvest descriptive metadata from local databases and make it accessible through an

OAI searchable repository hosted by the UCLA Digital Library Program.

Consortium member institutions have chosen to catalog their sheet music in different ways, but a very large proportion of the original sheets in participating collections have been digitized, allowing users direct access to the music and—in many cases—covers and advertisements that offer evidence of the cultural context in which the songs were published. The first phase of the project aims to establish the service as a gateway to these US collections, and contains over 60,000 records from Indiana University, Johns Hopkins University, UCLA Music Library's Archive of Popular American Music, and records from the Library of Congress.

UNIVERSITY OF ILLINOIS

The University of Illinois has developed a vertical, domain-specific digital library portal designed to search metadata describing manuscript archives and digital cultural heritage information resources. Metadata describing non-digital resources and resources of restricted availability is included along with metadata describing publicly available digital objects. The Illinois team has explored "best practices" for using the OAI-PMH to reveal resources contained in hierarchical metadata structures such as those expressed in archival finding aids displayed with Encoded Archival Description (EAD) metadata.

Materials in the Illinois' Cultural Heritage Repository include harvested metadata from the Library of Congress American Memory Heritage Collections and from the Online Archive of California. The American Memory and Online Archive of California resources are made searchable via their metadata alongside other materials gathered from over 39 institutions. Total number of discrete items searchable is over 2 million and represents public and research libraries, museums, archives, historical societies, and digitization projects focusing on cultural heritage materials.

Integrating TEI and EAD to Create Usable and Re-usable Archival Resources

SUSAN HOCKEY
University College London
s.hockey@ucl.ac.uk

ELIZABETH HALLAM-SMITH
Public Record Office of Great Britian
elizabeth.hallam-smith@pro.gov.uk

ANNA SEXTON
University College London
a.sexton@ucl.ac.uk

CHRIS TURNER
University College London
c.turner@ucl.ac.uk

The LEADERS (Linking EAD to Electronically Retrievable Sources) Project follows the terminology of the archives and records community in using the term 'archives' to indicate a sub-set of records which are preserved because they have long-term value. Records are created by organisations or individuals in the course of professional or personal activities and they form evidence of the activities which gave rise to them. Archival records provide a means of understanding human history, they are a basis for corporate memory and a source of community and personal identity and offer pathways to learning for people of all ages and backgrounds. (National Council on Archives, 2002)

The work of the archivist is concerned with providing the means through which individuals can access archives. Users of archives need to be provided with access tools that will describe the contents of archive collections. Such access tools are often called 'finding aids' and they are produced through a process of 'capturing, collating, analysing, and organising any information that serves to identify, locate and interpret the holdings of an archival institution and explain the contexts and record systems from which the holdings were selected' (Definition from the Society of American Archivists).

Encoded Archival Description (EAD) has brought many benefits to custodians and users of archives. It has provided a means for archivists to structure finding aids using technology that is independent of proprietary hardware and software platforms and it has enabled Internet delivery of those structured finding aids. However, as a stand-alone tool, it cannot give users access to the actual content of archival material. Therefore, there is real potential for the development of a resource that enables the integration of EAD with other tools that do allow for remote use of the contents of archival collections. The Text Encoding Initiative (TEI) enables electronic texts of all kinds to be searched and presented to users in a variety of different ways. When EAD and TEI are brought together alongside digitised images of archival material the potential benefits for users are vast. Within a single environment the user can find items in archival collections; learn about their contexts; view representations of the items themselves; and read, study, analyze and manipulate their content.

At present the two encoding systems, EAD and TEI, are often used independently of each other and no generalised environment exists for linking them. Where links do occur in online archival finding aids, they normally point to digitised images of transcriptions without provision for any analysis or manipulation of the source. The LEADERS Project is developing a generic computer-based toolset that can integrate EAD encoded finding aids with TEI encoded transcripts and digitised images of archival material. This toolset will enable users to link directly from finding aids to digital documents and to manipulate those documents in various ways.

This session will concentrate on three aspects of the LEADERS Project's research. The first paper will discuss how the needs of the user community have been taken into account by LEADERS. This discussion will be set in the context of user needs evaluations conducted at the Public Record Office (United Kingdom) with a focus on how the LEADERS work on users is relevant to the PRO and to the archival profession as a whole.

The second paper will highlight the presence of overlaps between EAD and TEI. It will define the problems that these overlaps can cause in a system that seeks to bring the two frameworks together, and will explore the solutions that are being employed by the LEADERS team. The final paper will explore the technical choices that have been made in building a demonstrator application from the LEADERS toolkit, in particular it will focus on how the TVS (Transport Validation and Services) Model has been used as the basis for the system's architecture.

FOCUSING ON THE NEEDS OF THE END USER COMMUNITY

This paper shows how the theoretical and practical work of the LEADERS project is supporting service developments at the Public Record Office (PRO), the National Archives of England, Wales and the United Kingdom, and, with an annual 300,000 on-site visits and 80 million page downloads from its website (**http://www.pro.gov.uk**) the most heavily used archival resource in the world. Since 1994 the Public Record Office has re-invented itself as a customer focused organization, and since 1998 has held the prestigious Chartermark (recognition from the UK Government indicative of high quality services). Service enhancements introduced in response to the stated wishes of customers include extended opening hours, the provision of more expert services, faster retrieval times for documents, opening the research library to readers, self service scanning, reader orientation tours and a cybercafe. New online services which have evolved since 1995 include the flagship PROCat, a sophisticated EAD-based archive catalogue containing 9.3 million items; PRO Online which delivers images of popular records; the 1901 Census Online website; searchable research guides; and online educational resources for adults and school students.

PRO users are consulted regularly through on-site and on-line surveys and through a number of focus groups and panels, and feedback from comments forms, complaints and new user surveys is regularly tracked and fed back into any immediate remedial action and programmes of service development. The system is well-established and effective for onsite surveys; because they were in the main cutting edge and designed to grow and cater for new markets, the PRO's e-services were to a degree speculative in their design (although the award-winning Learning Curve for schools relied on feedback from teachers and school pupils from its inception). The PRO is now beginning the process of further consultation and will re-design many parts of its site to meet the information-seeking needs of its online users more closely. Here the findings of the LEADERS user survey, which it is currently running, will be invaluable.

Major consultations on what users regard as priorities for service developments for the future have since 1998 produced consistent results: topping the wish list is access via high quality catalogues to the images of the documents and with them, searchable transcripts. UK government guidelines and the rules of funding bodies such as the New Opportunities Fund have in addition since 2001 mandated full transcriptions (and where appropriate translations) of documents to appear alongside the images in online learning resources. In large-scale imaging projects this approach would however add exponentially to the costs of creating the resource. Hence the PRO is not at present providing full transcripts in its PRO Online website,

and nor is this currently linked directly to PROCat. The LEADERS work linking EAD and TEI will consequently be very valuable, since it offers a toolkit to link directly and enhance the functionality of these two online resources in line with stated customer priorities.

The PRO has generally in the past followed a simple segmentation model categorizing users as personal interest and family history researchers; professional researchers; the educational sector; and others. Since 1999 it has worked with the Public Services Quality Group for archives and local studies in carrying out an annual survey of archive users across the UK. Here are the results for 1999 in terms of research segments; these results are typical.

Type of User	**% of Use**
Personal interest including family history	76%
Educational sector	9%
Professional researchers	7%
Other	8%

Table 1: Research Interests of UK Archive Users in 1999

Since 2001 the PRO has evolved this segmentation model as the basis for tailored service development, the user segments being academics, family historians and personal interest researchers, people in formal education and leisure historians, many of this last category are non-users (PRO Keeper's Report, 2002).

The LEADERS work on the segmentation of the user market and evaluation of user needs will however add a new layer of sophistication which will help the PRO to fine tune its service developments more closely. The current LEADERS questionnaire segments researchers under education/training, personal leisure, professional/occupational, personal obligation and others—the professional and personal obligation data correlated with research topics will enable the PRO and other archives to tailor research leaflets specifically to the very precise needs of people in some microsegments, where appropriate. In addition LEADERS examines research interests under names, places and topics, and looks also at levels of familiarity with research interest—even the initial findings supporting the relative popularity of topic-based research suggest that the archival community may need to rethink its approach to the front ends of knowledge management systems and resources.

How does the LEADERS research fit within the broader framework of academic work on information-seeking behaviour? A number of academic studies on how users of information access resources online have been undertaken on both sides of the Atlantic. A key issue in this debate is the changing expectations of archive and information users in a digital and web-enabled age (Cox, 1998). Here there is much work available on information seeking behaviour by users largely of digital libraries, but encompassing obstacles to access to both library and archive collections (Seaman, 1997; Peterson Bishop, 1998; Mates, 2000). But although Margaret Hedstrom has published an important clarion call to make archival resources on-line more accessible to users (Hedstrom, 1998), this is unusual: most such work in the archive-specific domain is detached from the real world of the coal face and focuses on narrow issues. While some useful findings emerge from it, it must be said that at times these are obvious from the point of view of the practitioner.

For example, one recent study expresses disappointment that many online users display a 'disappointing reluctance' to use sophisticated search mechanisms (Large et al, 1999); a second that many archive users are searching for a specific item (Bearman, 1989/1990); a third establishes that the vast majority of users of photographic collections want to be able to search and retrieve images by subject (Collins, 1998); and a fourth, that searchers in online catalogue want indexes, glossaries and help functions to help them navigate through hierarchically structured material (Dugg & Stoyanova, 1998). Others establish that using search engines to try to locate archival finding aids usually produces unmanageably large sets of results (Feeny, 1999; Tibbo & Meho, 2001). All this supports the conclusion that many on-line archival finding aids do not take their users' needs sufficiently into account (Rosenbusch, 2000).

This is where the LEADERS work in extrapolating from the results of research on user needs to the building of a practical application is of such significance, since it feeds back the fruits of academically rigorous methodology into the development of a demonstrator system and toolkit which will have a practical use and value to the archival community in the UK. Once this demonstrator is built it will be further trialled in a real environment and modifications made to take account of changes both in technology and the ever-expanding expectations of users.

The archive profession in the UK will also be able to use the results of research based on the customer segmentation model to help it tailor services more precisely to the needs and requirements of each subset. For Access to Archives (A2A), for example, an EAD-based catalogue of archival series drawn from numerous repositories across England and currently containing 4 million records from 199 record repositories, the

findings will enable existing work with the user panel to enhance searchability focused on the needs of each user segment to be further developed and refined (**http://www.A2A.pro.gov.uk**). And the toolkit integrating TEI and EAD will support the enhancement of A2A by adding digitized searchable texts and images to some of the series. Similarly work to integrate and enhance the website of the new National Archives of the UK, bringing together the Public Record Office and Historical Manuscripts Commission from April 2003, will focus on user needs and search patterns and will be able to use the results of the LEADERS research alongside the views of user panels and online surveys.

INTEGRATING EAD AND TEI: THE RESOLUTION OF METADATA OVERLAPS

The two encoding systems, EAD and TEI, were designed to serve different purposes, yet there is a degree of overlap between them. In the context of archival material we can begin to explore the overlap by considering the TEI and EAD in relation to the original archival document, as both encoding systems seek to represent the original in different but convergent ways.

The archive profession in the UK will also be able to use the results of research based on the customer segmentation model to help it tailor services more precisely to the needs and requirements of each subset. For Access to Archives (A2A), for example, an EAD-based catalogue of archival series drawn from numerous repositories across England and currently containing 4 million records from 199 record repositories, the findings will enable existing work with the user panel to enhance searchability focused on the needs of each user segment to be further developed and refined (**http://www.A2A.pro.gov.uk**). And the toolkit integrating TEI and EAD will support the enhancement of A2A by adding digitized searchable texts and images to some of the series. Similarly work to integrate and enhance the website of the new National Archives of the UK, bringing together the Public Record Office and Historical Manuscripts Commission from April 2003, will focus on user needs and search patterns and will be able to use the results of the LEADERS research alongside the views of user panels and online surveys.

Original archive records can be described as ‘objects of study’. They provide evidence of and information about the functions and activities that gave rise to them and are therefore used by individuals to aid learning and research of varying kinds. When archival material is classified in this way, then the purpose of EAD is to describe those objects of study. It is a metadata standard designed to provide a structure for holding data about the original material.

The archive profession in the UK will also be able to use the results of research based on the customer segmentation model to help it tailor services more precisely to the needs and requirements of each subset. For Access to Archives (A2A), for example, an EAD-based catalogue of archival series drawn from numerous repositories across England and currently containing 4 million records from 199 record repositories, the findings will enable existing work with the user panel to enhance searchability focused on the needs of each user segment to be further developed and refined (**http://www.A2A.pro.gov.uk**). And the toolkit integrating TEI and EAD will support the enhancement of A2A by adding digitized searchable texts and images to some of the series. Similarly work to integrate and enhance the website of the new National Archives of the UK, bringing together the Public Record Office and Historical Manuscripts Commission from April 2003, will focus on user needs and search patterns and will be able to use the results of the LEADERS research alongside the views of user panels and online surveys.

TEI on the other hand is primarily a content encoding framework for the creation of new ‘objects of study’. The objects of study created by TEI are usually derived from one or more other (original) objects of study. When TEI is used to encode archival material, the archive document is the original object from which the new object is derived. This means that in order for the TEI object to be understood the encoder must not only transcribe and encode the contents of the original, but must also provide metadata that can put the text in context. This metadata, provided in the <teiHeader>, must describe the newly created object of study (e.g. provide a description of the electronic file and the encoding process) but must also describe the original object. Furthermore, the actual encoding that surrounds the data within the TEI object can also be viewed as a kind of metadata because the aim of the encoding is to delimit parts of the text and describe what those delimited parts represent. Therefore, although the TEI is primarily concerned with content encoding, it also must include metadata to put the object into the context of why and how it has been created, what it has been derived from, and what the data within the objects represents. Therefore, the overlap between the two frameworks occurs in relation to metadata that:

- Identifies, locates and gives details about the creation of the original object
- Describes the physical characteristics of the original object
- Provides contextual information about the creator and the participants within the original object
- Interprets/describes the data in the object

The general solution to areas of overlap employed by LEADERS is the development of a Schema which contains namespaces to reference the relevant parts of the respective EAD and TEI DTD. It makes sense for us to work on the general principle that metadata about the creation of the electronic transcript is best described using <teiheader> elements, whereas information relating to the original object (which is the source of the TEI transcript) is best described using EAD elements.

However, one particularly contentious area for consideration in the integration process is where metadata for interpreting and categorising the data within the object is held. When the basis for a TEI encoded document is a 'literal' transcription of an original source then metadata that seeks to categorise and interpret the data within the text is equally applicable to the original document and the new electronic transcript. There is no clear dividing line between the two, and so the general principle being advocated by LEADERS becomes difficult to apply. Furthermore, both TEI and EAD offer the encoder a range of options over where and how such categorisations and interpretations are made. This flexibility means that as well as comparing EAD and TEI against each other, they must also be examined individually to assess the variety of encoding options contained within them. Conducting this assessment is vital because the decision over where to place this metadata will ultimately have an impact on search and retrieval possibilities across the EAD finding aid and the TEI transcript.

Within a TEI encoded transcript it is possible to mark up the content of a text within a range of different elements. The degree of content encoding and the range of elements used to do so varies according to the purpose behind the encoding process. In the development of our demonstrator application we have taken a preliminary decision to explicitly encode all instances of names, dates, places, and publications as they appear in the documents we are using as test bed material. This explicit markup allows the content within the elements to be used as index terms for the transcribed documents, and it opens up possibilities for various types of user analysis to be conducted across the content marked up within the elements.

As well as being able to categorise and interpret data by enclosing it in specific elements within the flow of the text, TEI also allows the encoder to categorise and classify the content using elements provided by the <teiheader>. Index terms can be created (preferably using a recognised classification scheme/taxonomy) and placed in the <keywords>, <classcode> or <catref> elements (within <textclass> in <profileDesc>). Furthermore, if the encoder has included the additional tagset for language corpora in the DTD then other classification elements such as <channel> and <domain> also become available.

There is a degree of overlap between content encoding that can occur within the text and categorisations that can be placed within the <teiheader>. This overlap leads to a consideration of whether it is acceptable or useful for the same information to be tagged up both in the header and the flow of the text. One argument may follow that the dual functionality offered by content encoding within the text makes it a more obvious choice for containing categorisations and interpretations on the data. If for example, every instance of a name is marked up within the TEI <name> element, then the name element can act as an index term and can also be used in user analysis. If names are tagged up in this way what advantage is there in having the name repeated again as an index term within the header? This point is particularly relevant to documents like a parish register where there are so many names occurring within the flow of the text, that transferring such information into the header would be an arduous task.

However using the content encoding for indexing is not altogether straightforward. Encoding index terms within the flow of the text becomes more complicated when you want to assign a term to a wide span of data that overlaps with other hierarchical divisions. For example, a particular subject classification may span over several paragraphs. Although various encoding techniques can be used to overcome the problems of overlapping hierarchies it may be easier simply to put the classification term in the <header> rather than trying to apply it to a particular chunk of data. Furthermore, the dual functionality of content markup demands that every single instance of the data types that have been selected for special encoding will be marked up. For example, if the encoder has decided to mark up names within the text, every single instance of a name will be tagged. If a particular person has been referred to several times in a single document then each instance will be enclosed in a name element. If the content encoding on names is acting as the document's index terms then the implementer must consider how a search engine should deal with documents with multiple instances of the same name within them. Care must be taken to avoid the situation where the same document appears on the user's hit list because it contains multiple instances of the search term used.

EAD has a similar dual approach to metadata that interprets and categorises data within the original documents. Within EAD, index terms into the finding aid can be provided through a <controlaccess> element. However, index terms such as names, dates, and geographical locations can also be marked up within the narrative of the finding aid descriptions when they occur in descriptive content holding elements such as <scopecontent> and <bioghist>. However, the argument for encoding index terms within the flow of an EAD finding aid is unconvincing. This is because such content encoding in EAD does not have the dual function of also opening up possibilities for user analysis and manipulation of the data within the markup, as is the case with TEI. The reason for this goes back to the fact that an EAD finding aid describes objects of study and is

not an object of study in itself, and therefore there is no demand for close analysis of the finding aid as a 'text'. When the argument of dual functionality is non-existent then enclosing index terms within a specific element like <controlaccess> seems preferable.

However, this issue is complicated by the hierarchical nature of the EAD framework. An EAD encoded finding aid is organised into multiple levels of description where the collection is first described as a whole, and then in component parts which get more specific at each level of description. At the lowest level (item), it is the archive record that is being described. The levels of description within EAD follow the rule of inheritance where the lower levels can 'inherit' descriptive data placed above them in the finding aid tree. If 'inheritance' is applied to index terms within the <controlaccess> element then it follows that terms should be placed at the highest level of description where they can be said to apply to that level and all levels below them. The rule of inheritance then dictates that all lower levels of description have their own <controlaccess> terms plus all the terms placed above them in the descriptive hierarchy. In practice this means that a sophisticated search engine needs to be employed to traverse the hierarchy and map the relationships between index terms and levels of description. No real research has been conducted into the effect or usefulness of levels of description within EAD on index terms and their searchability and so it is an area that LEADERS needs to explore further. Our research is complicated further by adding another layer to the problem which is the effective searchability across both EAD and TEI as an integrated unit.

Solutions to overlaps between EAD and TEI and our final integration method must avoid repetition of information. Archivists and other related professionals will not appreciate repeating information that has been recorded in one place a second time. Furthermore, the potential for confusion in a system that is trying to integrate the two encoding frameworks is increased when the same data is held in different places according to different principles. It is also important that both the EAD finding aids and the TEI transcripts are not integrated to the extent that one cannot be re-used independently of the other. Each should be able to be recovered and used in other systems/applications for other purposes as stand-alone digital objects, otherwise the potential for the re-use of data that comes when working with non-proprietary tools will not be fully exploitable. Finally, as a primary use of metadata is to facilitate 'resource discovery', our integration solution must support meaningful search and retrieval and presentation of results.

DEVELOPING A GENERIC TOOLKIT: ARCHITECTURE AND TECHNOLOGY ISSUES

The LEADERS project is charged with developing a generic XML-based toolkit for use on multiple projects and with a wide variety of archival source materials encoded using EAD and TEI. This paper discusses the issues involved in designing and implementing such a generic toolkit.

The generic goal has been a key influence on the criteria applied to the technical choices made on the project. When considering a generic development as opposed to a one-off project, criteria such as availability, support, sustainability and issues of flexibility versus standardisation become relevant. The relative immaturity and the openness of the XML environment, has lead to a proliferation of tools and utilities developed and supported by individuals or ad hoc groupings. Whilst these may be both innovative and acceptable for use on a one-off project, the requirements for a generic toolset mean that we need to focus on products which are within the technical mainstream and have the backing of a stable organisation.

With regard to standardisation, both EAD, and more fundamentally the TEI have been designed for maximum flexibility. The assumption behind the design of TEI and supporting technologies, and to some extent behind EAD as well, has been that these tools would be used on single projects with a particular aim or objective and a homogenous set of source materials. This flexibility is desirable when viewed from the perspective of supporting the widest possible use on the widest range of individual projects. However the consequence is that each project using the TEI needs to define its own DTD and make its own rules for interpreting the TEI when encoding. For a generic toolkit, the rules need to be tightened up so that the tools for transforming and exploiting the resulting encoded material can be standardised.

The project has also had to choose between the use of Schemas and DTDs, in particular in view of the need to combine TEI, EAD and the NISO MIX Schema for the visual images associated with the encoded materials. Unlike DTDs, Schemas offer the use of namespaces, central to the combination of schemes required by LEADERS. Also Schemas support data type validation, a central requirement to support re-usability of the encoded materials and associated stylesheets and applications. Having chosen to develop a schema, different Schema languages such as RELAX-NG, Schematron and W3C were assessed and with reference to previously published evaluations and applications of the tools in question. A basic feature comparison was undertaken, and previous experiences evaluated within their contexts. Most experiences relate to the use of tools on a specific project or series of projects. For LEADERS the generic aspect takes on a major importance. We have to be conscious of the fact that our toolkit is designed for others to use in a range of differing circumstances. Therefore criteria such as sustainability, support tools available and compatibility with other areas of the XML family of standards play a more important role in selection than

basic features and functionality.

In addition to the toolkit, LEADERS is creating a demonstrator application to exemplify the benefits of combining TEI and EAD and support the re-usability of the encoding. The Transport, Validation and Services (TVS) model is being adapted and used as part of the technical architecture for LEADERS.

The TVS model (Carvalho & Cordeiro, 2002) proposes a structured method for the exploitation of XML technologies in a bibliographic environment, which LEADERS is applying in an Archival context. The rationale behind the TVS Model is that use of XML within the library community has focussed on attempts to create an XML version of the ISO 2709 MARC Exchange format while at the same time developing a Schema to validate candidate MARC records, while little or no attention has yet been paid to considering the possible use of Web Services outside the Z39.50 Implementers Group. By separating out the transport, i.e. data exchange issue from the validation, i.e. is this a MARC record?, issue and by exploring ways to implement Web Services the TVS Model aids clarification of thinking and focus of development in the application of XML in the Library community.

In considering the actual and potential use of XML in the Archive community, the LEADERS project team have identified similar issues with regard to the mixing of transport and validation. The very flexibility that has aided the adoption of both EAD and TEI, has mitigated against their usefulness for data exchange and interoperability (Shaw, 2001).

Whilst LEADERS is not specifically concerned with data exchange in the same ways as ISO 2709, the need to create generic tools which can be reused on different sets of encoded materials while at the same time supporting the creation of materials which can themselves be re-used in different contexts, means that the requirements for a basic and consistent structural integrity are the same as for data exchange. This has led us to develop a Schema based on a simplification and a precise interpretation of the encoding rules in the TEI and EAD frameworks.

The Validation element of TVS takes the concept of validation beyond structural integrity, by proposing that a human readable commentary is incorporated alongside the syntactic and semantic rules normally expressed in an XML Schema. The ideal is that the resulting document can act as both a human and machine ‘validation’ mechanism. LEADERS employs this concept to produce documentation integrated with the schema to facilitate the consistent application of TEI and EAD encoding rules.

In the TVS Model the Services element describes the use of Web Services to make the encoded materials ‘self describing’ and thus open them up to exploitation by multiple remote applications. In the LEADERS project we are implementing this concept by incorporating SOAP (Standard Open Access Protocol) messaging within the toolkit and by using WSDL (Web Services Description Language) to describe the services available. Both SOAP and WSDL are XML based systems which provide an open and system neutral method for defining and describing ways in which Web applications can interact with the encoded materials so that information may be retrieved and displayed. Our demonstrator application is built on these principles using the Microsoft .NET tools. .NET was chosen on the basis that most working archives would be in organisations with Microsoft Windows Servers, however the open nature of the tools used means that Java AXIS tools can also be used to develop and host the application. The goal is to provide a set of encoded materials which can be exploited by multiple applications constructed using either .NET or Java AXIS tools and an application which can be used to exploit different source materials encoded to the same schema.

REFERENCES

Bearman, D., ‘User Presentation Language in Archives’, *Archives and Museum Informatics*, vol. 3, Winter 1989-90, pp. 3-7.

Carvalho, Joaquim de, Cordeiro, Maria Inês, ‘XML and bibliographic data: the TVS (Transport, Validation and Services) model’, IFLA 2002 Conference Proceedings, at **http://www.ifla.org/IV/ifla68/papers/075-095e.pdf**

Collins, K., ‘Providing subject access to images: a study of user queries’, *The American Archivist*, vol. 81, 1998

Cox, R.J., ‘Access in the Digital Information Age and Archival Mission: The United States’, *JSA*, vol.19, 1998

Dugg, W. and Stoyanova, P., ‘Transforming The Crazy Quilt: Archival Displays from a User’s Point of View’, *Archivaria*, vol. 45, 1998

Feeny, K., ‘Retrieval of Archival Finding Aids using World-Wide-Web Search Engines,” *The American Archivist*, vol. 62, 1999

Hedstrom, M., ‘How Do Archivists Make Electronic Archives Usable and Accessible?’, *Archives and Manuscripts*, vol.26, 1998

Large, A., Tedd, L.A. and Hartley, R.J., *Information Seeking in the Online Age: Principles and Practice*, Bowker Saur, London, 1999

Mates, B.T., *Adaptive Technology for the Internet: Making Electronic Resources Available to All*, American Library Association, Chicago and London, 1998
National Council on Archives, *Changing the Future of Our Past*, NCA, London, 2002
Peterson Bishop, Ann, 'Measuring Access, Use and Success in Digital Libraries', *The Journal of Electronic Publishing*, vol.4, 1998, at <**http://www.press.umich.edu/jep/04-2/bishop.html**>
Public Record Office, Keeper's Report 2001/02, PRO, London, 2002
Rosenbusch, A., 'Are Our Users Being Served? A Report on Online Archival Databases', *Archives and Manuscripts*, vol. 29, 2000
Seaman, D., 'The User Community as Responsibility and Resource: Building a Sustainable Digital Library', *D-Lib Magazine*, July/August 1997, at <**http://www.dlib.org/dlib/july97/07seaman.html**>
Shaw, Elizabeth J. 'Rethinking EAD: balancing flexibility and interoperability', *New Review of Information Networking*, vol. 7, 2001
Tibbo H. R. and Meho L.I., 'Finding Finding Aids on the World Wide Web', *The American Archivist,* vol. 64, 2001

Textual Critical Encoding

BARBARA BORDALEJO
De Montfort University
bbordalejo@dmu.ac.uk

At the beginning of 2001 work started on the Commedia Project, a research effort which will transcribe in full seven manuscripts of Dante's Divine Comedy in order to collate them and present them in an electronic format. As part of the prolegomena for the Project, we had to write its transcription guidelines, a task that appeared straight forward. However, as often happens, drafting the new guidelines developed into a task with implications beyond its immediate intended use.

Initially, it was agreed that the Commedia Project transcription guidelines should be based on those of the Società Dantesca for their Dante Online website (**http://www.danteonline.it/english/risorse.htm**). It soon became evident that although the *Società Dantesca*'s guidelines offered the advantage of having taken into consideration practical matters concerning spellings, punctuation, word division and the expansion of abbreviations, they did not deal with many other matters which were required for the Commedia Project. The *Società Dantesca* uses a form of symbolic representation—based on conventions—to convey the transcribers interpretation of what they believe to be in the manuscripts. For example, the Società Dantesca transcribes (Riccardiana 1005, Inferno, Canto I, 17):

```
<di +i0 del>
```

These symbols are used to represent a deletion. In this case, the deletion was carried out by the main scribe of the text—or by an indistinguishable hand—indicated by 0. The complete set of symbols is enclosed in angle brackets. The first word, in this case 'di' is the one which was originally in the manuscript, and the last word—'del'—is the one which replaced it. Next to the 0—representing the main hand or one which cannot be distinguished from it—the plus symbol is used—addition—and the letter 'i' which indicates that the correction has been introduced between the lines, i.e. it is interlinear. The *Società Dantesca* guidelines allow the possibility of marginal additions—'m'—or additions within the line—for which they do not use any symbol. In this specific case, according to the transcription produced by the *Società Dantesca*, the manuscript has the word 'di' which has been substituted by the word 'del,' creating the phrase 'del pianeta' instead of the original reading 'di pianeta.'

A second example can be found in Riccardiana 1005, Inferno, Canto I, 94:

```
<crede +i0 cride>
```

Here, the original reading "crede" is followed by the identifiers for the position and the scribe, and at the end, the modified reading "cride," again, all enclosed in angle brackets.

The Società Dantesca system also permits a symbolic representation of marginal additions—`[... +m1 o]`, an addition of the letter 'o' in the margin, which has been added by a second scribe (here represented by the number 1) to cover a lacuna—, interpolations or cancellations.

Although these guidelines were useful as a base for the Commedia Project's transcription system, a new encoding system was required for the encoding of the manuscripts. Other projects in which we are involved used very simple encoding systems. For example, the encoding for the Canterbury Tales Project's publications uses `[add][/add]` for additions and `[del][/del]` for deletions. Thus, an interlinear addition in the Merchant's Tale, line 219 (CTP lineation system) was tagged:

```
tree [add]is[/add] neydir
```

However, the verb is not in the same line as the other words, in fact, there is a caret indicating that the word ‘is’ is an addition to the line. Although the Canterbury Tales Project tags were useful when it started, they now seem to lack the flexibility which is required to present certain aspects of a scholarly edition. Given the nature of the corrections in the Commedia manuscripts, the Project required tags that were able to handle situations more complex than those of additions and deletions. One of the main aims of this project is to produce a CD-ROM with seven witnesses of the Commedia which Federico Sanguineti has identified as textually the most important. Sanguineti has already produced a critical edition of Dante’s Commedia, and the research he has already done is still being carried out at the Canterbury Tales Project. For this reason, the Commedia Project has a clearer conviction of which things are important and should be displayed in the CD-ROMs, and what the purposes of its transcriptions are, than the Canterbury Tales Project had when it was officially started in 1993. Because of this, it was possible to develop a encoding system which allows the distinction of different scribal hands or corrections made by the same scribe at different stages.

Hitherto, the encoding of projects similar to the Commedia Project, such as the Canterbury Tales Project, attempted to present simultaneously both ‘what is in the manuscript’ as a series of additions or deletion, and ‘what is in the text’, as a series of distinct readings. However, after months of discussion with Klaus Wachtel (Institute for New Testament Research, Munster) about the transcription of corrections of the manuscripts of the Greek New Testament, new ideas about how to encode different textual stages started to emerge. These discussions were the base of the encoding system developed for the Commedia Project. The main goal of this new transcription system is to present a clear distinction between what is in the manuscript and how the transcriber interprets the different stages of development of the text.

The Commedia Project encoding system aims to represent the different stages of variation in the text. When a transcriber finds a ‘place of variation’ in the manuscript, he or she can use the apparatus tag—`[app][/app]`. (We are using Collate-style encoding in the transcriptions: before publication, these will be translated into XML encoding). The apparatus tag contains three main components: the original reading (contained in the `[orig][/orig]` tag), the final reading (contained in a tag which specifies which copyist produced this `[c1][/c1]`, `[c2][/c2]`, `[c3][/c3]`), and what literally is in the manuscript (contained in the `[lit][/lit]` tag). If there are more than two stages in a correction, for example, in the case of having more than one corrector), these stages are presented in what is likely to be their successive order.

The following example is taken from Riccardiana 1005, Inferno, Canto III, 9:

The Commedia Project transcription guidelines indicate that we should transcribe as follows :

```
[app]
     [orig]dura[/orig]
     [c1]duro[/c1]
     [lit]dur[ud]a[/ud]o[/lit]
[/app]
```

Since the dot below the letter ‘a’ indicates deletion, the transcriber is faced with a place of variation—indicated in the transcription by the apparatus tag—`[app][/app]`. The first component inside the apparatus tag is the original reading—`[orig]dura[/orig]`. The second component is the final reading by the main hand, the reading `[c1]duro[/c1]`. These two components express different states of the text, but do not explain by which process the text change from one to the other. For this purpose we use the literal tag—`[lit][/lit]`—which indicates what literally is happening in the manuscript, in this case `[lit]dur[ud]a[/ud]o[/lit]`, that is, literally the letters ‘d’ ‘u’ ‘r’ are present, followed by an ‘a’ which has been underdotted and an ‘o.’ In the literal tag, there is less space for interpretation and the transcriber is required to postpone judgment (for example, whether the underdotting of the ‘a’ indicates cancellation or not).

In comparison with the lack of flexibility of the old encoding system of the Canterbury Tales Project, the Commedia Project's guidelines present several advantages. Firstly, the transcribers can defer interpretation of the stages of meaning, since the literal tag can be transcribed independently of the other components of the apparatus tag (this also gives the advantage of allowing the editor of a publication to make a final decision as to what happened at each individual place of variation). Secondly, the contents of the literal tag allows us to reconstruct what actually appears in a manuscript on the computer screen. Thirdly, the other components of the apparatus tag, such as original reading, final reading, and intermediate readings, can be collated separately from the rest of the text. The separate collation of multiple readings in a manuscript will be most useful when a scribe used a manuscript of different affiliation to correct his copy. In such cases, separate collation will allow the isolation of readings which originated in different manuscripts and which could hint at distinct affiliations in a single text. Separate collation might also be of help in cases in which conflation has occurred because a manuscript is corrected with another one from a different branch of a textual tradition.

The encoding system of the Commedia Project has also been implemented by the Canterbury Tales Project (for publications to appear after the Miller's Tale on CD-ROM, edited by Peter Robinson and the Nun's Priest's Tale on CD-ROM, edited by Paul Thomas) and, in the future, might be also adopted by other projects (notably, transcription of Greek New Testament manuscripts) for their transcriptions. This new encoding system also offers advantages when applied to authorial manuscripts, and although it was originally designed to deal with problems of corrections presented by medieval manuscripts, it should work as efficiently to distinguish different authorial versions of a particular text. This should translate into an easier reconstruction of these authorial versions and allow the distinction and separate reconstruction of different authorial versions.

Annotation and Electronic Scholarly Editions

CHRIS TIFFIN
Univ of Queensland
c.tiffin@uq.edu.au

GRAHAM BARWELL
Univ of Wollongong
g.barwell@uow.edu.au

PHILL BERRIE
Australian Defence Force Academy
p.berrie@adfa.edu.au

PAUL EGGERT
Australian Defence Force Academy
p.eggert@adfa.edu.au

This paper considers the purpose of annotation in scholarly editions and the methods by which annotation should be provided in the electronic environment.

There is a consensus that the most important task in scholarly editing is to prepare an accurate and reliable text, according to transparent criteria. The second task is to supplement that text with apparatus to enable the modern reader to read it more adequately. While some of the more influential guides to editing procedure (Center for Editions on American Authors, Committee for Scholarly Editions) suggest that annotation will not in all cases be required, the annotative process can be high on the list of editor's satisfactions and may therefore be accorded disproportionate effort. As Mary-Jo Kline reports, some of the volumes in such magisterial series as the Jefferson Davis Papers and the Madison Papers were criticised for the "plague" of overannotation. Any General Editor must quickly learn stealthy strategies for restraining the annotative enthusiasms of contributors.

The process of footnoting or annotating in the presentation of an edited text has a different role from the one it does in works of literary critical commentary. On a critic's page the primary end of the annotation is to provide the evidentiary basis for the commentator's assertions and argument, or, in one lurid account, to fight turf wars about academic status with other scholars (McFarland 1991). We concentrate here, however, on the principles and practice of annotating the primary text in a scholarly edition.

The purpose of annotation is usually framed with the reader in mind. It may be based idealistically on "render[ing] the author's meaning wholly intelligible" (Battestin 1981), or it may be based on the attempt to provide the modern reader with the information that would have been possessed by a reader on first publication. In the case of unpublished works, it may translate a general reader into an essentially private world. There is often a strong injunction that annotation should not be subjective or judgmental (CEAA, Hewett 1996). In practice, a further aim may be to present the accretion of scholarly knowledge and interpretation of the text up to the present (Ricks 1989).

Annotation in the print mode is arranged in a number of different ways: textual or explanatory notes at the foot of the page, notes at the end of the chapter or book, collations, glossaries, appendices, and "excursus notes" (McFarland 1991).

For reasons of cost, electronic editions can offer much richer annotation than print editions, and

targeted parts of this extra material can be made accessible from precise points in the text. Two important questions arise, though: how should this ancillary material be arranged, and how should it be linked, or rather, how should its availability be signaled from within the text?

One system, used especially by publications like journals that have both a print and a web publication, is to follow closely the endnotes practice used in print publications by offering the notes in a single file and signaling them by a numerical footnote indice. Other editions indicate the presence of the annotative note by a marginal indice (e.g. Thomas Gray Archive), by highlighting the span of text to which the note refers, or by enabling the highlighting of the text span through a mouseover to reveal the presence of the link.

Some of the most spectacular scholarly editing projects of the last decade such as the Rossetti Archive and the Blake Archive situate their texts in such a complex environment that the annotative strategy of immediate explication of individual terms in texts is abandoned in favor of more serious immersion in the contexts of the poems. Paradoxically, though, under this system precise assistance is not always available to the reader.

In the electronic edition of His Natural Life developed by the Authenticated Editions Project at its JITM website, we commenced with the list of annotations in a flat file transferred as legacy data from the printed Academy Edition of this text, and highlighted the text spans which were indicated in the lead-ins to the print edition's endnotes. But this seemed to make poor use of the amplitude and precision offered by the markup scheme of the electronic edition. Accordingly, we have analysed the content of the notes and reformatted them according to type into sub-glossaries. The result constitutes in itself a critical approach to the text.

Signaling the presence of annotations has been rethought in the light of the expected readership of the electronic version, and this has led to the provision of greater density of annotation links. However, providing for multitudinous non-sequential reading patterns can easily produce an absurdly over-linked text when the names of recurrent characters, places or motifs are annotated. This latter problem can be alleviated if users have a mechanism for enabling or suppressing the link indicators themselves.

We conclude that different strategies are required for the arrangement and announcement of annotative material in the electronic environment from the print one, but the major determinant of how annotations are to be arranged and offered (and perhaps still the most difficult thing to judge) is how readers will approach the edition.

REFERENCES

Authenticated Editions Project **http://idun.itsc.adfa.edu.au/ASEC/aueledns.html**

Battestin 1981. Martin C. Battestin, "A Rationale of Literary Annotation", *Studies in Bibliography* 34: 1–22.

Clark 2001. Marcus Clarke, His Natural Life, ed. *Lurline Stuart. The Academy Editions of Australian Literature*, (St Lucia, Qld: University of Queensland Press).

Furura 2002. Richard Futura and Eduardo Urbina, "On the Characteristics of Scholarly Annotations", *HT02*, June 11–15.

Hanna 1991. Ralph Hanna III, "Annotation as Social Practice" in *Annotation and Its Texts*, ed. Stephen A Barney (New York: Oxford University Press), pp.178–84.

Hewett 1996. David Hewett et al. *The Edinburgh Edition of the Waverley Novels: A Guide for Editors*, (Edinburgh: Edinburgh Edition of the Waverley Novels).

Kline 1987. Mary-Jo Kline, *A Guide to Documentary Editing* (Baltimore: Johns Hopkins University Press).

McFarland 1991. Thomas McFarland, "Who Was Benjamin Whichcote? Or The Myth of Annotation", in *Annotation and Its Texts*, ed. Stephen A Barney (New York: Oxford University Press), pp.152–77.

McGann 1998. Jerome McGann. "Textual Scholarship, Textual Theory, and the Uses of Electronic Tools" A Brief Report on Current Undertakings", *Victorian Studies* 41: 609–19.

Modern Language Association of America. Committee on Scholarly Editions. *MLA Guidelines for Electronic Scholarly Editions*. **http://sunsite.berkeley.edu/MLA/guidelines.html**

Modern Language Association of America. Committee on Scholarly Editions. *Draft Revised MLA Guidelines for Electronic Scholarly Editions*. **http://jefferson.village.virginia.edu/~jmu2m/cse/CSEguidelines.html**

Ricks 1987. Christopher Ricks, *The Poems of Tennyson*, 3 vols. (Harlow, Essex: Longman).

Rossetti Archive. **http://www.iath.virginia.edu/rossetti/index.html**

Suarez 2000. Michael Suarez S.J., "In Dreams Begins Responsibility: Novels, Promises and the Electronic Editor", in *Textual Studies and the Common Reader*, ed. Alexander Pettit. (Athens, Georgia: University of Georgia Press), pp.160–79.

Theory in Text Encoding

PAUL CATON
Scholarly Technology Group, Brown University
Paul_Caton@brown.edu

The existing body of theoretical work in text encoding suffers from two interrelated problems: firstly, confusion over the specific nature of the work; and secondly, the absence of any truly critical theory.

Theory's purest signified and a fundamental predicate of progress in the natural sciences involves a representation that meets discursively bounded criteria of "truth." According to Renear (1997) the electronic document encoding and processing community "has evolved a rich body of illuminating theory about the nature of text" (107), a claim that expands that of Renear, Durand, and Mylonas (1996) and anticipates Mylonas and Renear (1999) who assert the principal goal of the research community that develops and applies the TEI Guidelines and other text markup schemes is a greater "*theoretical* understanding of textual representation" (my emphasis). In making their case the authors invoke Lakatosian criteria which they confidently pronounce (some disclaimers notwithstanding) this research community's work meets. Yet the development of OHCOs 1-through-3, for example—surely one of text encoding's founding "theoretical" moments— demonstrably fails to qualify as a progressive problemshift in Lakatosian terms.[1]

Text encoding has not generated a rich body of illuminating theory because it cannot. Asking, for example, "what is text, really?" already poses the wrong question because it assumes an undiscovered essence the encoding community can find with a progressive research program. This will not happen because no mysterious core exists whose explanation falls to home-grown text encoding theory. Renear argues that despite falsification of all OHCO variants there is nevertheless "no reason to give up the common-sense view that texts do have an objective structure independent of our methods and theories about them" (1997, 122). Structure there may be, but discovering it is not like positing the double-helix of DNA; we need not strive to understand-by-modelling because of inadequate observational technology or limited knowledge. Praxis prompts reflection which generates principle to guide subsequent praxis. Poetry, for example, shows a self-conscious praxis continually reviewing, refining, and codifying (prescriptively) its methodology,[2] and this is as true of its encoding as its writing. Out of "successful" text encoding praxis comes not theory but principle.[3]

This is to say not that text encoding has no theoretical component but that pursuing the theory behind a principle takes us to another place: to linguistics or rhetoric or semiotics or the cognitive psychology of visual perception, and so on. Such pursuits can produce sophisticated and thought-provoking borrowings that unquestionably enrich the literature, but these are the exception, not the rule.[4] In particular, scholarly work in text encoding rarely positions itself with respect to work in modern literary/cultural theory, and even more rarely does it use text encoding as a springboard for a sustained engagement with such theory. This seems to me a significant and unfortunate absence, a product both of text encoding's desire to differentiate itself as a specific field of intellectual inquiry and of a positivist, utilitarian bias against the perceived negativity and self-imprisoning reflexivity of contemporary theory. Renear's narrative of the development of text encoding theory exemplifies the latter stance with its Realist/Anti-realist distinction, associating the former with common-sense and characterizing the latter as "consistent with post-structuralist epistemologies" (122). This imitates an exclusionary move Zavarzadeh and Morton (1991) identify in literary studies: positing deconstruction as the boundary of theory beyond which it is unthinkable to go because (supposedly) deconstruction represents the limits of the thinkable, the point where theory swallows itself in absolute relativism. Ironically, Renear's account positions itself precisely as theoretically reflexive while exposing its own refusal of genuine reflexivity.

Text encoding—indeed humanities computing as a whole—can too easily think of itself as related to every humanities discipline but also marginal or even external to all of them. This licenses attitudes such as Renear's contention that what he calls text encoding theory brings "a much needed fresh perspective on textuality" (107), as if text encoding occupied a different space from traditional disciplines. Contrarily I would argue that whatever its origins in non-academic praxes, text encoding forms itself *in* and *of* humanities disciplines. I should stress that I do not consider these disciplines stable sites: they can and should be

challenged. Text encoding therefore offers a locus for work that tries to think through the tensions, contradictions, and faultlines that constitute those disciplines *qua* humanities disciplines. However, unless it can stand on sufficiently equal terms to enter a critical dialogue with theory of exemplary reflexivity and philosophical rigor, the text encoding community's theoretical work will have limited significance and appeal—a fate already shared by much of the humanities computing literature (Corns 1991; Warwick 1999). My own position is that the historical materialism that comes down to us from Marx, enriched and updated by thinkers such as Lenin, Adorno, Althusser, and many others, offers us the best critical tools for a dialectical engagement with what it means to encode texts in the humanities. Althusser (1982) memorably describes the problem:

> Left to itself, a spontaneous (technical) practice produces only the "theory" it needs as a means to produce the ends assigned to it: this "theory" is never more than the reflection of this end, uncriticized, unknown, in its means of realization, that is, it is a *by-product* of the reflection of the technical practice's end on its means. A "theory" which does not question the end whose *by-product* it is remains a prisoner of this end and of the "realities" which have imposed it as an end. (171, emphasis in original)

Currently a prisoner of its pragmatic roots, theoretical work on text encoding has only its chains to lose.

REFERENCES

Althusser, Louis. 1982. *For Marx*. First published in French, 1963. Translated by Ben Brewster. London: Verso.

Buzzetti, Dino. 1999. "Text Representation and Textual Models." Paper presented at ACH/ALLC 1999, June 1999, University of Virginia.

Caton, Paul. 2001. "Towards a Politics of Text Encoding." Paper presented at ACH/ALLC 2001, June 2001, New York University.

Corns, Thomas. 1991. "Applications in the Study of English Literature," *Literary and Linguistic Computing* 6 (2): 127–30.

Mylonas, Elli, and Allen H. Renear. 1999. "The Text Encoding Initiative at 10: Not Just an Interchange Format Anymore—But a New Research Community," *Computers and the Humanities* 33, 1–2 (April 1999): 1–9.

Piez, Wendell. 2001. "Beyond the 'Descriptive vs. Procedural' Distinction." Paper presented at Extreme Markup Languages 2001, August 2001, Montreal.

Renear, Allen H. 1997. "Out of Praxis: Three (Meta)Theories of Textuality," in *Electronic Text: Investigations in Method and Theory*. Edited by Kathryn Sutherland. Oxford: Oxford University Press.

Renear, Allen H. 2000. "The Descriptive/Procedural Distinction is Flawed." Paper presented at Extreme Markup Languages 2000, August 2000, Montreal.

Warwick, Claire. 1999. "English Literature, Electronic Text, and Computer Analysis: an Impossible Combination?" Paper presented at ACH/ALLC 1999, June 1999, University of Virginia.

Zavarzadeh, Mas'ud, and Donald Morton. 1991. *Theory, (Post)Modernity, Opposition: An "Other" Introduction to Literary and Cultural Theory*. PostModern Positions, Vol. 5. Washington D.C.: Maisonneuve Press.

NOTES

[1.] I omit the argument of proof here for brevity.

[2.] In the hyper-self-consciousness of artistic praxis prescription inevitably invites its own negation—though anti-prescription is arguably still a principle.

[3.] See, for example, the Women Writers Project's online guide for marking up line groups. (**http://www.wwp.brown.edu/encoding/training/lg/page1.html**)

[4.] Recent productive borrowings include Buzzetti's from *Hjelmslevian Semiotics* (1999); Renear's from Austin's *Speech Act Theory* (2000); and Piez's from *Traditional Rhetoric* (2001).

Burrowing into Translation: A Case Study

JAN RYBICKI
Pedagogical University, Krakow, Poland
strybick@cyf-kr.edu.pl

While computational stylistics and text analysis seems to be in constant quest for the just right set of criteria (see e.g. David Hoover's presentation at the 2002 ALLC/ACH Conference in Tübingen), this presentation will try to apply what has already become a standard in statistical stylistic analysis to a (relatively) novel material. Taking for granted—very unoriginally—the usefulness of John Burrows's method that has been around since his 1987 Computation into Criticism, in discerning stylistic differences between individual characters in works by the same author, I will try to see if the same or similar patterns of similarity and difference travel across linguistic boundaries; if differences between characters' "idiolects" are preserved in translation.

In a typically Polish approach to the matter, I have chosen as my material the trilogy of historical romances by Poland's first literary Nobel Prize winner, Henryk Sienkiewicz, written between 1882 and 1888, and its two English (or, more precisely, American) translations by Jeremiah Curtin (completed between 1890 and 1893) and W.S. Kuniczak (1991). The reasons for this choice have been manifold. First, Sienkiewicz's three novels, With Fire and Sword, The Deluge, and Pan Michael, although set in the turmoil of 17th-century Poland, remain to this day a major classic of Polish literature and the country's most popular reading. Second, a trilogy, the subsequent parts of which share some of their characters, seems ideal for Burrowsian analysis (a fact confirmed by the interesting coincidence of John Burrows's undertaking of a study of Beckett's bilingual trilogy in a much later paper). The final reason was the difference between the two translations, evident both in their being separated by an entire century and, what follows, in the entirely different approaches and results obtained by the two translators: the largely word-by-word transcoding by Sienkiewicz's contemporary and the highly adaptative and "free" method of the modern Polish-American writer.

Faithfully maintaining the original Burrows model, the study of distances between the "idiolects" of the major characters has been based on relative frequencies of the 30 most frequent words in the dialogue of each version of the trilogy. The resulting correlation matrices were then used to produce two-dimensional multidimensional scaling charts of distances between such "idiolects."

This procedure has yielded, in Sienkiewicz's original, a very consistent influence of the personae's social status and ethnic background, especially in the first part of the series. It is particularly visible in idiolects of 'enemy' (non-Polish) collective characters, usually plotted at some distance one from the other. Curtin's translation is notable for 'de-clustering' idiolects of various Polish gentry characters, making their idiolects much less alike. There is also a visible tendency in Curtin to limit the distances between rival collective characters and making them markedly similar rather than divergent as in the original. Idiolects in Kuniczak are even more evenly distributed, with a general trend towards greater distances and less clustering observable in the graphs. The very high resemblance between the idiolects of two major characters, Zag#oba and Wo#odyjowski, in all three versions of The Deluge (almost identical in Sienkiewicz) is one of the most consistent traits of this portion of the analysis—an interesting illustration of the fact that the two personae's function of keeping the three volumes together becomes evident in the second part of the series. Sienkiewicz's social/ethnic idiosyncrasy has been confirmed in a plot for idiolects of characters involved in the Polish-Ukrainian conflict in With Fire and Sword, a feature slightly visible in Kuniczak and almost not at all in Curtin.

As perhaps the most consistent effect of all, the peripheral situation of female idiolects is a constant element in almost any configuration. The study of more detailed and thematic configurations of characters is also the source of interesting insight. Plots for female characters exhibit a tendency to group together young (and marriageable) Polish women; Helena's Ukrainian provenience is highly visible in Sienkiewicz, while the ethnic element is indiscernible in both translations. Among characters involved in each novel's eternal triangles, both of the above aspects are clearly visible in all three versions; differences between characters in the same triangle are quite considerable.

In an examination of characters that recur throughout the series, a good consistence has been observed between the three idiolects of the Polish Falstaff, Zag#oba, the most inveterate talker of the series, in each version separately: best in Sienkiewicz, worst in Curtin. Another character, Wo#odyjowski, is much more of a developing character, which agrees well with the evolution of the persona in the course of the

series, from a humble officer to the hero and spearhead of Sienkiewicz's ideology. This has been confirmed in a separate plotting of idiolects of those two characters.

A number of joint plots for Curtin's and Kuniczak's translations (based on frequent words common in both versions) has been made to investigate if there is a constant pattern in their respective differences. In agreement with some "intuitive" assessments as to the decreasing differences between Curtin's and Kuniczak's versions (mainly due to Kuniczak's gradual abandonment of adaptative procedures, especially on the microstructural level), the patterns become more consistent with time: fairly chaotic movement for the first part of the trilogy has become more ordered in the second and almost uniform in the third. The "stylistic drift" observed between idiolects in Curtin and in Kuniczak— divided, apart from their contrasting approaches to translation, by an even more significant difference of a whole century—is a vindication of Burrows's 'tiptoeing towards the infinite:' that visible and uniform shift in the configuration of the most frequent words in English texts with time.

REFERENCES

D.I. Hoover, 'New Directions in Statistical Stylistics and Authorship Attribution' Proc. ALLC/ACH 2002: 51–53.

J.F. Burrows, *Computation into Criticism: A Study of Jane Austen's Novels and an Experiment in Method* (Oxford: Clarendon Press, 1987).

J.F. Burrows, 'Tiptoeing into the Infinite: Testing for Evidence of National Differences in the Language of English Narrative' *Research in Humanities Computing* Vol. 2 (1994): 1–33.

W. McKenna, J.F. Burrows, A. Antonia, 'Beckett's Trilogy: Computational Stylistics and the Nature of Translation,' *Revue informatique et statistique dans les sciences humaines* 35 (1999), 151–71.

Jan Rybicki, *A Computer Assisted Comparative Analysis of Two English Translations of Henryk Sienkiewicz's Trilogy* (Kraków: WAP, 2003).

A Computer-Based Questionnaire for Hearing Impaired People

JOACHIM GERICH
Dep. for Sociology, University of Linz/Austria
joachim.gerich@jku.at

ROLAND LEHNER
Dep. for Sociology, University of Linz/Austria
rl@x-net.at

In this paper we present a self administered computer based method of survey research for hearing impaired people originally designed for the measurement of "quality of life" indicators.[a]

The application of standardized paper & pencil questionnaires as a method of measurement and data collection is a wide-spread and commonly used technique in empirical social research. Compared to other methods, considerable advantages are higher comparability of the single measurements, simple performance and speedy data administration. A self administered questionnaire needs only small resources and—due to the absence of involved interviewers— interviewer effects as a source of measurement-bias are reduced. Further on, it exits a large pool of established questionnaires in various fields of research and diagnosis. Therefore the implementation of the questionnaires is possible without further large-scale developments and test examinations. In addition, we have the opportunity to compare already collected data of earlier surveys with our results.

In our project, we examined indicators of quality of life for a representative sample of hearing-loss and hearing impaired persons in an Austrian region. The goals were to attain descriptive information about the specific aspects of the quality of life of this population and to gain knowledge of possible improvements. In detail, the questionnaire has to be designed as a long-term monitoring instrument of different strategies for therapies and medical or psychological attentions in the ambulance for hearing impaired persons of a local hospital. Furthermore, comparisons between a variety of dimensions of quality of life with regard to different subpopulations and to a diversity of nationalities have to be enabled. On the basis of these and some other requirements—which are not mentioned here—the project group decision was to use three instruments for the

administration of quality of life. These instruments are the WHOQOL-Bref (WHO-quality of life questionnaire[7, 1]), the BSI (Brief Symptom Inventory[3]) and the GHQ Questionnaire (Goldberg Health Questionnaire[5]).

All three questionnaires are originally constructed as self administered standardized paper & pencil instruments. One main criterion for applying standardized Questionnaires is that the respondents have hardly problems with the media of the literary question-answer type. As a result, it is not appropriate for persons with reduced reading abilities (e.g. [2] 164). Taking into account the group of hearing impaired persons this recommendation is not fulfilled by each interviewee but not due to illiteracy. The vast majority of the deaf community cannot be labelled as "illiterate", however the primary communication channel is not written or (the phonetic equivalence) of spoken language but mainly the sign language as a visual channel[6, 8]. So we have to consider that wide parts of this specific population have not got a comparable literary ability to persons with the ability to hear: "we have to deal with the phenomenon that many members of the deaf group are unable to understand written texts of normal complexity"[4] (208). Therefore, the principle task for the implementation of the questionnaires for the specific needs of hearing impaired persons is the creation of an instrument including the favoured form of communication namely ASL (Austrian Sign Language).

One possible adaptation is the face-to-face interview. In practice, the interviewers capable of sign language have to translate the questions and answers during "real time" of the interviews. This way of problem solving raises a lot of problems. Firstly, by changing the design of the instruments from the original paper&pencil to face-to-face mode, the results of the administration are not fully comparable any more. Some sensitive questions (e.g. about the sexuality and the satisfaction with the own personal outfit) included in those questionnaires causes a higher rate of biased answers due to effects like self disclosure and social desirability in face-to-face situations. Secondly, the typical situation in standardized face-to-face interviews is that the interviewer reads out the questions to ensure the similarity of the interview situations. Sign language cannot be read to someone. The only way is to translate it from text to sign language during each interview session, resulting in a variety of translations in different interview situations because of the use of different mimics and hence reducing the reliability of the data administration. Thirdly, the recruiting of interviewers with appropriate knowledge of sign language causes additional administrative efforts. Furthermore, the high degree of institutionalisation of deaf people increases significantly the probability that respondent and interviewer know each other. In other words it implies again a higher probability of biased answers. Recruiting interviewers from different regions is constrained by the considerable variations of local dialects of ASL. Finally, the situation of face-to-face interviews restricts the way of communication to a visual one. As a matter of fact, hearing impaired persons should have the opportunity to decide by their own the preferred way of communication. Due to all this considerations, we had to find a way in which sign and literary language are simultaneous presented and freely elective. To put it more precisely, we had to find a form of data administration allowing the self administration of the questions, differing as few as possible from the original form of paper&pencil presentation, giving the free and simultaneous choice between text and sign language and retaining the advantages of paper & pencil method such as simple administration and cost effectiveness.

Those recommendations can be fulfilled to a great extent through the implementation of the questionnaires as multimedia CASI (Computer Assisted Self Interview). Consequently, we developed the computer program Animaqu (Animated Questionnaire) as a result of interdisciplinary work of deaf people, sociologists, neurologists, linguists and psychologists. Animaqu contains a window for video as well as text and features for the answering and navigation through the questionnaire. Question and answer categories are presented as video sequences in sign language and as written text.

The specific pros and cons of this kind of multimedia CASI, the comparability with traditional methods for data administration and some first results of the data administration will be discussed.

NOTES

[1.] This research project is a cooperation between the department of sociology and the outpatient clinic for the Deaf and hard-of-hearing at the general hospital of St. John of God in Linz, Austria.

REFERENCES

Angermeyer, Matthias C. / Kilian, Reinhold, Matschinger, Herbert (2000): WHOQOL— 100 und WHOQOL—BREF. *Handbuch für die deutschsprachige Version der WHO Instrumente zur Erfassung der Lebensqualität*. Göttingen.

Atteslander, Peter (1993): *Methoden der empirischen Sozialforschung*. 7. Auflage, Berlin u.a.

Derogatis, L.R. & Spencer, P.M. (1982): *Brief Symptom Inventory: Administration, scoring and procederes manual I*. Baltimore

Dotter, Franz (1996): *Computer for the deaf (and hearing impaired): Towards an integrated solution from a linguistic standpoint*. In: Klaus, J. et al. (Eds.): *Interdisciplinary aspects on computers helping people*

with special needs. Linz, pp. 205-210.
Goldberg, David .P. (1978): *Manual of the General Health Questionnaire*. Windsor
Lucas, Ceil (Ed.), 1996: *Multicultural Aspects of Sociolinguistics in Deaf Communities*. Washington D.C.
The WHOQOL Group, (1998): *Development of the World Health Organization WHOQOL-BREF quality of life assessment*. Psychological Medicine, 28, 551-558.
Wilcox, Sherman (Ed.), 1989: *American Deaf Community*. Burtonsville.

Beyond Taxonomy: Digital Poetics and the Problem of Reading

TALAN MEMMOTT
Brown University
talan@memmott.org

*(Talan Memmott is a hypermedia artist/writer/editor from San Francisco, California. He is the Creative Director and Editor of the online hypermedia literary journal BeeHive (**http://beehive.temporalimage.com**). His hypermedia work appears widely on the Internet. In 2001 he was awarded the trAce/Alt-X New Media Writing Award for his work Lexia to Perplexia, which also received honorable mention for the Electronic Literature Organization's award in fiction. He is a tutor for the trAce Online Writing School, and has been a speaker, panelist, reader and performer at various Conferences and Universities. He is currently at Brown University as their first electronic writing graduate fellow.)*

What is digital poetry? The definitions are decidedly nebulous.

The term digital poetry has been applied to a variety of creative literary applications, from work developed in Flash and DHTML to MOO spaces and works that utilize Perl. From cybertext to web art, digital poetry is somewhat interchangeable with other terms used to describe what could be called creative cultural practice through applied technology. We can agree that digital poetry as hypermedia presents an expanded field of textuality that moves writing beyond the word, toward a relationship between signs and sign regimes, their integration, disintegration, and interaction one to another. But, how these relationships are established in digital poetry is as diverse and various as the practice itself.

The problems of developing any general typology, let alone taxonomy, are hinted at in Espen Aarseth's *Cybertext: Perspectives on Ergodic Literature*. Aarseth puts forward a number of models for the definition of various types of objects (buttons, actors, interactive, controller, layout) within creative applications—including games, interactive fiction, and hypertext. As the analysis expands, it is discovered that these typologies breakdown when any given piece is viewed as a whole. One moment a button may be a button, the next moment it may be an actor; or, any given element may carry the attributes of any number of types simultaneously.

Digital poetry does not properly define any specific type of expressive object. Because of this there are many problems that emerge for the reader/users of digital poetry and for those that deal critically with such work. Lacking a definite object of study, we must begin to move away from the idea of digital poetry as a genre toward an observation of applied poetics within the digital environment—a poetics that is based in an individual author's engagement with media technologies, as scripted, programmed and applied within a particular work.

Using Artaud's *The Theater and its Double* as a guide, this paper explores the mise en scène (or mise en screen) as a potential model for the close reading of digital poetry. The paper looks at a number of web-based digital poetry works that utilize a variety of technologies to demonstrate how the network and its technologies play into artist/writer intent to develop an applied poetics rather than poetry proper.

Katherine Parrish's *Oulipo* inspire web project *MOOlipo* is examined for its creative use of MOO technology to create a particpatory poetical space. In this work the user particpates by inputting text, which is parsed and filtered under certain rules to affect the output. The various rooms of the MOO have different produce different effects. In one room a mesotic is created from user input, in another room the word order of the input text is reversed.

Two other works that require user participation for the construction of content will be examined. Lisa Jevbratt's *Syncro Mail*, a web-based mail service, requires that a user input the email address of a second (perhaps unknowing) 'user'. Through this process, the second 'user' receives a random image and a random word in their email. This piece uses perl scripting for its functionality and presents a unique, if not mysterious method of poetic emergence. In the delivered email there is no explanation as to the relevance or connection

of image to word, nor any indication of how, or where the mail originated. The connections, the poetry must be made by this second 'user', independent of any knowledge of the process. Another project with much more immediate participatory poetic results is *You and We*, a collaboration of Seb Chevrel and Gabe Kean. *You and We* allows users to upload images and short texts. Using a combination of Flash and Macromedia Generator, the images and texts are randomly compiled in a somewhat cinematic, MTV-like display complete with music. Within seconds of uploading an image or text, it is incorporated into the collection. As of November 10th, 2002 there had been over 1,500 images uploaded and nearly 5,000 texts.

Additional works to be examined include; Brian Kim Stefans' *The Dreamlife of Letters* for its use of letterist animation, and a couple of "codeworks" by Ted Warnell—*VIRU2* and *BERLIOZ*—for their elegance and simplicity of interface as well as their transparency. Warnell's work allows the functions of code to play into the applied poetics of his work, at the surface. In *VIRU2* the actual code that drives the piece is made viewable as screenal text. The exposure of code at the surface and the integration of functionality, aesthetics and poetics in Warnell's work emphasizes the role of technology and an individuated ecounter with media in establishing what is inferred by the term digital poetry as well as applied poetics.

Rather than work from a model that hopes for a close reading through the abstraction of words from their media-rich environment, this paper proposes that critics and readers take a more choragraphic (to borrow a term from Gregory Ulmer) approach to reading that observes the entirety of a work—from interface design to interactivity, the written word and code—as something of a micro-cultural statement. By examining digital poetry objects as a whole we may begin to recognize how each work presents an individuated applied poetics and move away from overreaching taxonomic designations.

This paper also proposes that more critical work be developed in hypermedia environments as a way of diminishing critical/theoretical detachment from the realities of creative digital practice.

Beyond the Archive: Immersive Textuality for William Blake's Poetry

---

STEVE GUYNUP
Georgia Tech, IDT
steve_guynup@hotmail.com

MARCEL O'GORMAN
Detoriot Mercy University
marcel@e-crit.com

NELSON HILTON
University of Georgia
nhilton@room343.english.uga.edu

RON BROGLIO
Georgia Tech, LCC
ron.broglio@lcc.gatech.edu

In the past decade the Blake Archive transformed scholarship on William Blake by making his illuminated texts widely available online. The Archive delivers some of the best features of web edited editions, including the ability for the viewer to modulate the size and color of plates, to compare similar images, and to search for thematicly similar images. Nonetheless, there are many applications of new media that the Archive does not employ. The performativity of new media matches complex performative elements in Blake's work that are not accounted for in an archive. As poet, artist, and publisher, Blake is keenly aware of the horizons and limits of his media. As poet he creates narratives in which characters and spaces morph. Such transformation of people and places in turn structurally change the narrative. As artist, his visual puns and transformed characters take on Ovid-like metamorphosis, such as his tree/women that liberate and bind their children with their "limbs." As publisher, his very processes of etching, painting, and publishing become characters in the narrative and morph the story world at the same time that they call attention to themselves as media in our world. This conference panel is designed to look at the "other" Blake, one not available in the Archive but necessary in the immersive experience of reading Blake's work. As the visionary poet sets out to "open the

doors of perception," this panel presents applications for modelling and simulating the phenomenology of reading. While the archive presents a "transparent" text and a reader at a non-point outside the poem and distanced from the computer screen, immersive textuality attempts to place the reader inside the space of reading while both inside and outside the computer screen. The complex image-text, multivalent narratives of William Blake's poetry serve as an ideal model for thinking digital representation other than an archive. Blake's work becomes a means for formulating responsive new media spaces both for scholarship and for pedagogy.

Immersive texts present elements of gaming to literary criticism. The panel will consider how such environments bring gaming to scholarship. While Jerome McGann has begun to address these issues in his Ivanhoe game, recently published in Radient Textuality, the panel's participants have been engaged in similar efforts over the last decade. Play, embodiment, narrativity, and design—key elements of gaming—help Blake scholars to rethink the look, feel, and purpose of Blake's cosmologies. Unlike McGann's work, the present panel uses a variety of new media technologies and designs to think through the reader's position within the text rather than rendering an interpretation of the text from an "outside." The reader/user is asked to stand within the text rather than using various applications and discourses to post or discuss texts.

The panel will provide brief demonstrations of immersive digital texts accompanied by short papers and then roundtable discussion finally opening into a question and answer period. Panelists are from a variety of disciplines. Steve Guynup, a seven year veteran in web3D and internationally recognized digital artist, will present a web3D immersive illustration of Blake's "Crystal Cabinet" as a model that allows readers to understand Blake's sense of folds between worlds and Blakean embodiment. Marcel O'Gorman, head of Detroit Mercy University's Electronic Criticism program, will present "The Fourfolds of William Blake and Martin Heidegger: Minds, Bodies, Technologies." For O'Gorman the intersections between the four folds of both Blake and Heidegger serve as a path toward understanding how design and critique might intervene in the separation of body and mind wrought by technology. Nelson Hilton, a leading scholar in Blake criticism, head of the online text of Blake, and chair at University of Georgia, will present "Golgonooza Songs, or, Blake in a Flash." Hilton will show and discuss how Flash can be used to visualize multiple versions of Blake's poems, calling into question textual stability and unseating transparency of the text. Finally, Ron Broglio, associate editor of Romantic Circles online journal and noted for his experimentation in digitizing Romantic scholarship, will demonstrate and discuss the use of MOOs as graphically-oriented multi-user virtual environment that can incorporate Flash and Java applets to create experiential texts. Broglio has led a team of Georgia Tech design students in consultation with Blake scholars at other universities to create scenes of reading key moments of Blake's illuminated texts. The mix of demonstration with discussion and the variety of new media applications employed to represent the text provide a disjunctive synthesis of archival procedures and gaming environments. While much of humanities scholarship has employed computing toward editorial and archival ends, it is the hope of this panel to open traditional humanities texts to other applications, discourses, and play available in new media.

Research Library Collection Descriptive Frameworks

GUENTER WAIBEL
Berkeley Museum of Art,Research Libraries Group
guenter@uclink4.berkeley.edu

JARED CAMPBELL
University of California, Davis
jlecampbell@ucdavis.edu

NANCY KUSHIGIAN
University of California, Davis
njkushigian@ucdavis.edu

Libraries and archives have embraced the potential of the World Wide Web as a means of expanding access to their collections. Scholarly work in humanities computing has evolved as technologies have evolved. The early availability of technologies to store and manipulate online text spawned extensive work in linguistics, literature, and documentary history. As computer memory becomes cheaper and more widely available, this trend is mirrored in scholarly fields dealing with visual or audio or hypertext primary sources. Susan Hockey,

in her keynote address to the TEI Consortium this fall, predicts that we are experiencing a vast shift in how humanities scholarship is conducted, and that this shift to digital formats will continue. This state of affairs poses grand challenges for academic research libraries and museums, which are faced with the task of making the products of humanities computing research available to the academic community, and preserving them for the future.

Research libraries now store and provide access to vast online collections, and yet, like traditional manuscript or archival collections, these remain widely unfamiliar and unknown to most library users. To provide access by cataloging individual texts or images within these collections is cost prohibitive. As a response to this challenge, research and systems librarians have proposed and are considering "Access Integration" models, such as that proposed by Bernard Hurley to the California Digital Library. These models provide "points of access" through online library systems to customized or specialized local scholarly interfaces that provide users more granular descriptions and data manipulation tools. In such a model, collections of books, images, manuscripts and other groupings, would be represented in major library user interfaces by collection descriptions.

The papers in this panel discuss library metadata initiatives that seeks to understand, create, and use digital collection descriptive frameworks for many different kinds of library collections, both printed and digital. Guenter Waibel, curator at the Berkeley Museum Art, presents his work on the Making of America Project (MOAC) and with the California Digital Library's Online Archive of California (OAC) adapting the Encoded Archival Description as a framework for museum collections, thus allowing interoperability with other OAC finding aids. He goes on to discuss museum collection issues more broadly.

EAD is a specialized DTD built and maintained for an international archival community. Libraries have therefore begun to consider other frameworks for organizing and describing other kinds of collections such as published texts, digital images, sound files and hypertext. In his paper, electronic resources librarian Jared Campbell describes and evaluates some of these new developing frameworks, including OAIS, METS and RSLP. He argues that the creation of collection metadata needs to be considered by scholars on the "front end of a project" to ensure access through library search mechanisms for the digital texts or images they are creating.

Finally, research librarian Nancy Kushigian draws on her work with the British Women Romantic Poets' project to observe that no current collection description framework is sufficiently granular to describe legacy printed collection descriptions. She presents an overview and analysis of printed collection descriptions that currently populate the shelves of research libraries, and suggests a basic taxonomy for these catalogs and collection descriptions as a basis for contemplating a descriptive framework, possibly TEI based, adequate to preserve their richness and variety and to serve as a framework for the creation of new, web-based collection descriptions.

MUSEUMS IN THE MIX—COLLECTIONS ACCESS ACROSS COMMUNITIES

Guenter Waibel, Berkeley Museum of Art/California Digital Library/Research Libraries Group

Museums, Libraries and Archives increasingly collaborate to bring their collections online within one integrated system. The integration of access is fueled by the realization within the different communities that all cultural heritage institutions hold similar objects, as well as by end-user demands for searches across institutional boundaries. The statewide project Museums and the Online Archive of California (MOAC) adapts archival and library standards to integrate museum collections into the greater space of cultural heritage.

Long before the Internet, each individual cultural heritage community developed its own unique access model suited to its' particular collection and mission. Libraries have developed the Online Public Access Catalog (OPAC) to facilitate searching of their book holdings. Archives have used the finding aid as an access tool for researchers even long before finding aids became synonymous with their electronic incarnation as Encoded Archival Description (EAD). Access to collections in museums revolves around thematic exhibitions, which bring together objects from within the institution's collection as well as borrowed objects from the outside to form a coherent experience for the visitor. In this way, museums provide value-added access through grouping the objects into meaningful exhibitions, plus they enrich the offering through educational materials and programming. However, this form of access remains incomplete: over 95% of a museum collections will be in storage rather than in the galleries at any given time. Historically, museums have not developed a model for providing access to materials not on display in their galleries.

With the advent of networked access, museums started to image their collections and put versions of their in-house collections management systems online. While these online databases provide a more exhaustive view of an institutional collection, they usually restrict interoperability by locking the data into a proprietary system. Ultimately, access systems of individual institutions need to be integrated to provide maximum service to researchers and cultural tourists alike. A project led by the UC Berkeley Art Museum

called Museums and the Online Archive of California (MOAC) addresses this issue by encoding museum materials using archival and library standards, which allows the integration of data into a larger cultural space opened up by the Online Archive of California (OAC), hosted by the California Digital Library (CDL).

The project adapts the standards developed by other communities because the OAC union database had already been established around those specific submission information packages. Furthermore, the museum community had not yet (and still has not) established a file exchange format for collections information. Standards used in MOAC are Encoded Archival Description (EAD) xml, Metadata Encoding and Transcription Standard (METS) xml and Text Encoding Initiative (TEI) Lite xml.

The archival standard EAD describes collections hierarchically down to the item level. The xml mark-up allows them to successfully integrate their collections with library and archival materials. Since the EAD has no other uses for museums than to integrate data, archival principles pertaining to other internal functions the EAD fulfills in archives, such as authentication, documentation and collection management, lose their relevance in a museum setting. A typical museum EAD collection guide provides item-level access to a thematically structured collection enriched with educational materials.

Since EAD makes few provisions for encoding rich multimedia content at the item-level, the collection mark-up is extended by the digital object mark-up Metadata Encoding and Transmission Standard (METS). METS allows the encoding, transfer and display of hierarchically structured digital objects which mimic the behaviors of their analog counterparts. An artist's book, for example, may be presented to an end-user as a navigable set of digital images. METS has been developed by the library community, but its extensive use of extension schemas makes the mark-up flexible enough for easy implementation in other communities. A good example for METS adaptability through extension schemas is its approach to descriptive metadata. Each implementer may choose the descriptive metadata schema they feel most comfortable with to describe the information content of their METS object. The METS object also becomes a coordinating hub for other services provided alongside digital media surrogates, such as TEI Lite transcriptions.

From a museum perspective, this development may lead to an interesting shift in the role of collections for online access. Rather than providing an absolute home for items / digital objects, collections may now be viewed as services built on the digital object repository. Any single digital object may be part of multiple collections, which contextualize the digital object in a unique way. For example, a painting by abstract painter Hans Hoffman may be part of a provenance-based finding aid; a collection guide pulling together highlights from the UC Berkeley Art Museum's holdings; a collection guide thematically arranged around works from the particular school of abstract painting Hans Hoffman belongs to; etc. Each of these collections provides a different access path to the individual work. Viewed from this perspective, collections may move closer to the exhibition paradigm favored by museums—objects appear in ever-changing constellations of other objects in order to tease out particular facets of meaning.

DESCRIBING COLLECTIONS DIGITALLY: METADATA FRAMEWORKS FOR RESEARCH LIBRARY COLLECTIONS

Jared Campbell, University of California, Davis, Shields Library

Researchers and students have come to expect an increasing number of services via the World Wide Web from libraries and archives. Over the last two decades libraries have vast staff and financial resources to developing online catalogs and digitizing materials for remote use. The result has been the development of increasingly more powerful research tools for students and scholars that can be accessed from a simple networked computer twenty-four hours a day and from any geographic location. In addition to this, researchers are also expecting greater unmediated access to these collections. Despite this improved access, librarians and archivists continue to be faced with the challenge of how they help potential users find these collections and how they can describe them in a way that integrates access to digital and traditional physical collections.

This paper surveys current work being done in collection level descriptions for traditional and digital information resources. Focusing on the library and archival communities, we will look at several descriptive schemas that have been developed to provide access to collections of online resources, including the MARC bibliographic record format, the RSLP Collection Description Schema, the OAIS Harvesting protocol, and Encoded Archival Description (EAD). Focusing on resource discovery and analytic content description, the paper will compare the types of access each of these schema allows in terms of the detail of description, how easily data can be mapped or compared to other record schemas, and the ability for each to be integrated with one another.

The metadata formats examined are those concerned primarily with resource discovery: records that provide a summary or detailed description of the collection and its contents. The oldest and most widely used

of these is the Machine-Readable Cataloging (MARC) record. Created as a means of storing and transferring bibliographic data, it has been the standard data file format for library and many archival online catalogues. MARC, along with its content standard the Anglo-American Cataloging Rules 2nd ed. (AACR2), has been used to describe everything from books to multimedia web pages. The combination of MARC and AACR2 is useful because of their emphases on strong content and formatting standards. For instance, in a MARC record, the main title of a work will always be transcribed in a 245 field using AACR2 to formulate the entry. Though a patron may never see a record in its raw MARC form, they will recognize just about any web output that is the product of that record. Additionally, MARC and AACR2 place a strong emphasis on the need for name and subject control. Name headings and subject terms must come from an established thesaurus (such as the Library of Congress Headings, Library of Congress Name Authority File, or the Arts and Architecture Thesaurus) that has been identified in the record.

Though useful from an end user perspective, this rigidity causes problems when describing library collections rather then individual bibliographic works. This problem of scalability stems from the fact that MARC and AACR2 were developed to handle description of individual items. The collection level MARC record tends to consist of large unstructured textual note fields that make it difficult to systematically compare it to other collection level records beyond controlled subject vocabularies and names.

To remedy these problems, the archival community developed the Encoded Archival Description (EAD) as a means providing more detailed access to archives and manuscript collections. EAD is a SGML/XML DTD designed to capture both content and structure of archival finding aids. The use of SGML/XML allows archivists to encode the hierarchical document structure at a very granular level, facilitating cross collection searching. Another strength of the EAD is the ease with which it can be converted into a web friendly html document. The development of Extensible Style Language (XSL) allows institutions and individuals to present their collection data on the web in a variety of different ways. The major problem in implementing EAD is in the time and resources required to encode finding aids. Integrating EAD into an archives' or special collection department's descriptive program requires a tremendous amount of training and planning to fully utilize the DTD's potential. Moreover, the EAD DTD was designed specifically to be used with archival and manuscript collections. As a result, the structure and semantics used in these records come from that tradition and are not all that easy to adapt for other types of collections.

Another current means of gathering and disseminating collection level descriptions has been through the Open Archives Initiative (OAI) Protocol for Metadata Harvesting (PMH). This protocol is designed to automate the process of searching out scholarly information that may be hidden from traditional web browsers (Yahoo, Google, etc.). Common targets for the metadata harvesting include structured XML documents, finding aids, and resources that may be living inside of a database (photo images). Once harvested, records are then created and maintained in a central database. Depending on the metadata that is being harvested the OAI-PMH can create both collection level records and analytic records for the individual items within the collection. A main goal for OAI-PMH projects (University of Illinois, University of Michigan, and Virginia Tech) is to provide a central repository of structured data that allows for item level searching across institutional collections. Theoretically, a user could find records for two different transcriptions of the same text that may have been digitized and stored in two separate project text-bases.

The United Kingdom's Research Support Library Programme (RSLP) has produced a more recent development in collection level description. In an attempt to provide a more standardized cross-institutional collection level description, RSLP developed the RSLP Collection Description Schema. Developed as both a collections management tool and a resource discovery mechanism for scholars and students, the schema provides much more detailed description of collections then currently available in MARC. The result is a schema that brings together, in a single record, descriptive attributes about the collection itself, the location or locations of the resource, resource creators, agents (i.e. collector, owner, and administrator), and external relationships with other collections. Unlike MARC the RSLP schema is scalable in that it can be used to describe anything from a small collection of digital images to an entire library collection.

Based on these comparisons, the paper concludes with a discussion of the importance of early and continued planning of descriptive practices as the most important means of building and implementing flexible and robust collection level metadata. The paper suggests that as early as possible in this planning phase, project planners consider their text or image-base's primary and secondary audiences, the level of detail required to make sense of the collection, the scope of the materials. Creators of scholarly projects and collections should consult early with library metadata specialists early in the project planning process to discuss issues relating to how the existence of collection is made known, how collections can be linked (thematically, geographically, etc.) with others, and how potential researchers will be able to successfully navigate through various query interfaces.

Finally, potential collection users need to be included during the planning stages to take a more active roll in the development of useful metadata. This is especially true of descriptions of digital collections in which the goal is to provide unmediated access.

In providing an overview of current thinking about collection description in the library and archival world, this paper aims to provide a context for students, researchers, and other users of library and archival materials think think about their information needs as they undertake humanities research. These needs should determine the extent to which current and new schemas are appropriate for their own research projects or whether new frameworks need to be created.

DESCRIBING WHAT'S THERE: TOWARD A DESCRIPTIVE FRAMEWORK FOR LEGACY LIBRARY COLLECTION DESCRIPTIONS

Nancy Kushigian, University of California, Davis/California Digital Library

In the days when library catalogs consisted of card files, most research libraries housing "special" collections of material found ways to describe those collections in more detail than a MARC record could provide. "Collections," broadly defined as groupings of discrete library items, originate for a variety of reasons: a scholar takes an interest in a topic and buys every relevant book; a bookseller decides that a genre or author text is more interesting in large quantities than in scattered instances, so he collects, for example 14,000 volumes of British poetry and sells them as a group to a research library; a small special library goes "out of business" and gives its' collection of radical pamphlets larger library. In all these cases, groupings of books or pamphlets or photographs are often described in collection descriptions. The originating principle of the collection is known as its' "provenance."

Collection descriptions are created by many different kinds of writer: booksellers, collectors, documentary editors, bibliographic scholars, librarians, archivists, and others. They serve as an aid to researchers in approaching and understanding a particular group of material. In library departments of special collections, they often replace the library catalog, as the main printed source of information about to the contents of a particular collection.

Traditional printed collection descriptions vary widely. Sometimes, copies of card catalogs are produced together in book format and supplemented with commentary or introductory material. Sometimes, glossy collection descriptions are created for prospective donors. Sometimes, these descriptions consist of descriptive bibliography. Sometimes, as in exhibit catalogs, they are meant to serve as a surrogate, as a way of preserving an "artificial collection" brought together from many different libraries.

Printed collection descriptions exist in great number, and form a significant percentage of the collection of every research library. While traditional printed collection descriptions do not describe each item in a collection exhaustively, they do provide a rich source of descriptive and bibliographic "finding" information to a researcher who may be considering whether or not to visit a particular library or archive. Increasingly, however, these storehouses of information gather dust, as students and scholars turn to online sources of information, often proceeding no further in their research efforts.

Thus, there is a need to preserve and make accessible online these many historical or "legacy" printed collection descriptions. Clearly, such a text-base, were it encoded in such a way that it could be searched across repositories, would provide a rich and valuable resource for students and scholars seeking access to rare or primary materials. Current metadata collection descriptive frameworks such as METS, OAI, and RSLP do not provide descriptive elements adequate to capture the variety and complexity of these "legacy" printed collection descriptions.

Based on the observation that the Encoded Archival Description, narrowly designed for a particular type of collection description, is now widely adapted for non-archival collections, the paper argues that there is need of a broader, more flexible, descriptive framework that could be used to encode both new and legacy collection descriptions that vary widely in their content and structure. EAD, while suited well for the encoding of a particular type of description, the archival finding aid, lacks basic bibliographic and structural elements necessary to encode complex texts.

Because of its scope and descriptive granularity, the Text Encoding Initiative seems a promising candidate for such a descriptive scheme. But in order to ensure interoperability, the TEI encoding of legacy collection descriptions needs to be regularized. Indeed, we seek a solution that transcends the limits of the EAD, but constrains the possibilities of TEI.

To begin to develop such a best practices framework, the author of the paper analyzed 50 legacy collection descriptions found on the shelves of the University of California, Davis Shields Library, chosen to represent the widest possible range of formats and types of collection. The paper presents a basic taxonomy of functional types of information in these legacy printed collection descriptions and proposes that the library community work to develop a "TEI-Like" dtd, best practices standards, and implementation guidelines to allow these and other descriptions to be encoded and made accessible.

Finally, the paper argues that such a descriptive framework could be used as a guide for the creation of new online collection descriptions, and that the existence of such a framework would allow special

collections libraries and curators to create standard, familiar, and extensible user “web page” descriptions of their major collections. Such online descriptions could form part of a large, inter-operable text-base that included legacy collection descriptions, and could thus form the basis for a seamless bridge between our intellectual inheritance and future collections.

XML Schema 1.0: A Language for Document Grammars

C. M. SPERBERG-MCQUEEN
World Wide Web Consortium / MIT Laboratory for Computer Science
cmsmcq@w3.org

The Standard Generalized Markup Language (SGML) and its offspring the Extensible Markup Language (XML) appear to be fairly well established as methods of representing texts in electronic form.[1] One of the characteristic features of SGML and XML which marks them as an advance over earlier systems of textual representation is their notion of document grammars: formal specifications of rules for distinguishing valid documents from other data streams. Document grammars prove useful in routine quality assurance (finding and cleaning up dirty data), in documentation of agreements between data providers and data consumers or of the contents of data flows, and as a means of specifying the contents of messages in client/server protocols.[2]

In some respects, however, the notation defined by SGML and XML for document type definitions (DTDs) has proven to have some shortcomings.

- The use of a specialized notation rather than SGML or XML itself means that standard tools like XSLT and XPath processors cannot be used straightforwardly to work with DTD files.
- The availability of data typing for attributes, but not for *#PCDATA* content of elements, introduces an unnecessary complication and lack of parallelism into the comparison of elements and attributes.
- From a programming-language or database point of view, the selection of data types available for attributes may charitably be described as eccentric: it has strings (more or less) and various abstrusely specialized forms of tokens, but lacks integers, floating-point numbers, dates, and other standard types.
- Although a number of published DTDs (e.g. that of the TEI) rely explicitly on notions of class and inheritance similar to those used in object-oriented systems, DTD notation lacks explicit support for inheritance.
- Even if DTD notation did support inheritance, there is no standard way for applications to ask SGML/XML systems for information about the DTD used to validate a document.
- DTD notation does not do at all a good job of supporting XML namespaces, which are increasingly important as a means of supporting compound documents and the mixture of different XML vocabularies in the same document.

For these and other reasons, there has in recent years been a good deal of interest in new languages for specifying document grammars [Bourret et al. 1999, Bray et al. 1998, Frankston/Thompson 1998, Layman et al. 1998, OASIS 2001]. XML Schema 1.0 is a non-proprietary schema language developed by the World Wide Web Consortium; work began in 1998, the specification became a W3C Recommendation in May 2001 [W3C 2001], and further development continues today. This paper will offer a brief introduction to XML Schema 1.0 and describe its salient features.

Unlike DTDs but like most recent schema languages, XML Schema 1.0 uses an XML vocabulary, rather than an ad hoc specialized non-XML notation to represent document grammars. This makes XML Schema documents more verbose than equivalent schemas in DTD notation but also makes them much more easily processable.

XML Schema provides explicit support for XML namespaces and for combining XML vocabularies from different namespaces into a single composite schema. Given the increasing use of namespaces to minimize name conflicts between vocabularies, the inability of DTDs to handle this task adequately has become a more and more distressing deficit.

DTDs intermingle several functions: in addition to defining constraints on the logical structure of marked up documents, they also include entity declarations, which affect the initial scanning of the XML data stream. XML Schema, by contrast, assumes that a standard XML processor has already processed the XML document before schema-validation is started: the input to an XML Schema validator is not an XML

document in the strict sense, but an *XML information set*, which may be produced by parsing an XML document or by other means, such as the construction of a data structure in memory through function calls to an API. The output of an XML Schema validator is the same information set, augmented with information about the validity of each element and attribute in the document and about the validation episode itself. Defining schema validation as a mapping from an input information set to an output information set has advantages for the conceptual clarity of the specification, but it has also proven unpopular with some users, because it means that DTD notation must still be used to declare human-readable names for special characters, and there is no prescribed XML form for the additional information about validity and datatyping produced by the XML Schema validator.

XML Schema provides a basic set of predefined "simple" datatypes, which can be associated with attribute values or with elements whose content is a simple character string without sub-elements. In addition to the legacy types inherited from XML, XML Schema provides types which correspond to those most commonly found in programming languages and database management systems: exact decimal numbers and integers, floating-point and double-precision numbers, dates and times (in the standard notation defined by ISO 8601), and some other more specialized datatypes. Schema authors can define new simple types by restricting existing ones in certain well defined ways. They cannot, however, create new primitive types; this has advantages for interoperability and disadvantages for the expressive power of the language: the TEI `date` element, for example, can use the XML Schema date type to describe the value of its `value` attribute, which is required to use the ISO standard date format, but not to describe its content, which also denotes a date but which does not use the standardized notation.

In addition to *simple* types, schemas can also define*complex* types, for elements which can contain sub-elements; complex types correspond to the content models and attribute declarations of DTDs. From object-oriented systems, however, XML Schema has adopted the concept of class inheritance: it is possible to derive new complex types from existing complex types, just as it is possible to derive new object classes from existing classes in an object-oriented programming language. Experience with DTDs shows that two quite separate kinds of inheritance may be needed for document grammars: one in which the derived type inherits some properties of its content model and attributes from the ancestor types, and another in which what is inherited is the ability of an element to occur in particular locations. (The TEI models these two different kinds of inheritance by distinguishing attribute-classes and model-classes.)

Perhaps the most important innovation in XML Schema 1.0 is that schema-based validation provides much more information than the simple yes/no is-this-valid? information provided by DTD-based validation. Information about the simple or complex type assigned to an attribute or element is provided by an XML Schema processor as part of the standard post-schema-validation information set (PSVI). The validity of each element and attribute is checked and recorded separately; this entails a distinction between the concept of full validity, which is recursive and requires that all descendants also be fully valid, and of local validity, which is not recursive. Since schema validation need not start at the root element of the document, and since a schema can direct that the contents of particular elements are not to be validated, or that the elements encountered in particular contexts need not be declared, XML Schema 1.0 can be said to have introduced a coherent concept of partial validation; whether it can be exploited to handle problems of structural variation in historical documents [Birnbaum 1997, Birnbaum/Mundie 1999] remains to be explored.

The paper will conclude with a brief account of current work on XML Schema within the World Wide Web Consortium.

NOTES

[1.] This is not to ignore the recent work done by Patrick Durusau and Matthew Brooke O'Donnell on Just-In-Time Trees [Durusau/O'Donnell 2002a, 2002b], by Wendell Piez and Jeni Tennison on LMNL (Layered Markup and Annotation Language) [Piez/Tennison 2002], by Andreas Witt on the representation of concurrent markup structures in logical form [Witt 2002], or by Claus Huitfeldt and C. M. Sperberg-McQueen on TexMecs (Trivially Extended MECS (Multi-Element Code System)) [Sperberg-McQueen/Huitfeldt 2001]. All of these projects retain their interest, but at the moment most appear to be experimental systems rather than fully developed alternatives to SGML and XML.

[2.] This last usage is now prominent in work on the Simple Object Access Protocol and other Web-services work, but the ideas predate the current interest in Web services [Catteau 1999].

REFERENCES

Birnbaum, David J. *In defense of invalid SGML*. Paper given at ACH/ALLC 1997.
http://clover.slavic.pitt.edu/~djb/achallc97.html

Birnbaum, David J., and David A. Mundie. "The problem of anomalous data: A transformational approach". in *Markup Languages: Theory & Practice* 1.4 (1999): 1–19.

Bourret, Ronald, et al., ed., “Document Definition Markup Language (DDML) Specification”, Version 1.0, Submission to the World Wide Web Consortium, 19-Jan-1999. **http://www.w3.org/TR/NOTE-ddml**
Bray, Tim, Charles Frankston, and Ashok Malhotra, ed., Document Content Description for XML, Submission to the World Wide Web Consortium 31-July-1998. **http://www.w3.org/TR/1998/NOTE-dcd-19980731**.
Catteau, Tom. “An SGML system for the budget of the European Union”. in *Markup Languages: Theory & Practice* 1.3 (1999): 41–59.
Cowan, John, and Richard Tobin, ed. 2001. “XML Information Set”. W3C Recommendation 24 October 2001. [Cambridge, Sophia-Antipolis, Tokyo]: World Wide Web Consortium. **http://www.w3.org/TR/xml-infoset/**
Durusau, Patrick, and Matthew Brooke O’Donnell. “Visualizing overlapping hierarchies in textual markup”. Paper given at ALLC/ACH 2002, Tübingen, July 2002. **http://www.uni-tuebingen.de/cgi-bin/abs/abs?propid=100**
Durusau, Patrick, and Matthew Brooke O’Donnell. “Coming down from the trees: Next step in the evolution of markup?” Paper given at Extreme Markup Languages 2002, Montréal, August 2002. **http://www.idealliance.org/papers/extreme02/author-pkg/2002/Durusau01/EML2002Durusau01.zip**
Davidson, Andrew, et al., “Schema for Object-oriented XML 2.0”, W3C Note, 30 July 1999. **http://www.w3.org/TR/NOTE-SOX/**
Frankston, Charles, and Henry S. Thompson, ed., “XML-Data reduced”, Draft Version 0.21, 3 July 1998. **http://www.ltg.ed.ac.uk/~ht/XMLData-Reduced.htm**
Huitfeldt, Claus, and C. M. Sperberg-McQueen. “TexMECS: An experimental markup meta-language for complex documents”. [Working paper of the MLCD project at the University of Bergen]. Bergen: [n.p.], 2001. **http://www.hit.uib.no/claus/mlcd/papers/texmecs.html**
ISO (International Organization for Standardization). ISO 8601. Representations of dates and times. 1988–06–15. Available at: **http://www.iso.ch/markete/8601.pdf**
Layman, Andrew, et al., “XML-Data”, W3C Note [Acknowledged submission], 05 Jan 1998. **http://www.w3.org/TR/1998/NOTE-XML-data-0105**.
OASIS (Organization for the Advancement of Structured Information Standards). “RELAX NG Specification”. Committee Specification 3 December 2001. **http://www.oasis-open.org/committees/relax-ng/spec-20011203.html**
Piez, Wendell, and Jeni Tennison. “The Layered Markup and Annotation Language (LMNL)”. Paper given at Extreme Markup Languages 2002, Montréal, August 2002. Project home page at **http://www.lmnl.org/**
Text Encoding Initiative. Guidelines for electronic text encoding and interchange (TEI P4), ed. C. M. Sperberg-McQueen and Lou Burnard. XML-compatible edition prepared by Syd Bauman, Lou Burnard, Steven DeRose, and Sebastian Rahtz. Oxford, Providence, Charlottesville, Bergen: TEI Consortium, 2002.
Witt, Andreas, “Meaning and interpretation of concurrent markup”. Paper given at ALLC/ACH 2002, Tübingen, July 2002. **http://coli.lili.uni-bielefeld.de/Texttechnologie/Forschergruppe/prolog/allc2002-witt.html**
W3C (World Wide Web Consortium). “XML Schema Part 0: Primer”, ed. David Fallside. “XML Schema Part 1: Structures”, ed. Henry S. Thompson, David Beech, Murray Maloney, and Noah Mendelsohn. XML Schema Part 2: Datatypes, ed. Biron, Paul V. and Ashok Malhotra. W3C Recommendation, 2 May 2001. [Cambridge, Sophia-Antipolis, Tokyo: W3C] **http://www.w3.org/TR/xmlschema-0/**, **http://www.w3.org/TR/xmlschema-1/**, **http://www.w3.org/TR/xmlschema-2/**

Text Markup—Data Structure vs. Data Model

ALLEN RENEAR
GSLIS/University of Illinois
renear@uiuc.edu

SUMMARY

For over ten years there have been several carefully argued critiques of the descriptive markup approach to text representation that have for the most part not been adequately answered. Recently an integrated development of some of these criticisms has been presented at length, and with considerable ingenuity and

extension, in a widely read article by Dino Buzzetti, published in *New Literary History* (2002, 33: 61–88). In our paper we analyze some of Buzzetti's criticisms, accepting some but resisting others.

BACKGROUND

When the SGML or "descriptive markup" approach to text began to be actively promoted in the mid-1980s, quite a number of criticisms were developed in response to this approach, or, more exactly, to the accompanying theoretical claims, express or implied. In retrospect, it appears today that some of these criticisms have been answered by the proponents of SGML descriptive markup, others have simply faded in apparent significance, and still others, while live issues to some scholars, seem neverless to generate only unproductive repetitive discussions.

However there is a particular family of related criticisms that upon careful re-consideration still seem fresh, deep, and, at least given their positive reception, plausible—and yet these criticisms have also been, strangely, largely ignored by the markup community. Some of these arguments were first presented in the late 1980s in a paper titled "Markup Considered Harmful" privately circulated by Darryll Raymond in response to Coombs et al 1987. They were later presented again, in more detail in 1992 in "Markup Reconsidered" (Raymond, Tompa, and Wood 1995); and then were further refined in "From Data Representation to Data Model: Meta-Semantic Issues in the Evolution of SGML" (Raymond, Tompa, Wood 1996).

In brief these papers argue that "Markup is not a data model" but rather a "data representation" (Raymond et al 1992), and that while "SGML provides standard representations for documents" a "standard semantics" is required as well (Raymond et al 1996). And then, generalizing from this conclusion, the authors go on to suggest that many of the expectations for SGML-based markup systems, such as TEI, are misplaced.

Despite the fact that these were pointed and clearly argued positions, the response from the TEI community, and perhaps more importantly from the theorizing researchers whose ambitious claims were particularly targeted (Coombs et al 1987, DeRose et al 1990, Sperberg-McQueen 1990), appears to have been largely to simply ignore these criticisms.

Recently Dino Buzzetti has been re-presenting these arguments, supplementing them with further theoretical extensions, and integrating them with other criticisms of descriptive markup approach, in particular those of Jerome McGann, most recently presented and extended in a new book *Radiant Textuality: Literature Since the World Wide Web* (New York 2001)—which Buzzetti takes as supporting Raymond's and his own criticisms. Buzzetti has in fact presented his concerns in several venues over the last 5 years (e.g. Buzzetti 1999a, Buzzetti 1999b), but it is in "Digital Representation and the Text Model" (*New Literary History*, 2002 33) that he presents these criticisms in a form, and a forum, and a language, in which they can no longer be ignored.

Interestingly, this article comes precisely at a moment when some of the theorizing descriptive markup proponents are themselves very actively working in a similar vein. Sperberg-McQueen, who has called for the development of a SGML/XML semantics since at least 1992 (Sperberg-McQueen 1992), is leading a project to develop such a semantics now, and Renear, who is member of this project, has also recently claimed, in language similar to that of Buzzetti and Raymond et al, that an SGML document instance per se provides "only a data structure, not a theory" (Renear 2001).

THE CRITICISM

Raymond et al argue, as described above, that markup is not a "data model", but only a data representation that is "fully dependent on external information for meaning". They note several characteristics that they say confirm that markup is not a data model in the sense that that expression is used in database theory. These include inadequate notions of equivalence, lack of redundancy control, and lack of defined algebraic operators. This claim should not of course be taken simply as a narrow assertion that SGML markup does not meet criteria for being a certain sort of thing as defined in some field or other—the larger point, and in the context of the discussion a plausible one, is that SGML markup will be inadequate for the kinds of theoretical tasks it is being assigned in virtue of failing to have these characteristics. Raymond et al (1995) state that this failure can be remedied either by adding a "formal structure on top of markup systems" or by defining a new abstraction from scratch. Raymond et al prefer the latter because they believe that markup systems are not only a representation rather than a data model, but they are in fact a relatively poor technique for representation. Raymond et al 1996 characterizes this addition of a formal structure as adding "semantics" to SGML.

Buzzetti accepts Raymond et al's criticism of SGML markup. But he goes much further, both in the extent of his theorizing and in the specific nature of his criticisms. With respect to the latter Buzzetti argues that descriptive markup theorists are actually confused, and profoundly so: they conflate the "structure of the representation" with the "structure of the object represented", or, alternatively, they confuse the "expression" with the "content" of that expression. Supplementing Raymond's criticism from the perspective of database theory with an opposition from Saussure Buzzetti argues that descriptive markup is an expression whose form is a data structure. But the form of the content of that expression is a "data model". Neither the expression

itself nor its form however, is a data model—as mistakenly believed by the promoters of SGML descriptive markup and the defenders of the OHCO model of text. (Buzzetti goes on to make this point with juxtapositions of other theoretical terms: “format” vs. “formalism”, “syntax” vs. “interpretation”). As examples of this confusion Buzzetti singles out in particular the TEI Guidelines, Coombs et al 1987, and DeRose et al 1990, and Renear et al 1996.

“In summary,” Buzzetti says, “we may assert that strongly embedded markup systems [are] inadequate in reference to both the exhaustivity and the functionality of the text representation and model.”

OUR RESPONSE

We believe that that there is some truth to both Raymond, Tompa, and Wood’s criticisms, and to Buzzetti’s application of the criticisms. However we believe that both Raymond et al and Buzzetti substantially overstate the case, Raymond somewhat and Buzzetti much more. Such a response may seem too mild to deserve public airing, but it is for two reasons much more important than it sounds: first because of the radical form the criticism takes in Buzzetti, and second because our response might clarify some long confused issues in text encoding.

There can be little doubt that many SGML markup enthusiasts have been now and then confused, and, even more frequently, confusing, about the representational nature of SGML document instances. This is mostly because representation itself is conceptually difficult, and subject to a kind of semiotic oscillation between expressing, referring, and exhibiting. And it may also be partly because of carelessness and genuine lapses of clarity and understanding.

But Buzzetti makes too much of the fact that we are sometimes awkward about what document instances are, representationally, and exactly how markup does what it does. Notice that for the most part everyone manages the ambiguities quite well, and gets on, successfully more often than not, with various tasks and projects. How could this be the case if there were such a consistent systematic widespread conflation of expression and content? And, in fact, a close examination shows that Buzzetti does not produce evidence for such a systematic conflation, other than the various awkward or unfortunate characterizations mentioned above—and those, while admittedly signs of some failure to fully conceptualize key notions, are far from convincing evidence for systematic conflation.

Moreover, there is no evidence that this confusion, such as it is, is connected, as either cause or effect, with a radical unsuitability of SGML markup to express theories about text. (According to Buzzetti “SGML is a data structure representation language” and does not provide a data model; and he echoes the views of Raymond et al that it is, moreover, a bad data structure representation language.) Part of the problem is that Raymond and Buzzetti rely too much in their argument on the specific notion of a “data model” from database theory, and therefore over-react to the failure of SGML to have precisely that sort of data model, or to have a good (i.e. completely or exactly defined) one. Part of the problem is that they demand formalization before admitting that there has been sucessful representation. It cannot be denied that we use SGML document instances to express express facts and theories about texts rather effectively: that is simply an empirical fact. Part of the problem is that in claiming that SGML and XML lack semantics and a defined set of operators, Raymond et al and Buzzetti ignore the fact that both SGML and XML define the semantics of the markup vocabulary in use as an intrinsic part of the application's document type definition. Unlike Raymond et al and Buzzetti, the SGML standard does not assume that the full semantics of an SGML application are captured in full by the SGML formalisms used to declare elements and attributes. Nor does a demand for an exhaustive set of algebraic or other operations seem a promising start for a critique of a markup discipline intended to encourage the reuse of data and the separation of data representation from processing concerns.

In this connection Buzzetti’s criticism of the ordered-hierarchy of content objects (OHCO) hypothesis is particularly illuminating, because strikingly unconvincing. Unlike SGML markup, OHCO is obviously not intended as a representation syntax, but rather an abstraction claimed to be, structurally, the general form taken by all texts. As such it actually appears to be roughly the kind of thing that Buzzetti wants for a data model, though without the specific features characteristic of data models in database theory. Buzzetti’s interpretation of OHCO however has it as “not a model of the text, but a possible model of its expression”, just as SGML on Buzzetti’s account, describes not the text, but “the structure of the text’s expression”, consistent with its distinctive role as “a data structure representation language”. But whatever the failures of OCHO as a data model, this is no more than a tortured effort to make things seem worse then they are. The data structures in question have untyped purely logical parent/child relationships, whereas “hierarchy” in OHCO clearly refers to *containment*—and that is not a purely data structural relationship, despite the obvious formal similarities (asymmetry, transitivity, etc.) that makes tree data structures convenient for representing hierachical containment.

It is true that additional formalization of the semantics of SGML markup is a good thing. Formalization will assist in precision, clarity, and computation. It is for that reason that Sperberg-McQueen

and others are working to explicate the semantics of XML markup. But while we may regret that such work has not made more progress, or that not everyone understands the need for improved formalization of semantics and "data models", or that some of our colleagues are sometimes confused about subtle matters of representation, or that we ourselves are now and then confused, or even often confused, we should not conclude that things are worse than they are. Scholars working in SGML descriptive markup are working with languages that do have semantics, and do have "data models", even if inadequately formalized, or not quite the sort of thing database researchers prefer. And these scholars are at least as often as not quite clear on the difference between expression and content.

CONCLUSION

The criticisms of inadequate formalization that began with Raymond et al and that have recently been so intriguingly developed and extended by Buzzetti are important ones that have been mysteriously neglected by the markup community. And the more ambitious criticisms made by Buzzetti of our failure to fully theorize, or at least attempt to clarify, the complexities of representation are perhaps even deeper and more important, and also unfortunately neglected. But the claim that in these failures we see the signs of either a systematic category mistake endemic in the text encoding community, or evidence of a disastrous inadequacy in current techniques, simply has not yet been proven. On the contrary, as we have suggested here, that claim itself seems to be based upon the mistaken notion, ironically reminiscent of the third man argument, that one more formal system will finally bridge the gap between thought and object. There may or may not be a gap, but if there is it will not be closed simply by another formal system, however useful that system may be.

REFERENCES

Buzzetti, "Text Representation and Textual Models," ACH-ALLC'99 Conference Proceedings (Charlottesville, 1999).

Buzzetti, "Digital Representation and the Text Model", *New Literary History*, 33. 2002.

James H. Coombs, Allen H. Renear, and Steven J. DeRose, "Markup Systems and the Future of Scholarly Text Processing," *Communications of the ACM*, 30:11, November, 1987.

Steven J. DeRose, David G. Durand, Elli Mylonas, and Allen H. Renear, "What Is Text, Really?," *Journal of Computing in Higher Education*, 1:2, 1990

Darrell R. Raymond, Frank W. Tompa and Derick Wood, "Markup Reconsidered," *First International Workshop on Principles of Document Processing*, Washington, D.C., October, 1992).

Darrell R. Raymond, Frank Tompa and Derick Wood, "From Data Representation to Data Model: Meta-Semantic Issues in the Evolution of SGML," *Computer Standards and Interfaces*, 18, January, 1996.

Steven J. DeRose, David G. Durand, Elli Mylonas, and Allen H. Renear, "What Is Text, Really?," *Journal of Computing in Higher Education*, 1.2, 1990

Allen Renear, David Dubin, C. M. Sperberg-McQueen, and Claus Huitfeldt, "Towards a Semantics For XML Markup" *Proceedings of the ACM Symposium on Document Engineering ACM*, November 2002.

Allen Renear, Elli Mylonas, David G. Durand, "Refining Our Notion of What Text Really Is: The Problem of Overlapping Hierarchies," in *Research in Humanites Computing*, (Oxford 1996).

Allen Renear, "Raising the Bar, Text Encoding from a Logical Point of View", CLiP 2001.

C. M. Sperberg McQueen, "Back to the Frontiers and Edges," Closing remarks at "SGML '92: the quiet revolution" conference sponsored by the Graphic Communications Association (GCA), Boston, 29 October 1992. Available on the Web at **http://www.w3.org/People/cmsmcq/1992/edw31.html**

C. M. Sperberg McQueen, Claus Huitfeldt, and Allen Renear, "Meaning and Interpretation in Markup," *Markup Languages: Theory and Practice* (MIT Press 2000).

C. M. Sperberg-McQueen, Allen Renear, Dave Dubin, Claus Huitfeldt, "Drawing Inferences on the Basis of Markup" Proceedings of Extreme Markup 2002. Jerome McGann, *Radiant Textuality: Literature Since the World Wide Web*, (New York, 2001).

Anastasia: A New XML Publication System

PETER ROBINSON
De Montfort University
peter.robinson@dmu.ac.uk

Over the last decade, many humanities scholars have been persuaded by the promise and the power of encoding schemes for electronic texts to create texts, sometimes very large and complex, encoded using these schemes. This is specially true of SGML/XML based encodings, with the implementation of the Text Encoding Initiative being particularly influential in the community. However, scholars who have made such texts have typically discovered that software to publish them is too expensive for their limited budgets, or too difficult to use, or lacking essential facilities, or all three of these.

The *Anastasia* electronic publishing system, developed in the last five years in a partnership between the Centre for Technology and the Arts, De Montfort University, and a new electronic publishing company, Scholarly Digital Editions (SDE), attempts to supply this deficiency. *Anastasia* stands for 'Analytic system tools and SGML integration application'. As this implies, it is able to handle all valid SGML and XML documents, with no limits on their complexity.

Particularly, *Anastasia* has been designed to meet the needs of humanists, and especially textual scholars. It is a common complaint of humanists that SGML/XML systems constrain a single hierarchical view of a document, while humanities texts can be seen as containing many overlapping and competing hierarchies. SGML/XML publishing systems usually cannot support facilities which cut across the primary document hierarchy, and so cannot satisify even such simple needs as display of a single page of transcription, or display of a tabular list of key word in context search results with formatting of all returned search strings according to the embedded encoding. *Anastasia* seeks to escape these limitations by adopting a document processing model that sees the document as made up of a series of events which are defined not only by their hierarchical relation, but also by their left to right relation in the document stream. As a result, Anastasia provides tools which allow the document to be manipulated according to alternative hierarchies implicit in the element relations. Thus, one can very easily extract views of the text by column or page, or indeed start a display at any point in any element and continue to any point in any other element. A KWIC display, for example, requires that we display an arbitrary number of characters before a hit, then display the characters in the hit themselves, and then display an arbitrary number of characters after the hit, all with complete awareness of the document encoding within those spans of characters: *Anastasia* can do this. Then, one should be able to click on a link from the KWIC display to the document itself, and see the hits highlight in the full-text context: once more, Anastasia has been designed to make this easy. One can also manufacture virtual texts by extracting and combining multiple and even overlapping segments.

Anastasia is also designed to fill another need: for a mode of publication which is identical on both CD-ROM and the internet, on the major Windows and Macintosh systems. Typically, the scholar will prepare a body of SGML/XML documents for publication using the *Anastasia* GroveMaker application, which compiles the documents into a binary database. The *Anastasia* Reader then serves the documents to an internet browser, either over a network or from a CD-ROM. Control of all aspects of the publication's display and behaviour (including fully SGML/XML aware searching) is achieved through a series of Tcl script files.

A key factor in the development of *Anastasia* has been the desire to achieve publication without compromise. That is: if it is possible to achieve a certain kind of computer display effect, then *Anastasia* will allow this. For example, we might want to use some of the advanced dynamic HTML features permitted by *Javascript*: pop-up menus, text which changes colour as the mouse passes over some other part of the document (for instance, to show that a word or phrase in one window is a translation of, or is otherwise related to, a word or phrase another window), synchronous scrolling or separate windows, and more. Practically, this means that we should be able to generate streams for display in any format whatever, directly from the XML: in pdf, SVG, rtf, any variant of HTML and XML, and send it directly to the display engine. We have concentrated on using Anastasia to generate HTML with Javascript: an example of the effects possible through this can be seen in the work on the digital 28th edition of the Nestle-Aland Greek New Testament, accessible through nestlealand.uni-muester.de. Other instances can be seen from the SDE website, www.sd-editions.com/anastasia.

Anastasia is designed to work as a Apache webserver module. It also requires C-language support, and the Tcl (Tool Control Language) libraries. In theory at least, this means *Anastasia* can operate whereever

Apache operates: our main development is on Macintosh OS X and Windows machines; there is also a Linux port. The search systems in *Anastasia* are based on SGREP, written by Jani Jaakkola and Pekka Kilpelainen of the University of Helsinki: we have heavily customized the SGREP code to improve its performance with large texts. Perhaps one of the most distinctive (if not controversial) features of *Anastasia* is that the style sheets we use to control exactly how the source XML is sent to the browser are written in Tcl, and not in any of the various XML-based systems which have appeared in the last years (such as XSLT, XPATH, and others). In part this is historical: the roots of *Anastasia* lie some distance back, as far as the first work done by myself on the *Canterbury Tales Project* with Elizabeth Solopova and Norman Blake) in 1993, long before even XML made an appearance. In part, it is because those systems themselves remain in a start of flux. But it is also because there is room for argument about the efficiency of such schemes. There is no doubt that XML is superb at representing textual structure. But this does not mean it is suitable for use as a programming language, requiring ease of use, rapid development, efficient maintenance, and widespread support across many different computer systems. Tcl does offer all these.

Anastasia is not intended to be the tool of choice for everyone who works with XML. It is designed for situations where the very best possible presentation is required of highly complex XML. A single screen of the digital Nestle-Aland, for example, may draw XML from hundreds of different places within the source, reformat into HTML interwoven with Javascript commands, and spread this across a series of frames nested within the browser display Ð all in a fraction of a second, in response to a request from the reader. It is also designed to run identically on CD-ROM and over the internet. Reports of the death of CD-ROM appear rather exaggerated: indeed, the availability of cheaply priced publication tools such as *Anastasia* may make it possible for high-quality CD-ROMs to be made available at much lower prices than hitherto, and so create a market which has been previously elusive. Finally, my hope when designing *Anastasia* was that a single scholar, with reasonable dedication, good knowledge of XML and with no more computer resources and support than are commonly available within university departments, would be able to use it to make high-quality XML based publications. There have been some encouraging signs that *Anastasia* can indeed be used in this manner. In the same context, it should also be appropriate for use by smaller academic publishers.

This is the first conference presentation of *Anastasia* as a mature publication system. There has been one previous conference presentation of the system, at the DRRH conference in Sydney in September 2001, when only a preliminary version of the software was available.

New Ways in Using and Creating Lexicographical Resources

MATTHEW S. GIBSON
Electronic Text Center, University of Virginia
mgibson@virginia.edu

UTE RECKER-HAMM
University of Trier
recker@uni-trier.de

THOMAS SCHARES
University of Trier
schares@uni-tirer.de

FRANK QUEENS
University of Trier
queensf@uni-trier.de

The three paper session focuses on the lexicography of German, on dictionaries to Middle High German and their primary textual sources, and on the 33 vols. of the Deutsche Woerterbuch by the brothers Jacob and Wilhelm Grimm and now electronically available.

The first paper by Matthew S. Gibson and Ute Recker-Hamm on “Middle High German Interlinked: A Comprehensive Text Archive” reports on an US-German collaboration project funded by the NSF and the DFG. The aim of it is the creation of an XML-encoded comprehensive archive of Middle High German

(MHG) texts representing the best scholarly editions. These are made available electronically together with their editorial additions, especially with their glossaries which are interlinked with an already existing compound of MHG dictionaries (Mittelhochdeutsche Woerterbuecher im Verbund, **http://www.MWV.uni-trier.de**). The MHG Text Archive with its links to glossaries of editions and to the major MHG dictionaries offer on the one hand most extensive reading helps to the student of MHG texts, and on the other hand an ideal working environment to the lexicographers working on the new MHG Dictionary.

The second paper by Frank Queens and Ute Recker-Hamm on "An Internet-based working environment for the production of dictionaries in distributed locations" focuses on the preparation of tools for lexicographers working in different places and making extensive use of the opportunities the computer and Internet have opened up during the last decade. Generations of lexicographers before our time put much effort into the creation of slips and sorting and storing them in boxes. The electronic environment allows the lexicographers to concentrate much more on their proper work, that is the examination of quotations and the writing of entries. The preparation of a citation corpus is now much easier by using electronic texts; the publication of the installments of a dictionary in print no longer involves the drudgery of thorough quotation checking and multiple proofreading. Moreover the electronic publication of dictionaries opens up new dimensions of reading technologies.

The third paper by Thomas Schares on "Electronic Dictionaries and Metalexicography: The Digital Version of the Deutsche Woerterbuch (=DWB) by Jacob and Wilhelm Grimm as a Basis for Metalexicographical Research" focuses on the history of the great national dictionary to the German language, the Deutsche Woerterbuch by the brothers Grimm. It took over 100 years to complete this work. It was begun by the two brothers, who are also famous for their collection of fairy tales. After them generations of lexicographers have worked on this monumental dictionary. Therefore it has no consistent overall structure, but shows the predilections of the various lexicographers in their time and thus mirrors the history of German philology. However, the inconstistencies and peculiarities can exactly be studied and pointed out now as we have the electronic DWB (**http://www.DWB.uni-trier.de**), and as its thorough encoding opens up new ways of metalexicographical research.

MIDDLE HIGH GERMAN INTERLINKED: A COMPREHENSIVE DIGITAL TEXT ARCHIVE

Matthew S. Gibson, Ute Recker-Hamm, M.A.

Electronic texts of all kinds can be found on the World Wide Web today, even texts in medieval German. However, the quality and reliability of these texts vary. Some websites offer texts with typographical markup (e.g. Gutenberg.de or Bibliotheca Augustana) and some provide texts with elaborate retrieval techniques where the results of searches are electronically linked to more robustly encoded full texts (e.g. the Mittelhochdeutsche Begriffsdatenbank). But regardless of the level of encoding, most electronic text archives concentrate only on primary texts; editorial additions (such as paratextual material such as introductions, critical apparatuses) are omitted as well as the lexicographical aids in glossaries and commentaries to the edited texts. In the case of medieval texts, such editorial components are essential for the academic to gain the most valuable reading and understanding of medieval vernaculars.

The central goal of the project, Middle High German Interlinked/ Digitales Mittelhochdeutsches Textarchiv, is to provide comprehensive textual editions (including introductions to each text, the primary texts themselves, word glossaries, and other editorial additions) and to link these electronic editions to external lexicographical resources. In the end, the project will house an XML-encoded electronic text archive of about 100 Middle High German texts and their glossaries. The texts will be interlinked with their glossary entries and those glossary entries will interlink with the existing Middle High German dictionaries. This interlinking system will provide numerous discovery and retrieval strategies to maximize the reading and etymological contexts of each work. This project, a collaboration between XML specialists at the Electronic Text Center of the University of Virginia and linguistic experts in the German Department at the University of Trier, is funded by the National Science Foundation (NSF) and the Deutsche Forschungsgemeinschaft (DFG).

The starting point of this collaboration is the compound of Middle High German dictionaries consisting of the following components:

- G.F. Benecke/ W. Müller/ F. Zarncke, Mittelhochdeutsches Wörterbuch (1854-1861)
- M. Lexer, Mittelhochdeutsches Handwörterbuch (1872-1878); and its Nachträge (= supplements) (1878)
- K. Gärtner/ Ch. Gerhardt/ J. Jaehrling/ R. Plate/ W. Röll/ E.Timm, Findebuch zum mittelhochdeutschen Wortschatz (1992)

All three dictionaries are closely related to one another: Lexer's dictionary is an alphabetical index and supplement to the older dictionary by Benecke, Müller and Zarncke and refers to it by explicit references. The Findebuch represents a compilation of glossary references to texts that have appeared since 1878 when

Lexer's Handwörterbuch had been completed. The Findebuch contains headwords and siglas that point to full glossary entries. These three dictionaries have been freely available on the Web since 1999 as an interlinked dictionary compound (**http://www.mwv.uni-trier.de**) and is also available from S. Hirzel (Stuttgart) on a CD-ROM which features more sophisticated retrieval facilities than the internet version.

For Middle High German Interlinked all editions upon which the Findebuch is based have been digitised and will form one of the largest text archives in and on Middle High German on the Web. While the texts of the archive can be read and studied as individual objects, the multidirectional interlinking to glossaries and dictionaries enriches the lexicographical contexts of these works and augments the researcher's own work in linguistic history and analysis of the Middle High German period. The project provides a model realization of the multiple opportunities of linking and electronically merging information gathered from large text corpora and dictionaries. In this presentation at the 2003 ACH/ALLC, the authors will present problems that such a large undertaking has posed, discuss the solutions to these problems, and finally demonstrate a version of the system on the Web to enhance future work in this field.

TOOLS FOR LEXICOGRAPHY, RETRIEVAL, MIDDLE HIGH GERMAN

Frank Queens, Dipl.-Inf., Ute Recker-Hamm, M.A.

STARTING POSITION

A new Middle High German Dictionary is being worked on by two teams of lexicographers: one team in Göttingen (funded by the Academy of Science in Göttingen) and a second team Trier (funded by the Academy of Science and Literature in Mainz). The aim of the project is to develop a new historical citational dictionary of the German language for the period from 1050 to 1350, consisting primarily of a printed version of four volumes, approximately 1000 pages each. Moreover, an electronic version is planned including comprehensive tools for accessing and searching the dictionary and its sources. The new dictionary is based on a 'core corpus' of 150 source texts, an open 'extended corpus' of about 500 sources (so that the composition of the corpus is well-balanced with regard to text type, date and location) and a so-called dictionary corpus which contains a set of dictionaries offering citations that are not covered by the core or extended corpus. The texts belonging to the core corpus have been digitized and lemmatized, and an electronic archive of 1,2 Mio. citations that are electronically linked to their full texts of the sources was built. The extended corpus exists either as non-lemmatized electronic full texts or as printed texts. The Digital Middle High German Text Archive, the subject of the preceding paper, is an essential part of the extended corpus. Being one of the most recent projects in the field of historical citation lexicography in Germany, it was possible from the very beginning to base all work for the new Middle High German Dictionary on electronic data processing. Needless to say, this is of great advantage for the making of a dictionary. But what kind of working environment is in fact needed at a lexicographer's desk? Which tools meet the requirements of the two teams working together on the same material in places that are far apart? Which essential features serve a print edition as well as an electronic edition without taking the lexicographer's attention off his proper work by the necessity of technical encoding? How to ensure the longevity of data and data structures in view of long-term dictionary projects? To answer these questions, the Trier team of the new Middle High German Dictionary, in connection with the Competence Center for Electronic Retrieval and Publishing Techniques in the Humanities, applied for a research grant from the Deutsche Forschungsgemeinschaft (DFG). The project Internet-based Working Environment for the Production and Publication of Dictionaries at Distributed Places was granted, and work started in March 2002. The first goal of the project is to set up a working environment for the two teams, working in different places. The second goal of the project is to adapt the system to the reqirements of other dictionary projects. In this paper the authors present the technical concept and the realization of the system in its present state, and give a brief perspective for the future work.

TECHNICAL CONCEPT

The concept intends to set up a client-server architecture: In its center is an Internet compatible relational database system where all dictionary data is stored (register of headwords, source texts, bibliography to the source texts, dictionary entries etc.). The working environment itself is installed on the lexicographer's computer as client software which contains all the needed tools and features and also manages the data exchange with the database via the Internet. By use of this central database, all the dictionary data is available at the places where the two teams of lexicographers work simultaneously. An XML export device forms the interface to different output devices as typesetting and electronic publication on the WWW. In addition, the XML capability ensures the longevity of the data processed in the project.

REALIZATION

The lexicographical desktop environment consists essentially of four components: the citation corpus, the bibliography to the source texts, the storage and management of entries, and the entry editor. The citation

corpus offers the opportunity to search the database for quotations to illustrate the meaning fo a certain lemma. Of course, truncated lemma input is allowed as well as search restriction on a user-defined source text selection. The result is given as KWIC concordance that can be arranged by various criteria such as source text, word form or date of source text. In the concordance the length of the quotation is freely selectable by the lexicographer, and it is also possible to go to a full-text readout of the source from the concordance directly. This feature is useful to review a quotation in its original full context. Furthermore, quotations can be copied from the citation corpus into the core component of the environment, the entry editor.

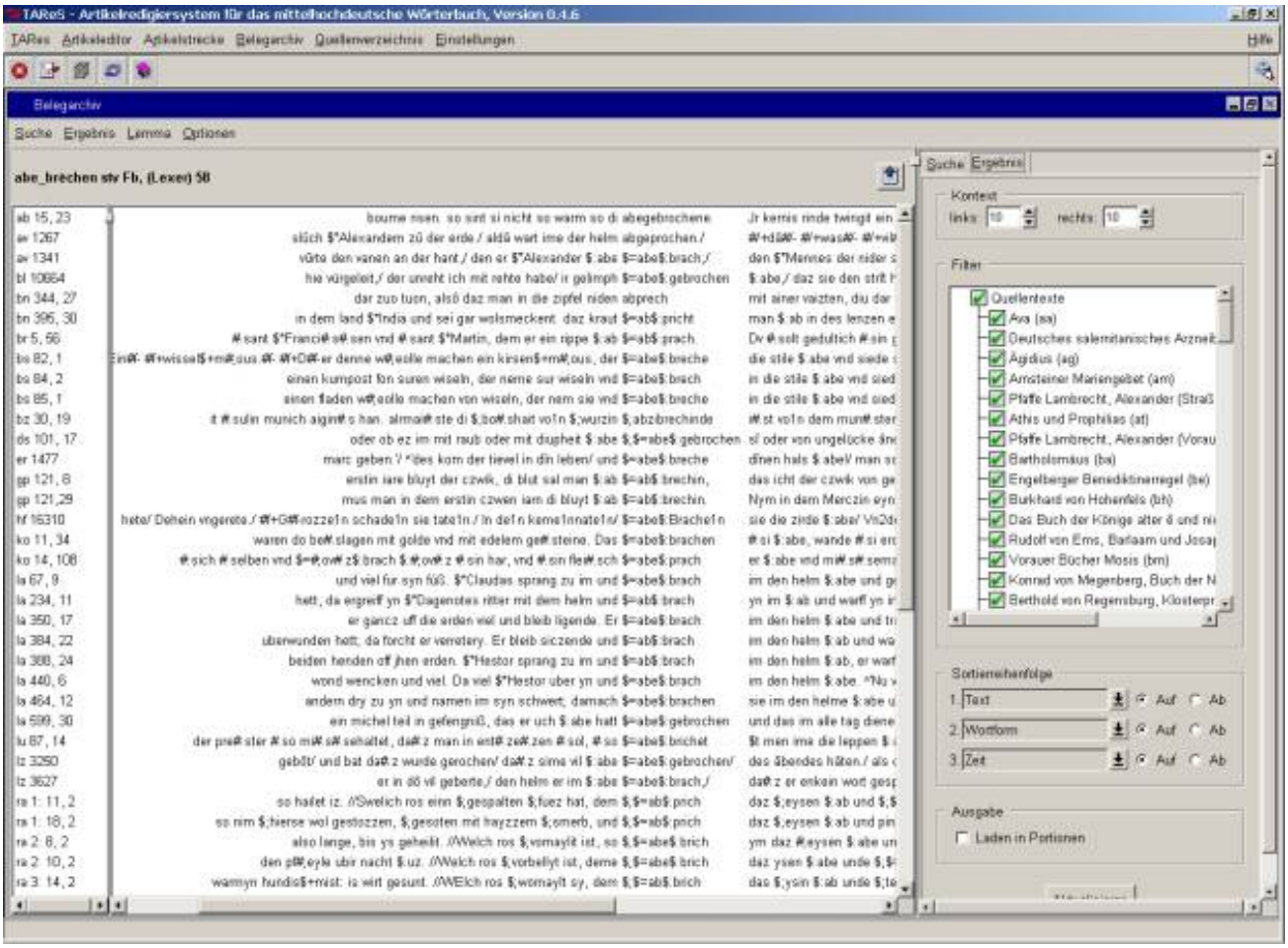


Figure 1: Citation Corpus

The bibliography to the source texts is an informational tool for the lexicographer, which helps to maintain and retrieve master data about the quoted edition, text type, date and localization of the source texts, and information about transmission and earlier editions. Furthermore the whole management of siglas is done in this component of the system.

The third component is used for managing and storing the entries. It consists of an alphabetical list of lemmata/headwords and the corresponding entries. The lexicographer can check the status of an entry: whether an entry to a headword exists already, whether it is ready for publication or not, or whether a headword serves only as a link to another entry. As soon as a certain sequel of entries is ready for publication in an instalment, it is possible to typeset the whole instalment or a number of selected entries by a TUSTEP typesetting routine. The result of this is a postscript-file which represents the printed page of the dictionary in two columns. Since the lexicographers work on XML-tagged data, it is important for them to have a typesetting routine at hand to ensure that an entry meets the measures of entry length required by the space available in a printed dictionary.

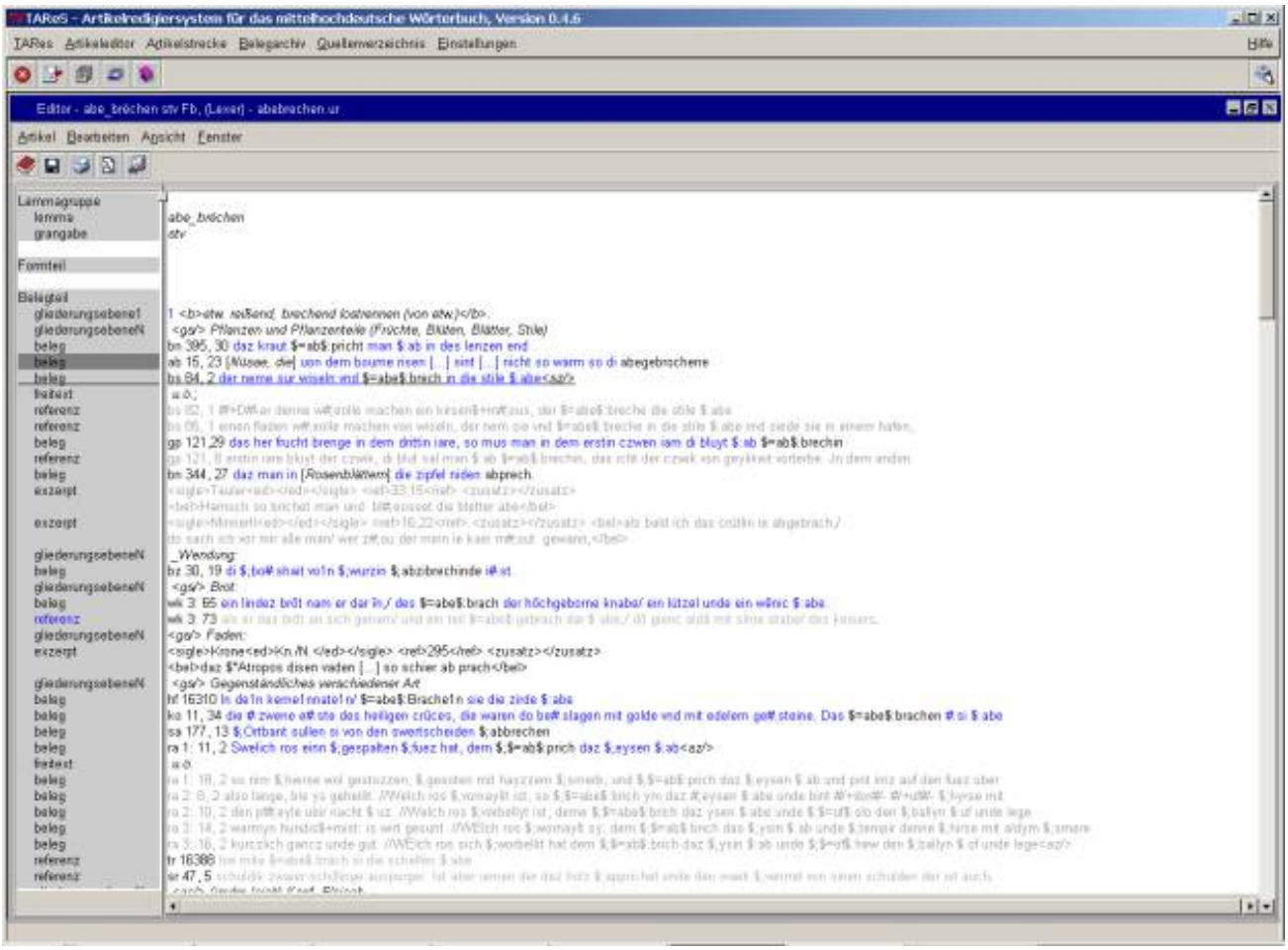


Figure 2: Entry Editor

The main component of the lexicographical environment is the entry editor. This tool allows the manipulation and the commenting of quotations, their grouping according to definitions, and their chronological ordering etc. The wording of a quotation is always protected against irregular handling; therefore it is not necessary to re-check the quotations before an installment is published. However, it is possible to resize the context of a quotation or go back again to a full-text readout of the source text. The entry editor is tag-based which means that all required information is marked up using XML-elements. The entry editor inserts most of these tags automatically, and they are invisible to the lexicographer thus helping him to maintain a clear overview of the structure of an entry. Certain elements can be inserted manually by the lexicographer either via keyboard or via macros. The XML-tagged entries are the basis for different output formats, for the printed and the electronic publication.

PERSPECTIVES

During the last months, the work of the project has concentrated more on the refinement of the entry editor than on the routines of publication. The teams of the new Middle High German dictionary are working with the system in their different places. It has been tested and improved continually, but it is still under construction. Experience shows that more functions are needed, for example a function that ensures a consistent cross-reference system and handling of links, or a workflow component that keeps track of an entry from its first draft to its final publication. The typesetting routines work sufficiently effective and will certainly convince the committees of the Academies who expect convincing results on the printed page. However, future work will also focus on electronic publishing on the Internet and, of course, adapt the system for the requirements of other dictionary projects.

ELECTRONIC DICTIONARIES AND METALEXICOGRAPHY: THE DIGITAL VERSION OF THE DEUTSCHE WÖRTERBUCH BY JACOB AND WILHELM GRIMM AS A BASIS FOR METALEXICOGRAPHICAL RESEARCH.

Thomas Schares M.A.

The electronic version of the Deutsche Wörterbuch (=DWB) by Jacob and Wilhelm Grimm (33 vols., 1854-1971), the largest dictionary of the German language, has been on the Internet for over a year; an offline version will also be available soon.

In recent years, electronic dictionaries have become a familiar tool for scholarly work. Access to dictionaries is much easier and very comfortable once the huge quantities of lexical information which was hitherto stored in weighty multi-volume works has been transformed into immaterial data streams within a computer. So far, the electronic version has proved to be ideal for users of such reference works. But the electronic form can do much more: it allows us to pursue forms of metalexicographical research into the structure and the history of the DWB which has not been possible before. The DWB, like the OED, is based on scholarly principles and reflects the history of philology and linguistics over a long period. Thus it forms an ideal object for the study of the changing methods and predilections in the field of lexicography and lexicographical research.

Not very long ago, the scholar often had to apply dubious methods when studying the way a dictionary is compiled. For example, to inquire into the frequency of quotations of a given author, only small “representative” portions of the dictionary were usually searched. The results were then used to calculate the total of occurrences (e. g. SCHLAEFER 1999 and SCHULZ 1999 use both one entry as a starting point). Factors like the inconsistency or the historicity of the dictionary were often left aside. Similarly, before the electronic version of the DWB was available, the exact number of entries was unknown. In a book published in 2001 the author estimates the number at about 400,000 to 500,000 entries (HAß-ZUMKEHR 2001). Now that we have this dictionary in electronic form, we know that it comprises exactly 249.545 main entries, adding the subentries we have a total of 320.505, this number showing how far previous estimations have been off the mark.

In this paper, I would like to show how we can explore more reliable ways of metalexicographical research on the basis of the electronic DWB and its TEI compliant structural markup. It is already a commonplace that it takes no more than a few seconds to find out how often Shakespeare is quoted in the OED, or how often Goethe appears in the DWB. But electronic dictionaries do not only lend themselves to such comparably simple full-text searches. The underlying text-databases contain the dictionary's content enriched with markup that represents the structure of the dictionary. The SGML/XML-encoded structure permits complex and sophisticated searches as are necessary for studies in metalexicography.

The structural base unit of a dictionary is the entry: one could call a dictionary a (usually) alphabetical collection of entries or articles. All information to a certain headword is collected in the related entry, and organized and presented in a certain way. Metalexicography is interested in finding out how this information is organized and presented, that is how it is structured. In order to gather information about the macro- and micro-structure of the DWB, I chose a new approach, sorting all entries according to their length. Each entry was given a fixed value consisting of its number of lines. These values were retrieved automatically from the SGML-tagged data of the DWB by using TUSTEP scripts.

Table

length	A	B	D	E	F	G	H	I	J	K	L
year(s) of publication	1852-53	1853-55	1855-60	1859-61	1861-72	1872-1935	1868-76	1876-77	1877	1864-73	1877-81
1-5	71,2	78,7	72,4	72,1	73,1	52,5	71,5	66,5	75,2	73,1	69,6
6-10	13,4	9,7	12,3	12,5	13,5	15,8	14,7	13,3	11,4	12,8	15,1
11-51	12,7	8,8	11,7	11,7	10,2	20,8	10,5	14,0	9,7	10,1	11,7
51-29999	2,7	2,8	3,6	3,7	3,2	10,9	3,3	6,2	3,7	4,0	3,6
length	**M**	**N**	**O**	**P**	**R**	**S**	**T**	**U**	**V**	**W**	**Z**
year(s) of publication	1881-1885	1881-1885	1885-86	1886-89	1886-92	1892-1942	1890-1952	1913-1936	1886-1951	1901-1960	1913-1954
1-5	67,8	68,3	74,2	74,2	73,1	62,8	55,3	17,8	48,0	43,8	33,9
6-10	15,5	15,8	13,0	12,6	13,7	16,1	15,4	16,4	18,0	17,0	15,3
11-51	12,6	12,8	10,6	10,8	9,6	15,7	21,7	36,1	21,1	24,7	30,6
51-29999	4,1	3,1	2,2	2,4	3,6	5,4	7,6	29,7	12,9	14,5	20,2

Table 1: Entry length, percentages

The table gives an exact account on entry length. One can see that an average of 63,4 per cent of the entries of the dictionary are shorter than six lines. One can also perceive that the percentages for some letters deviate considerably from this average. Older parts of the dictionary have higher percentages than younger parts. This suggests that the average entry is longer in younger parts of the DWB, which supports the assumption that the reinforced collection of slips after the turn of the century affected the structure of entries: more slips led to longer entries.

Entry length can also form the basis for a structural analysis of dictionary entries. Dictionaries like the DWB are famous for their very large entries like GEIST, GEWALT etc., and entry analysis in metalexicography usually focuses on larger entries with complex structures and rich semantics (e. g. SCHMIDT 1986). However, given the fact that almost two thirds of the entries of the DWB are shorter than six lines, I consider it essential to take a closer look at these small entries. By doing this, general statements about entry structures can be made, but at the same time peculiarities in the work of individual lexicographers become evident. In the parts prepared by Jacob Grimm, for instance, one specific type of entry appears which consists of a lemma, a grammatical designation, a definition and a Dutch equivalent:

```
BLONDHEIT, f. color flavus, nnl. blondheit. (DWB 2, 143)
```

In Jacob Grimm's parts of the DWB (letters A-C and E-FRUCHT), this type of one line entry occurs 137 times, in all the other parts only five times. Another look at the electronic DWB shows that a total of 4300 Dutch equivalents and examples are given. 3400 of these can be found—again—in the parts written by Jacob Grimm. In contrast, in the D-section, which was prepared by Jacob's brother Wilhelm, we find less than ten Dutch examples. Jacob Grimm is the only lexicographer of the DWB who has given such a large amount of Dutch equivalents. Peculiarities of lexicographers can be traced down with great exactitude; thus some lexicographers extensively quote material from German dialects, others are are not interested in dialects at all. These features of the DWB reflect the heterogeneous methods and attitudes of lexicographers in their time. They can only be revealed by using dictionary data with structural markup that proves to be a reliable basis for metalexicographical research.

REFERENCES

Haß-Zumkehr, Ulrike: *Deutsche Wörterbücher im Brennpunkt von Sprach- und Kulturgeschichte*. Berlin/New York 2001.

Schlaefer, Michael: *Zur Darstellung wortgeschichtlicher Zusammenhänge des 17.–20. Jahrhunderts in historischen Wörterbüchern*. In: Sprachwissenschaft 24.2 (1999), 195–220.

Schmidt, Hartmut: *Wörterbuchprobleme. Untersuchungen zu konzeptionellen Fragen der historischen Lexikographie*, Tübingen 1986.

Schulz, Matthias: *Der lexikographische Informationsgehalt in älteren Bedeutungswörterbüchern: Zugleich Überlegungen zum Nutzen einer Retrodigitalisierung älterer Wörterbücher*. In: Sprachwissenschaft 24.1 (1999), 47–73.

Confronting the Challenges in Collaborative Editing Projects: The *Dickinson Electronic Archives* File Management System

LARA VETTER
Maryland Institute for Technology in the Humanities, University of Maryland, College Park
LV26@umail.umd.edu

JAROM MCDONALD
Maryland Institute for Technology in the Humanities, University of Maryland, College Park
jmcdon@glue.umd.edu

This year, the *Dickinson Electronic Archives* (DEA) began work on an edition of *Emily Dickinson's Correspondences* (EDC), a multi-volume TEI-conformant XML scholarly edition of individual collections of letters and poems sent to and from Dickinson to her various correspondents that will include manuscript images, transcriptions, and critical annotation. Overseen by general editors Martha Nell Smith, Ellen Louise Hart, Lara Vetter, and Marta Werner, and coordinated by Vetter, EDC volumes are guest-edited correspondence-by-correspondence by Dickinson scholars across the country. The project is ambitious in scope; Dickinson sent over 1800 letters and poems to about 100 friends, family, and acquaintances, and conceivably many individual guest editors could be working simultaneously on various sets of documents.

The DEA faces several challenges in managing the concurrent editing of the voluminous correspondence, not the least of which is geographical. All four general editors are dispersed geographically: Smith in Maryland, Hart in California, Vetter in Missouri, and Werner in New York. The DEA project manager, Jarom McDonald, and encoding staff are located in Maryland at the Maryland Institute for Technology in the Humanities, while guest editors live and work all over the country. The documents are located primarily in two different libraries in Massachusetts with scattered manuscripts elsewhere, and the encoded data resides on a server at Institute for Advanced Technology in the Humanities in Virginia. Clearly, we need a system by which the geographically remote project coordinator can track the status of multiple documents in multiple correspondences at various stages of editing and encoding, and work with a staff and editors at a distance. We also face the utter impracticality of training innumerable guest editors in advanced TEI and XML, particularly given the encoding difficulties posed by the experimental nature of Dickinson's page, yet we need to maintain consistency in encoding and editing across sets of correspondence, all the while preserving the integrity of the individual editor's work.

Originally conceived by Lara Vetter and implemented by Jarom McDonald, the *Dickinson Electronic Archives* Manuscript Project File Management System is designed to allow editors and staff working from different locations across the country to collaborate on the production of XML files for web delivery. The system will accommodate multiple editors and staff working on different subsets of documents.

The DEA Manuscript Project File Management System facilitates the following processes:

1. For each manuscript, the editor completes a php-driven submission form collecting relevant information about that manuscript. When the form is submitted, it writes an XML file that parses against the DEA DTD, pulling additional data from SQL regularization databases, and deposits the file into the directory for newly submitted documents; whatever data cannot be automatically tagged is placed in a comment tag to provide instruction to the encoder. When the form is deposited into the directory, the editor receives a copy via email, and the project coordinator is notified via email that new manuscript data has been submitted. Additionally, an entry in the SQL-driven management submission database is generated that contains the filename, the editor's name, and the date the file was submitted. Finally, a <revisionDesc> statement is generated and placed automatically into the XML file, containing information about the nature of the action, the person who performed it, and the date.
2. The project coordinator logs into the administrative section of the system and is presented with a complete list of all submitted documents pending assignment to encoders. The coordinator can assign each document individually to any of the current DEA encoders via a drop-down menu; doing so will automatically generate and send an email to that encoder, notifying him/her of the new assignment. The corresponding entry in the submission log will be updated to reflect the encoder assigned to the project.
3. The encoder logs into the "Document Check-in/Check-out" section of the system and is

presented with a complete list of all documents pending to be downloaded and those which have previously been checked-out for encoding but not yet uploaded. The encoder clicks on the file to be downloaded and is taken to the download page, where he/she can save the file locally for encoding. This step updates the submission log entry, noting the date that the file was downloaded for encoding, and moves the particular file from the "pending check-out" section to the "pending check-in" section for the given encoder. This step also automatically updates the <resp> attribute in the header that signifies the encoder. The encoder's job is to perform post-processing on the XML file by utilizing the editor's notes, comparing the XML file to an electronic image of the manuscript, and tagging everything that cannot be automatically generated by the form.

4. When the encoder has finished with the file locally, he/she can log back into the system and click on the entry to "check-in." The encoder is then taken to the upload page and deposits the newly encoded file in a directory separate from that which the editor-submitted file is located. Checking a file in through the system will update the submission log to reflect the date of encoder uploading, as well as generating another <revisionDesc> entry into the document itself. The project coordinator/proofreader is also notified via automatic email that the encoded file has been uploaded.
5. Upon logging back into the system, the project coordinator is presented with a list of files that need to be proofread, in addition to seeing which files need assignment to encoders (see item 2 above). These files can be checked-out and checked-in for proofreading in the same manner as described above for encoders; each action generates the relevant entry in the submission log. Downloading modifies the <resp> element that signifies the proofreader's name, and when the final proofed file has been uploaded, it is deposited in a third directory; hence, all archival copies of the document are preserved. Once the proofreader has uploaded the file, a third <revisionDesc> entry is generated and written into the file and the coordinator is notified via e-mail.
6. Finally, when the proofreader is done, an automatically generated e-mail is sent to the editor(s) with a URL for a transformed version of the file. The editor then visits the URL indicated to perform a final proofing of the document; the HTML page gives the editor the option to accept or modify the file, and the coordinator is notified of the status of the document. If the file is accepted, it is copied to the public directory and becomes available online; if it is modified, the editor's modifications (again, submitted via a php-form) are routed back to the proofreader who can make the final emendations and subsequently upload the document into the public directory.

Additionally, the project coordinator has access to a report feature, which returns the entire SQL database of the submission log. This allows her to monitor the status of all ongoing projects and documents and to track what documents individual editors, encoders, and proofreaders are currently working on.

Though there will, undoubtedly, be issues with certain edited documents that cannot be addressed through the type of automation described above, the document management system will do much to streamline the editing and encoding process and increase the quantity and quality of work done by the dispersed staff. Decisions about what to tag and how to tag complex features of Dickinson's manuscripts are made by the general editors and enforced by the editorial submission form, so documents are encoded consistently across the various sets of correspondence. Editors can focus on editing, without having to learn advanced XML and TEI (although the system, as designed, will accept XML tags as part of submitted editorial data for those editors who are familiar with the TEI and the Dickinson project DTD). Encoders can focus on encoding, without being called upon to make difficult editorial choices. Ultimately, the entire process will facilitate integrity in editing, quality control, and institutional memory.

Texts into Databases: The Evolving Field of New-Style Prosopography

JOHN BRADLEY
King's College London
john.bradley@kcl.ac.uk

HAROLD SHORT
King's College London
harold.short@kcl.ac.uk

The Centre for Computing in the Humanities at King's College London is involved in three broadly Prosopographical projects: the *Prosopography of the Byzantine Empire* (PBE) (recently renamed *Prosopography of the Byzantine World*—PBW), the *Prosopography of Anglo-Saxon England* (PASE), and the *Clergy of the Church of England Database* (CCE). (All are funded by the UK's Arts and Humanities Research Board.)

The goals of these three projects at King's are ambitious. PBE's goal is "to record in a computerised relational database all surviving information about every individual mentioned in Byzantine sources during the period from 641 to 1261, and every individual mentioned in non-Byzantine sources during the same period who is 'relevant' (on a generous interpretation) to Byzantine affairs." (from website, see references). PASE's aim is "to provide a comprehensive biographical register of recorded inhabitants of Anglo-Saxon England (c. 450-1066)." (from website). CCE intends to create a "database of clergymen of the Church of England between 1540 and 1835." (from website).

The sources of information for all three projects are surviving manuscript records of many kinds. The central computing tool that all three projects employ is the relational database. Obviously, there is a significant issue involved in taking the textual source material, often presented discursively, and presenting it in the structured form that a database requires. Central, then, to the design of the database, and to the continuing process of putting data into it, is ensuring that the scholarly interpretation essential to this transformation is properly accommodated.

In traditional Prosopography (see, for example, the well-known *Prosopography of the Late Roman Empire*, or the more recent *Prosopographie der mittelbyzantinischen Zeit*—PmbZ(Lille et al 1998-2002)), the central organising principle is the person. The information about the person is formed into an article by the scholar, and the articles themselves are organised by the person's name. There is often some degree of more structured information attached (in PmbZ, for example, there is, among other lists, a formal list of sources in which information about the person could be gleaned). However, the primary source of information for the user is the article. The article is presented as a narrative in which we find a complex blending of quotation from sources and scholarly interpretation. Some of the assertions are made without further justification, in some cases an argument is provided to support why an assertion has been made, in some cases an issue is left unresolved and only alternatives are outlined. The scholar's task is to take the evidence provided by the sources s/he has read and to represent in the article the shades of certainty about any of them.

All three of our Prosopographical projects take a radically different approach.[1] The final publication will not be a set of volumes containing articles, but an online database. Furthermore, all three projects agree that there will be none or very few articles about persons in their database, and they will be written *after* the data collection process is complete, rather than being central to it. Instead, the evidence data will be recorded as a series of *factoids*—assertions made by the project team that a source "S" at location "L" states something ("F") about person "P". *Factoid* was first applied to this kind of information by Dion Smythe and Gorden Gallacher, and is not a statement of fact about a person; it is an assertion that a source says something about him/her. In Figure 1 you can see some sample output from the PBE database, showing factoids derived from the source *Skylitzes Continuatus*; for Emperor Alexios I Komnenos (identified in the DB as Alexios 001) in 1078. As the illustration suggests, there are several different kinds of factoids provided. In PBE, factoid data is collected for things such as activities or events in which the person took part; physical, spiritual or physiological descriptions applied to them; dignities or offices they held; ethnic group to which they belonged; kinship relationships with other people; locations with which they were associated; occupations they took up; possessions they owned; and religion they professed.

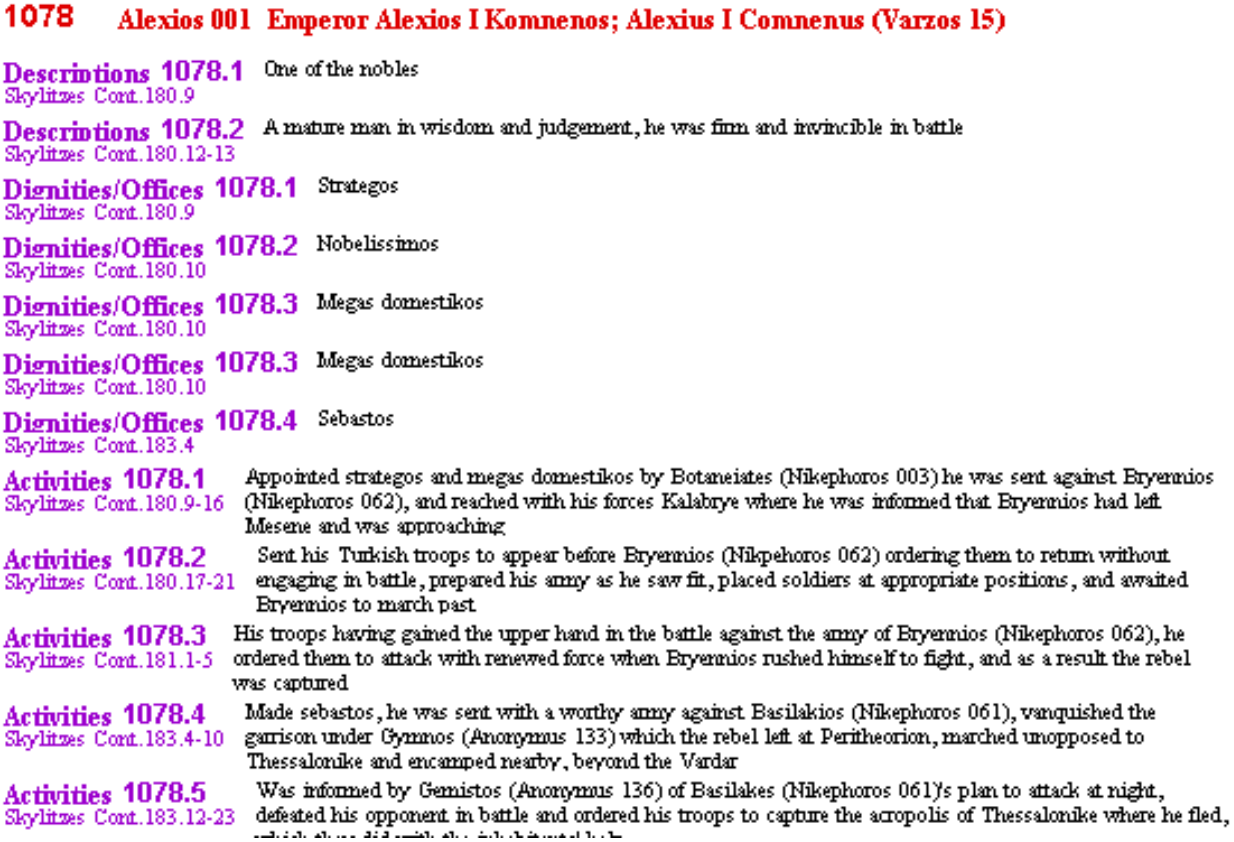


Figure 1: A Factoid List for Alexios I Komnenos (PBE)

Factoids are modeled as entities in the prosopographical database, and each factoid type contains both an explicit and implicit structure. The explicit structure can be relatively complex. Figure 2 shows the data capture screen displaying one of PASE's event factoids—the event being the tonsoring of Guthlac by Aelfthryth (as recorded in the *Vita Sancti Guthlaci*). Not only is there a description field that contains information about the act itself (here only the beginning of the full text recorded in the field is visible), but the event is:

- categorized in the Term field,
- attached to a place (the place described using the word found in the original text, the type of place it is and its modern day location), and
- linked to the two people involved, one identified as the recipient and the other as the agent.

Furthermore,

- the textual source for the event, and location in the text where the act is recorded is entered in the database, and
- there is space for recording a scholarly date recording when the event is thought to have occurred, and space to record whatever dating information is given in the source (only partially visible in this figure).

Finally, there is a place in the “Notes” and “Problems” field where the researcher can record in free text some commentary on the factoid that s/he considers important but does not fit the structured fields associated with the factoid itself.

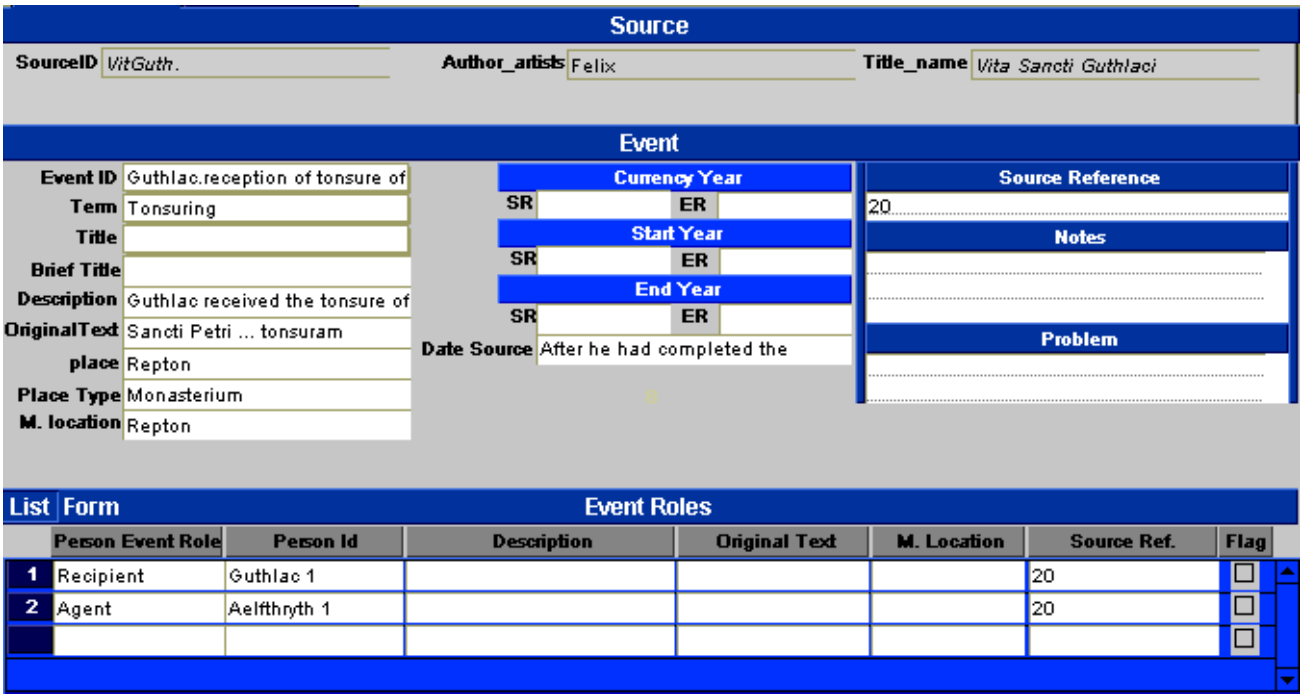


Figure 2: An Event Factoid in PASE

There are, in addition, elements of implicit structure that must be recorded in the textual elements—a reference to another person in the database in the description of an event would be an example of this. Textual fields in our projects are, therefore, often structured as mini-XML documents, with XML being used to handle the structure that they contain and making it available for further machine handling.

A relational database is most useful when the data it contains is highly structured. The capability of the relational model is often underestimated in scholarly circles, and both Greenstein (Greenstein 1994), and Townsend et al (in the *AHDS Guides to Good Practice: Digitising History*) begin their discussion of databases by acknowledging that a single table in the relational database as often too limited for large scale historical use (it is described as the "matrix straitjacket" in Townsend). Greenstein, however, recognises that the *relational* aspect of the relational model—which allows material from more than one table to be linked into a single logical entity —allows for richer collections of information to be formed. Once an entry (say, for the Person) in a table is changed from being a text string containing the person's name to a link to a row in another "person information" table it is possible to record a great deal of richer information about that person. Thus, PBE, PASE and CCE databases contain not only structured data in the form of factoids, but they also contain complex structures spread over more than one table each that represent other important "objects" in the database related to the factoids such as persons, geographic locations and possessions.

The challenges associated with our Prosopographical databases are many. First, there is a constant struggle to be sure that, for each field one enters, one is clear about to what extent the field represents simply what is in the source, and to what extent it is actually represents a scholarly interpretation of that source. Even the "original source" fields holds text out of context, making it a matter of scholarly decision about exactly what fragment should be included.[2]

Furthermore, for all three projects data is collected on a source-by-source basis by more than one researcher. The issue of consistency between researchers is constantly on our minds. Consistency issues are dealt with during the editorial phase of the project, where editors will use tools to assert, for example, that a person A in one source is the same person as person B in another. We expect to have more to say about these issues during our presentation at the conference.

An article in a traditional prosopography provides a well organised bundle of information to its user in the form of a narrative. What happens when the prosopography contains large collections of factoids instead? As figure I suggests, the factoid model used in these projects provides a way that the machine can generate "micro-narratives"—to use the term presented in (Ramsay 2002). These narratives will be richly linked to other data in the system, and different micro-narratives will be generated when one enters from different starting points (say, from a location, rather than from a person), or by traversing links that connect materials in each factoid to the broader database contents. Clearly, a web access mechanism for these databases will need to be significantly more complex than a simple search form which results in a list of selected items. We believe that the best interface to this will provide the blending of a searching and browsing paradigm that is now characteristic of large websites on the WWW. All three projects are now reaching the stages where some detailed exploration of these presentation and selection issues can begin. We will describe

some of these issues in more detail during the conference presentation.

In moving from article-based to factoid-based prosopography, all three of our projects are participating in the development of a radically new approach to the field. The role of the scholar has clearly altered in our projects. In addition to being ultimately responsible for the projects' "content", all project members must work together to ensure that new methodologies are developed which ensure both quality and consistency of information across the large number of individual factoids. The role of the scholarly end-user may also have to change, and here too the project has to work to ensure that the experience of the end-user searching through these factoids is positive and useful. In the end, we are finding new ways that technology can serve these essential scholarly goals.

NOTES

[1.] Those familiar with the CD publication of PBE I may have observed that the approach taken there was actually a transitional one—somewhere between the article-oriented and factoid-oriented approach described here. The current phase of PBE work has fully taken on the approach described below.

[2.] Linking of our databases to full-text representations of the source texts have been considered by all projects. In all three projects, however, this has been rejected for various reasons. There was no funding to take on the preparation of a fresh electronic edition, and, in general, no scholarly reputable electronic editions of the texts were available elsewhere. PASE might yet explore some options in this area, however, for recording data from charters.

REFERENCES

Bradley, John and Short, Harold. (2002) "Using Formal Structures to Create Complex Relationships: The Prosopography of the Byzantine Empire—A Case Study" in. Keats-Rohan K.S.B (ed.), *Only Connect: The Use of Computers in Developing Prosopographical Methodology*, Oxford. Unit for Prosopographical Research, Linacre College. Preprint online at **http://pigeon.cch.kcl.ac.uk/docs/papers/pbe-leeds/**

Clergy of the Church of England, at **http://www.kcl.ac.uk/humanities/cch/cce**

Greenstein, D. I. (1994). *A Historian's Guide to Computing*, Oxford: Oxford University Press. pp. 268

Lille, Ralph-Johannes et al. (1998-2002), *Prosopographie der mittelbyzantinischen Zeit, Abteilung I: 641-867*, Prolegomena and volumes I-VI. Berlin: Walter de Gruyter.

Prosopography of Anglo-Saxon England, at **http://www.kcl.ac.uk/humanities/cch/pase**

Prosopography of the Byzantine Empire, at **http://www.kcl.ac.uk/cch/PBE**

Ramsay, Stephen (2002). "Relational Ontologies and the New Historicism" in session Pitty Daniel et al. *Multiple Architectures and Multiple Media; The Salem Witch Trials and Boston's Back Bay Fens Projects*, at ALLC/ACH conference: July 2002. Abstract online at **http://www.uni-tuebingen.de/cgi-bin/abs/abs?propid=55**.

Smythe, Dion C. (2000). "Prosopography of the Byzantine Empire", in Deegan, M and Short, H. (Eds), *DRH99: Selected papers from Digital Resources for the Humanities 1999*, London: Office for Humanities Communication, pp. 75–81

Townsend, Sean et al. (1999) *AHDS Guides to Good Practice: Digitising History*. Oxford: Oxbow Books. Online at **http://hds.essex.ac.uk/g2gp/digitising_history/**.

The Suda On Line: Applying Computer Technology to Ancient and Byzantine Studies

ROSS SCAIFE
U of Kentucky
scaife@uky.edu

RAPHAEL FINKEL
U of Kentucky
raphael@cs.uky.edu

Using highly interdisciplinary methods we have built a collaborative infrastructure for translation and annotation of ancient texts. This generalizable infrastructure is now fully deployed in the Suda On Line

http://www.stoa.org/sol/). The Suda is a 10th century Byzantine Greek lexicon of some 30,000 lemmata. After four years of continuous development we have implemented a complex yet effective and practical system. Our goal is not only to provide the SOL as a useful tool for researchers, but also to explore and facilitate the modes of scholarship now made possible by open source technology and the internet: this effort is cooperative rather than solitary, communal rather than proprietary, worldwide rather than localized, evolving rather than static. Our international team of managing editors, editors, and translators has now worked up approximately one third of the material in the Suda, quite a satisfactory rate of progress. ACH/ALLC in Glasgow had an initial presentation concerning this project; we feel that substantial further development and our positive results warrant an update at this time.

In order to encourage the participation of translators and editors, and in order to make the SOL database a useful scholarly resource as quickly as possible, we make our materials available to users as soon as it is submitted. We acknowledge that this philosophy raises concerns. One of the major issues with electronic publication of scholarship is the potential it has for circumventing time-tested procedures for quality control. While we do not want simply to add to the sea of uncontrolled material on the Web, at the same time we insist on our right to experiment, and we have no desire to replicate the print-publication paradigm in electronic format. Many of the advantages that electronic publication offers, including immediacy, accessibility and adaptability, are seriously handicapped by traditional editorial processes, where chronic bottlenecks frequently develop in the effort to keep the publishing house's imprimatur off of anything with any detectable shortcomings. In order to exploit these advantages of the web while at the same time maintaining a reasonable level of quality control, submissions to the SOL database undergo the following process of editorial evaluation and monitoring:

1. Initial submissions immediately become accessible to users searching or browsing at the SOL site, but their "draft" status is clearly marked.
2. Once a submission has been carefully vetted by one of the SOL editors for errors and significant omissions, its status as part of the SOL database may rise from draft into one of two categories: low or high. At every stage of this process, the editors who participate in vetting and improving the entry will be prominently identified to the user, along with any descriptive comments they may provide concerning their editorial work.
3. Most importantly, even an entry that has achieved high status will not be considered perfect and immutable. At the discretion of the editors, improvements, changes and additions of links and bibliography can continue indefinitely.

While this way of doing things puts more of the burden of quality control on the end user, our system of marking editorial status gives researchers significant assistance in coming to an informed decision about the reliability of the material in SOL. In fact, our system offers definite advantages over the canonical paradigm of peer review from the consumer's point of view. In print scholarship (and electronic scholarship that merely follows the traditional model) the number, identity, and qualifications of reviewers remain hidden, and one must usually base one's estimate of the reliability of the scholarship solely on the identity of the author and the general reputation of the venue. In the standard paradigm, moreover, the end product is more or less fixed, whereas our database is being improved continuously.

This presentation will describe our project from various perspectives, including the following principal points of discussion. (1) An overview of the Suda itself, including a few examples that illustrate its diverse composition and unique value for several fields of humanistic scholarship, despite its flaws and peculiarities. (2) The multiyear interdisciplinary collaboration among computer scientists, historians, and philologists that has produced our results so far. (3) The academic ideology that guides our production of a freely-available and open-ended e-text, including significant ways in which our editorial practices diverge from more traditional ones. (4) The most important features of the online site available on a hierarchical basis to the participants and the general public. (5) The specific applications and programming technologies that enable those features. (6) Our most recent effort: generation of a unified, complete, and self-documenting XML snapshot of our data. This latest ability addresses our responsibility to ensure the long-term archival security and viability of our results, and it also allows us to experiment with powerful new technologies centered around XSLT programming and the Cocoon environment for the transformation and publication of electronic documents. The presentation will include a demonstration of these experiments and conclude with the prospects for future developments.

Great Expectations, Expectant Implementations—or, What We Expect of Our Electronic Resources and How We Meet Those Expectations

RAY SIEMENS
Malaspina U-C
siemensr@mala.bc.ca

GEOFFREY ROCKWELL
McMaster U
grockwel@mcmaster.ca

PATRICIA CLEMENTS
U Alberta
patricia.clements@ualberta.ca

ANDREW MACTAVISH
McMaster U
andrew.mactavish@mcmaster.ca

MICHAEL BEST
U Victoria
mbest1@uvic.ca

Our session's focus is on what our community expects from electronic scholarly resources and the ways in which we attempt to implement our resources in such a way that they respond to our expectations. The session will begin with individual presentations by Clements, Rockwell, Mactavish, and Best on this topic within the specific context of their ongoing research projects—with the intention of leading to panel and then open discussion and debate.

EXTENDING THE COLLABORATION

Patricia Clements (U Alberta)

In the last decade or so, humanities computing has introduced a new model of research in the humanities. More than that, it has undergone so rapid a development that it has brought us to the brink of major changes in our institutional practices both in research and in teaching. In this panel about expectations, or what the future may hold for humanities computing, this paper will address both research and institutional issues. This interdisciplinary practice has created new challenges in collaboration in both spheres.

From the outset, humanities computing has been an experiment in collaboration. Its introduction of genuinely interdisciplinary conversation into the core of the humanities has had a transforming impact on the ways in which we approach much of our work. Of course, some kinds of collaboration have already been a key methodology in the humanities and social sciences, and the work of the twentieth century was in significant part shaped by great collaborations—resulting in great editions, histories, and historical dictionaries by many hands. By and large these collaborations brought together like-minded scholars working in more or less uniform methodologies, and by and large they functioned as a perceived exception to the preferred norm of the single scholar producing the monograph.

The nature of the collaboration generated by humanities computing is quite different from that. At the core where the new scholarly resources originate lies a conversation between distinct disciplines and very different research methodologies. There can be no Orlando literary history, for instance, without the specific disciplinary expertise of the research collaborator who is a Professor of Computing Science or without the MSc in Computing Science who works at the side of the literary historians to build the deeply encoded Orlando Project, prepare its delivery system, and develop its interface for academics and other users. In this project, the interdisciplinary collaboration reshapes not only the notion of history, and of the reader of history, but also of the ways in which history is made. The Orlando collaboration has involved a training and research

partnership with graduate students which has become a new model for graduate education in the departments in which it is housed.

At this stage in the development of humanities computing, which is its move from the margins of our humanities disciplines to somewhere much closer to the centre, we are faced with the need for new kinds of collaboration. One of these must happen within our research mandate of humanities computing; the other in the practices of the institutions in which we are housed.

As the work of humanities computing is progressively mainstreamed and the need expands for electronic resources whose scholarly quality and authority is undeniable, it will be important for scholarly projects to reinforce one another's effectiveness by developing tools and tactics for convergence. As projects of scholarly quality and reliability emerge, they can expand their utility and resonance in new partnerships. For instance, conversations have been initiated on the convergence of the several literary projects dealing with women's writing in English. Together with the Brown Women Writers Project, the Victorian Women Writers Project, British Women Romantic Poets, the Dickinson Archives, and the Perdita Project, the Orlando Project has discussed the possibility of networking our various but related resources. Great synergy and enhanced utility would result, we believe, from bringing together a collection of independent text collections united by a central architecture that translates among their different knowledge representations. To allow for a single point of access to these established electronic text projects, however, will require a metadata standard that will allow the information they contain to be searched, retrieved, and shared effectively. Only then will a unified means of interrogating them be possible.

Provision of tools for this kind of integrated access to existing projects of reliable scholarly quality will be a major challenge to institutional and international collaboration in the next stage of our development. Now that humanities computing infrastructure is being strengthened and solidified, in Canada particularly with the national and multi-institutional support for TAPoR (the Text Analysis Portal for Research), we need to work towards tools that permit increased convergence and interlinking of currently discrete materials. Numerous fields in addition to women's writing or literature generally—history and law, for instance, in which the separation between source and secondary materials remains strong—stand to benefit from the investigation of possibilities for relating primary and secondary material which are involved in the kind of collaboration I have just sketched. We need to find ways of fostering such collaboration that overcome the various institutional and economic constraints, as well as the varied trajectories, of related projects.

But challenges in collaboration will not be restricted to research nor to the development of new tools to enable effective research and new knowledge production in the humanities. They are also being posed – in spades— in the area of our institutional presence and practices. So successful has the humanities computing project been in the decade of the Web that it is becoming mainstreamed into the teaching agenda of graduate and undergraduate programs. While only a few years ago this experiment in the bringing together of texts and technologies was a small specialization within the broad range of humanities research and a very marginal operation in relation to the teaching mandate of departments, it is now poised to become a curriculum requirement at both the graduate and undergraduate levels. The need for highly qualified personnel in humanities computing will increase enormously in the next decade, and our free-wheeling interdisciplinarity will need to move into the phase of building new institutional structures for its work and of finding funding for the training of a professional cadre to support teaching at both levels of curriculum. As reliable scholarly on-line resources multiply and links between them make them increasingly useful instruments of mainstream teaching and research, the humanities computing experiment will be mainstreamed as an important element in cultural literacy.

TAPOR

Geoffrey Rockwell (McMaster U)

Rockwell will speak on the multi-institutional Canadian TAPoR project, the Text Analysis PoRtal, and its first-phase development, part of which has included an extensive survey of what those in the international humanities computing community expect of their textual analysis resources. (A more detailed precis is forthcoming.)

DIGITAL GAMES RESEARCH, ELECTRONIC RESOURCES, AND THE CRIMINAL SCHOLAR

Andrew Mactavish (McMaster U)

The digital games research community in the humanities and social sciences has been enjoying impressive activity in recent years. Conferences and special sessions on digital gaming have become common across a range of disciplines; publication on digital games is taking off with the launch of the peer-reviewed journal Game Studies in 2001 and with several essay collections and books recently or soon to be published; the Digital Games Research Association (DiGRA) was established in 2002 to provide researchers with a formal network for knowledge exchange; and new courses and academic programmes covering digital games are

running and being developed. As the evidence suggests, academe is recognizing digital games as important works of culture.

While these advances in digital games studies help to legitimize a field of research potentially tainted within the serious halls of academe by the unseriousness of play, digital games researchers face an even more challenging set of problems around access to primary texts. It sometimes seems like a dozen new video games hit the market every week, but the truth is that quantity of titles does not necessarily make for ease of access. Digital games, especially commercially marketed games, are expensive to purchase, sometimes impossible to rent, and seldom available in public or university libraries or archives. The situation is even worse for legacy and abandonware games that are no longer available for purchase on the new or used markets. This problem shows few signs of going away.

All digital games, and many other forms of digital-born artifacts, become legacy texts as new computing technologies supplant older ones. Support for today's games will inevitably be dropped as new platforms, operating systems, form factors, storage media, and physical interfaces are adopted. Obsolescence helps drive the computing industry. But in a quarter century from now, when scholars are conducting historical research on "classic" games from the turn of the millennium, they will want to view works like Grand Theft Auto: Vice City. How will they do this if they cannot find operational consoles and original copies of the game? In 50 years, when there are no working PlayStation 2 consoles and no copies of the original disks, will they have a legal means for viewing the texts?

The humanities computing community, among others, has been working hard for decades on a similar set of problems around preservation of print texts. Its strategy has been to promote standardized mark-up languages to preserve texts in digital form for as long as text is supported by computers. Similar efforts against the tyranny of technological obsolescence are necessary if we hope to archive digital-born works that do not easily fit within the relatively narrow scope of text markup languages like XML.

One of the most promising means for preserving digital games comes from the development of game system emulators. The gaming community has been building, updating, and using emulators for years. One of the best known and most ambitious is MAME (Multiple Arcade Machine Emulator), which was first released in 1997. Like other game system emulators, MAME reads and plays game ROMs originally stored on the component boards and cartridges for arcade and console systems. Thousands of these ROMs are available on the Internet for download.

Building emulators, ripping game ROMs for Internet distribution, and using ROMs to view games are activities that raise several difficult legal issues around copyright. Playing copyrighted game ROMs for which one does not own a license or have permission seems clearly to contravene copyright law. But what is an academic to do if there are no means to acquire legal permission to play game ROMs? Given recent reforms to copyright law in the US and soon in Canada, some digital games research activities straddle the gray line between legal and illegal. Are digital game researchers who use emulators without licenses performing criminal activities in the name of knowledge building?

In an effort to decriminalize digital games research, scholars of digital games and of culture in general need to be developing and promoting strategies for archiving digital game materials to support current and future research in the area. In addition to presenting the problem of digital game archives, this paper will propose the following solutions:

1. Libraries and Archives: If academe is going to include digital games within the purview of its goal to preserve cultural artifacts, then it needs to build collections, provide viewing facilities, and maintain legacy platforms. This section of the paper will also give a quick summary of current archives.
2. Emulation: New forms of emulation need to be developed to support legacy systems and copyright legislation needs to be developed covering use of emulators for games unavailable on the market for academic research. These strategies raise their own set of questions around whether academics have inherent rights to view cultural works if copyright owners have chosen to make the works unavailable.
3. Negotiation with Game Publishers: Academics and university libraries need to negotiate licensing agreements with game developers to legally store emulators and legacy ROMs for digital games research.

These proposed solutions provide only partial answers to the larger problems surrounding preservation of digital works that do not fit well into the model supported by current markup languages. But if the fields of humanities computing and digital media hope to build and support research on multimedia texts, then we need to broaden our strategies or, at the very least, distribute "get-out-of-jail-free" cards to scholars forced to live the research life of a criminal.

A MOST RARE VISION: THE INTERNET SHAKESPEARE EDITIONS

Michael Best (U Victoria)

The Internet Shakespeare Editions (ISE) were created in 1996, with the simple, but ambitious, aim of making scholarly, fully annotated texts of Shakespeare's plays available in a form native to the medium of the Internet. The first tasks were to establish both an academic infrastructure and a design for the site as a whole. The complexity of the process of integrating an advanced academic structure with the new medium was such that I chose to designate myself the "Coordinating" Editor: one whose job was to provide connections between the academic and the technical.

To ensure high quality academic input into the Editions, I created an Advisory Board, with representation from various flavours of Shakespearean editorial traditions, as well as some members whose expertise was in the area of electronic texts. As with any major editorial project, each major work will be edited by individual scholars, or teams of scholars in collaboration. On the technical side, there were two initial considerations: the nature of the tagging of the electronic texts, and the design of the site. I took the risk of creating a special, simplified tagset for the editions, designed to make the process of tagging sufficiently uncomplicated that it could be used by a scholar more versed in the intricacies of Shakespeare's texts than in the then-standard TEI-SGML. At the same time, the site adhered to the general principles concerning electronic texts as established by early work in the field by Faulhaber and Schillingsberg. Thus the tagset was designed in such a way that it could at a later stage be automatically converted to a more standard format, and indeed ISE-tagged texts can now be converted to well-formed XML by a Perl script. The design of the site itself was structured to make the most of the expectations of the Internet, while at the same time sending a clear signal concerning the academic credibility of the refereed materials published. Using the metaphor of a library, the site was divided into a Foyer for introductory and explanatory materials, the Library itself, where only fully peer-reviewed materials would be published, and an Annex, where draft texts and other less formal materials could be published.

The basic academic and design elements are unchanged after seven years, and have stood the test of experience. There have been, however, some significant developments. The ISE now has a General Textual Editor, a distinguished young editor, Eric Rasmussen, who is also involved in the highly demanding, part electronic New Variorum edition of Hamlet. All Shakespeare's plays are now transcribed as initially printed; several plays are represented by more than one text, since they were initially published in one or more Quartos before the appearance of the First Folio (1623). The range of texts has also been expanded to include six plays first published in the Third Folio of 1663, wrongly attributed to Shakespeare. It has also become clear that an important service that can be supplied by a site like the ISE is the publication of reference works that are no longer in print. John Velz's important Shakespeare and the Classical Tradition (originally published in 1968) is now published on the site in a graphic facsimile, thanks to his generosity in making it available. A further significant addition is the major reference work on Shakespeare and film by Kenneth Rothwell. Again, the author has made his work available; in this instance, the text has been rescued from ancient 5 1/4 inch floppy disks, and is in the process of being transformed into database format to permit rapid searches and regular updates.

It is clear that the future direction of the ISE will involve a more general movement towards the use of relational database structures in other areas. The site has recently acquired permission to link images of the complete Folios 1–4 to the transcribed texts; to do this we are developing a database structure based on the transcribed XML texts.

The most important further development of the site will be in the area of performance. One of the much-vaunted capacities of the electronic text is its capacity to link to "texts" of a different kind: graphics, sound, and video. Shakespeare is a perfect vehicle for experimentation in this area, since his plays are filmed and performed with such frequency that a whole discipline of criticism depends on discussing his plays in performance. One of the great challenges in developing a performance database of this kind is copyright. For copyright reasons, it is unlikely that much material from popular films will be accessible; the "workaround" is to turn to the remarkably rich source of staging documents of various kinds created by the inventive and creative work of the many professional companies that perform Shakespeare around North America, especially in open-air forums. A parallel challenge is to create a data model that will reveal patterns within the materials as they are added, and allow for both simple and advanced searches by students, scholars, and actors. This section of the ISE has the potential to become a major research tool in its own right, and will in due course require that a further specialist editor be appointed to oversee the process of ensuring quality in the kinds of materials stored in the database. A dedicated section of the site, the Theater, will lodge the performance database.

The ISE is pleased to be contributing both its texts and images to the developing *Text Analysis Portal for Research*, at present under development by a consortium of six universities in Canada. Together with our

sister institutions, we look forward to enhancing the kinds of readings we can offer visitors to the site. As we work towards integrating multimedia texts with encoded transcriptions and modern, annotated editions, the design of sensitive and inclusive metadata, of the kind under development in TAPoR by the team at the University of New Brunswick will be central, as will the expertise of Ian Lancashire's group at the University of Toronto, as further tools for the online analysis of texts, along the lines of TACT and PatterWeb, are developed.

Ambiguity, Technology, and Scholarly Communication

WENDELL PIEZ
Mulberry Technologies
wapiez@mulberrytech.com

JULIA FLANDERS
Brown University
Julia_Flanders@brown.edu

JOHN LAVAGNINO
Kings College, London
John.Lavagnino@kcl.ac.uk

PART I

Jerome McGann very subtly puts his finger on a point of stress between humanities computing (especially as pursued through text encoding) and traditional literary studies:
[Computer-facilitated] methods, however, cannot concern themselves with aesthetic issues because they forego any engagement with the 'minute particulars' of specific works. More crucially, while these approaches view their materials of study as indeterminate and non-transparent, the critical instruments they deploy are not. Computers and computer programs may be (and often are) extremely 'complex'; nonetheless, their functionality depends upon their determinate and self-transparent structures. [from "Radiant Textuality", **http://www.iath.virginia.edu/public/jjm2f/radiant.html**]

The apparent incongruity between computational approaches and encoded data, on the one hand, and literary meaning on the other, becomes more poignant and more interesting as our digital tools become steadily more powerful and nuanced, and as the community of scholars with access to such tools broadens to include those with traditional "aesthetic" or critical, rather than linguistic, interest in the text. For this community, the problem of ambiguity and its related terms—indeterminacy, multivalence, uncertainty, disagreement—is central not only to their own work, but also to their perception of the new digital tools, which will seem alien and beside the point unless they can accommodate themselves to these qualities. Partly in recognition of this, at ALLC2002 in Tuebingen, Stephen Ramsay suggested a "ludic" approach (building on McGann's "deformative" criticism), in which the computer is not so much turned to the purposes of a literary "panopticon" (if this may serve as a figure for the text encoder's ideal of transparent access to a text, indexed, concordanced, and marked up for any kind of processing or analysis), but is more like an instrument of play, gambling or divination.

Yet the larger question remains. This assumption of incongruity bears reexamination, and not only from the standpoint of digital tools and literary method but within the entire economy of scholarly research and communication. When we ask precisely why—or whether—digital methods cannot accommodate the detailed textual insight on which literary criticism is built, we also raise several larger issues which this session will seek to articulate and address.

First, as Wendell Piez will argue, a careful inspection of the problem of "ambiguity" (and digital technology's presumed incapacity with it) reveals this is a problem that subsists not at a single level, but at every level of the system. Just as we feel there to be a difference or stress between an "ambiguous" literary text and a "disambiguated", cleanly marked up representation thereof, so also we insist there is a difference between how an electronic interface presents traditional textual scholarship from how a critical edition in print does it—and so also we find our work is difficult to evaluate and credit by traditional norms. A close consideration shows these ambiguities and destabilizations not to be a characteristic of electronic work per se,

but rather of poles within scholarly work in general, which is always dedicated both to ambiguities and their resolution—poles whose magnetic tension is being energized by the solvent effect of the new technologies on traditionally stable institutional roles.

Next, Julia Flanders will explore the assumption that text encoding cannot accommodate the kinds of ambiguity that are essential to scholarly textual representation and study. For reasons stemming from the history of scholarly textual study, text encoding is greeted with ambivalence as a tool for representing the subtler aspects of textual meaning. But not only can markup technologies, in principle, describe a much wider and less determinate range of textual phenomena than is presently acknowledged, but in addition they will need to do so in order to respond to and represent the real thinking and work scholars do with texts. This paper will consider how such a model of text encoding might fit within the larger environment of scholarly communication.

Finally, John Lavagnino reflects on the "scholarly economy" and the always-ambiguous efforts, and occasional successes, of scholars in reaching audiences outside their own narrow circles. It turns out that just as we wonder whether there is any audience left at all, it turns out new forms of distribution and access create new kinds of connections across boundaries. This, in turn, prompts one to consider the question of "ambiguity" rather in light of how we make our own language(s) and concerns of interest to readers who do not bring our own presuppositions to the work. Broader audiences have problems with ambiguous language or oblique references, but to counteract leveling tendencies of the electronic medium we can expect scholarly publications to feel impelled towards greater explicitness in some respects anyway.

REFERENCES

William Empson. "Preface to the second edition". *Seven Types of Ambiguity*. London: Hogarth, 1984.

Rob Kling, Lisa Spector, and Geoff McKim. "The Guild Model". *Journal of Electronic Publishing* 8:1, August 2002.

Jerome McGann. "Radiant Textuality". **http://www.iath.virginia.edu/public/jjm2f/radiant.html**

PART II
SCHOLARLY TRANSGRESSIONS

Wendell Piez

One might imagine a number of questions we could pose at the intersection of electronic text, and specifically electronic text that makes sophisticated use of markup technologies, and traditional problems or areas of interest of literary criticism. For this session we have agreed to consider the concept of "ambiguity" in the light of e-text technologies and humanities computing projects (or the humanities computing project in general), and/or vice versa: e-text in the light of approaches to "ambiguity".

Examining a particular text (my paper selects a randomly-found snippet from a pulp horror short story [example 1]) to see how ambiguity might manifest itself in literary language (that is to say, where even the most traditional critic might turn for such an example), it is apparent that ambiguity comes with the territory (as it were) of reading: indeed, reading itself (particularly the reading of narrative) is an engagement with a continuous chain of ambiguities, ambiguities suggested, modulated, and ultimately resolved (or not). This movement, in fact, can be observed at several levels at once in the course of reading, from the lexical level of the senses of words on up through various figurative structures into the narrative itself (and sometimes into wider contexts than that). It is intriguing to note that already, examining a text in this way, we can discern a dimension of textual experience which is the very stuff of literary criticism, but which hitherto, markup systems have not tried to describe. (To my knowledge. In part, this may be due to the predisposition of markup to model a text as a synchronous artifact, whereas the shifts and eddies of senses in a sentence or paragraph occur diachronously, i.e. through time, and subjectively and variably so.) And already we have noticed something we can call "ambiguity": ambiguity is when two or more possibilities are in play, and which of them will hold true depends on factors unresolved, unknown or unknowable. More often than not, these factors are part of the *context* within which the ambiguity occurs.

This is why when we look at something on the opposite extreme—an example of a markup language identifying ambiguities or uncertainties in or regarding its content (and as an example here I have a fragment of a DTD for a biographical encyclopedia in which several forms of dates are given, including various kinds of "unknown" dates [example 2]), we may laugh to suppose that this should be taken to be an example of "ambiguity in markup". It is the opposite: a systematic (and hence, unambiguous, at least if well-designed for its purpose) *representation* of ambiguity. TEI certainty attributes fall into this same category.

Yet when we look at electronic technologies (such as the encyclopedia that provides the context for the "naturalized" or "domesticated" ambiguity just cited, a work accessible in all kinds of ways besides print) nonetheless we see something deeply "ambiguous" about them. Considering a series of scans of print and

electronic scholarly productions (I'll show scans of a literary anthology, a critical edition with commentary of an Ancient Greek text, and of two or three not untypical electronic interfaces [examples 3-7]), they apparently share some striking qualities: all of them have their particular ways of represented and "resolving" ambiguities by drawing attention to them and finding some actual or supposed resolution. (This generally involves some simple protocol to be followed by the reader, such as consulting footnotes or marginalia, or selecting with a pointer). In fact, this seems of the essence of scholarly work (perhaps in contrast to the publishing of theoretical tracts). So why are electronic media so "hot", so scary, so cool or so retrograde? Why is it even an issue whether and how they represent the world (or scholarly research) differently from more established media such as journal articles and bound monographs? What accounts for the "culture gap" between traditional literary studies and "humanities computing"?

E-text technologies, it is apparent, both create new contexts for the reception, examination and study of literary or historic artifacts, and represent missing contexts in new ways (for as it turns out, the representation of missing contexts is much of what scholarship has always had to be concerned about). Yet I don't think this is the real reason why they raise questions with such apparent urgency, since more traditional forms of scholarly work do much the same; at least it does not account for the urgency fully. Rather, if one broadens one's view again to yet a greater context, a deeper reason for both excitement and anxiety appears. Scholars, publishers, marketers, audiences, and librarians have all played quite distinct roles in a highly-developed, elaborate information economy [see attached diagram]. In reality, of course, this economy is incredibly complex and layered—the diagram is only the most schematic representation of it and much could be said to extend and qualify what it can only hint at—but even in this simple view, it is apparent how both stresses and opportunities could arise with the introduction of technologies (such as the web) that both accelerate the movement of information ("stuff") through the system as a whole, and circumvent established channels and relationships within it. A world in which librarians or scholars or writers can become, effectively, publishers—using electronic media to go directly to an audience—and this is just the beginning of the disruptions caused by e-text—is a very different world from the old one. No more is the scholarly economy one in which value and status are directly based on the scarcity of a narrow set of resources: the available inches of pages in the name journals, or the attentions of the marketing department in a university press, who will assure that the product becomes "visible" to a readership or acquisitions department. Rather, where everyone can be a publisher it will not be the fact that someone selects you that connects you with an audience, but something else: quality, topicality, timeliness, record of success.

That is, in order to ask whether e-media can address or represent "ambiguity", we might need to move beyond a merely pedagogical, phenomenological, or aesthetic critique of an electronic resource or tool in itself, to anchor our question within the larger context of the ambiguities that the very *existence* of such a tool introduces into the dusty realm of scholarly practices and folkways: ambiguities that are heightened, not reduced, to whatever extent the new electronic resource manages to stimulate and support something recognizable as serious scholarly work. E-media as such may be no better, nor worse, at representing ambiguity in general than any other media or format: yet they are problematic—we raise questions about them—because they introduce ambiguities where before there were none ("Is this guy qualified for the job?"). Rather, now we come to a point when the explosion of available information is finally balanced by the explosion of available ways of participating and contributing, a kind of gift economy. It could be we are coming to a moment when the long-cultivated specialization of the literary scholar actually plays against itself: what was once an advantage (as institutions of academic departments and publishers grew better defined and rigid in their roles and categories), comes to be a liability in an age when the premium is on (superficially) some facility with machines and (more deeply) the particular intellectual capacities that are required to work with emerging media. As we know, these are capacities such as versatility, adaptability, imagination, a bent for cooperation and teamwork, and the broad view towards new possibilities (even while e-text also continues to support established media formats)—without ever requiring that a serious scholar change what she or he fundamentally does, the questioning, searching and synthesizing.

REFERENCES

Epstein, Jason. *Book Business*. New York: W.W. Norton, 2002. [See also an excerpt, "Reading: The Digital Future", at **http://www.nybooks.com/articles/14318**]

PART III
AMBIGUITY AND TEXT ENCODING

Julia Flanders

Text encoding and technology enter as interlopers into the complex and tense arena of scholarly publication, in which scholars are in fact conflicted about whether they want their technologies of representation to be

transparent or not. Within this arena, text encoding is either too factual, too empirical, to be useful to the humanistic enterprise—or else, in trying to be otherwise, it trespasses on a domain in which it is seen as an alien force. On the one hand, if markup is merely a means of representing true or at least widely accepted facts about a text, then does it not also have the effect of stiffening the text, making it less supple, reducing its fruitful human indeterminacy, limiting the reader's interaction with the text?

And on the other hand, if markup is more than this—if it is a means of intervening in the text, mapping or mimicking scholarly subjectivity and experimentation—then does it not usurp or destabilize the role of the scholar? If it mimics our own interventions, it does so with a difference: that is, the ambiguities scholars want are the ones that emerge from our own human sensibilities, not the ones that come from some other, non-human domain.

This conflicted relationship with digital tools for textual representation stems from the historical roots of modern literary publication and textual editing. In this model of textual production, the text has immanence: its meaning is the quintessence of human insight and wisdom, the distillation of what makes humanity rich and deep and complex and morally sound. This is most obvious in the case of the poet, but inasmuch as the scholarly editor is editing cultural texts which carry this weight, the editor is the modern surrogate of the poet, bringing the poet's wisdom back to vibrancy by restoring and representing his or her text. The scholarly editor must have a wisdom and sensibility which matches that of the poet in order to be able to fulfill this role—must have insight into the poet's likely meaning, habitual language use, taste, and so forth. The reconstruction of the original text (regardless of one's preferences as to copy text, treatment of variants, and so forth) depends on the deployment of expert judgment to both express and control the presence of ambiguity in the text: to make that ambiguity the field of the editor's expertise, rather than a challenge thereto.

While text markup has been widely accepted as an editorial tool for the preparation of scholarly editions of many sorts (as evidenced by the existence of efforts such as the Model Editions Partnership, the Walt Whitman Archives, the Canterbury Tales Project, the Piers Plowman Archive, and many others), its domain is assumed to be limited to expressing the text's determinacies, not its indeterminacies. Indeed, this suitability for expressing textual structure and behavior in a consistent, rigorous way is taken as text encoding's chief virtue, and the development of high-quality encoding schemes focuses on establishing methods that will minimize ambiguity and indeterminacy. This approach has brought text encoding methodology to a high degree of effectiveness in representing the kinds and aspects of texts that lend themselves to this treatment.

But what of the aspects of textual communication which on the contrary *require* an attentiveness to ambiguity itself? And what if instead of simply representing the text we aim to provide a scaffolding or additional musculature that can support our readerly and critical activities—a text encoding which more actively intervenes in the textual economy being established? Consider the following brief list of concepts and domains in which ambiguity or multiplicity of meaning could play a central role in our textual work:

- the representation of scholarly disagreement within a given edition (whether at the level of the interpretation of a particular mark on the page, or the ordering of sections, or the interpretation of the meaning of a given passage)
- the representation of aspects of the text for which a controlled (i.e. disambiguated) vocabulary cannot provide sufficient nuance: aspects which, in effect, cannot be "digitized", which have infinitely fine granularity
- the representation of textual variation in a way which does not merely capture the existing readings, but also the suspension of their resolution, the ways in which they do not simply displace one another but coexist

If text markup seems to operate at a different level of abstraction from these domains, it is because of what we assume about its empirical commitments, its alliances with determinacy and fixity, with ascertaining and clarifying meaning rather than allowing it to hover before the reader. Our current understanding of text encoding is as a powerful tool for presenting alternatives, for allowing us to *choose*, rather than for helping us to probe a more complex domain in a more hesitant or searching manner.

Can text markup (and its companion tools for digital representation of text) be used to redirect our attention and ambition towards a subtler textual economy? And would such a tool find acceptance within the scholarly community? This paper will consider the possibilities within the current environment of humanistic scholarly communication.

PART IV
AMBIGUITY, LANGUAGE, AND THE SCHOLARLY ECONOMY

John Lavagnino

We hear a great deal about the difficulty of getting certain kinds of scholarship published, notably printed monographs. Yet in other respects academic work is becoming more readily available to a large public, reversing the trend of the period from the 1970s through the 1990s when there was a contraction of availability. It's not just that scholars, like many other people, can now publish on the web; it's that web publications of current and back issues of journals can have the side effect of exposing scholarship to people in other fields who wouldn't have thought of consulting them, and it happens whether the authors of the scholarship thought they were doing Web publishing or not.

JSTOR is one example of this, as it seeks to cover a broad range of fields and makes it a simple matter to decide that you'll look for your topic in political-science journals as well as the literary-studies journals you really expect to have what you want. But many online journal organizations are similar: in this field you usually see an organization doing either one journal or dozens. Economically, the logic pushes such an organization to do as many as possible once they've done a few; for the reader, it's another instance of the Web effect—if it's quick and simple to take a look at something you may well do so.

These are journal systems that cater to academic audiences; but there is already some overlap with wider audiences, who sometimes have access to such systems and who constitute another market for the future. And from the point of view of any individual field most of the academic world is a popular audience, of people who do not know a particular field's history and practices.

One of the many effects of the rise of the Web, then, is that our academic writing is likely to be encountered by people in distant fields, or perhaps not even in academic life. They may take an interest in our subjects matter without knowing the history or conventions of our fields; they may find the presentation offputting or puzzling.

One response to this situation is to change nothing. This is the arXiv.org approach: this very large online collection of preprints of scientific papers adapts an existing practice in scientific communication to the new medium, and does it well; but without changing the nature of scientific writing and without changing the audience for it. (See Kling et al for some astute discussion of this system and some other models for academic publishing that aren't journal-oriented.) The papers on a topic such as astrophysics are no more comprehensible by those outside the field than the usual run of published papers, and indeed these are mostly papers that will shortly be published in the usual journals. And it works: many people use it and it serves its purpose within the field, though it doesn't make any effort to help other audiences understand the work.

We could adopt such an approach, but scholars in the humanities will find it harder to do, because outsiders to most humanities fields still assume that the literature is (or ought to be) comprehensible to them: that academic writing about astrophysics needn't make any sense to outsiders but academic writing about history or literature should. Where arXiv.org doesn't need to make much of an effort to scare away outsiders, the equivalent in the humanities would have to take positive steps in that direction.

We should also think about whether it's in our interest to repel people interested in our work. There is an ethical argument against such a practice: in the end our work is financed by the general public and it ought to be available to them. And there is a strategic argument: we might see more funding for our work if it was more visible; if our work is likely to be more visible anyway, then seeking to make effective presentations of our work for this larger audience would be wise, as a way of inviting interest and sympathy rather than hostility.

Discussions of problems that outsiders have in reading academic writing tend to focus on jargon, but such discussions tend to be superficial; as in learning entirely foreign languages, acquisition of a particular critical language involves gaining an understanding of suggestions and implications in words that are not easily encapsulated in definitions. We can compare a comment that William Empson made: when teaching in Japan and China he warned his students away from his famous book "Seven Types of Ambiguity", because without a native understanding of the language it was misleading: his approach was not based on tracking down every possible ambiguity, but only those that had some relevance in context (xii). It's characteristic of language use within a subculture that a lot is going on that the outsider just can't see.

It is easy but ineffective to suggest that scholars should simply write with an eye to a larger audience; it can diminish the effectiveness of a contribution for its scholarly audience (by diluting its originality through the need to take up space explaining familiar things) and it is a difficult task to state the general principles underlying one's practice. We may see some movement in this direction, but it will be the aspect of scholarly writing that will be the slowest to change.

But change in other aspects of the publication of scholarly work are more open to change, in part because they're changing already in the electronic world. We know that some conventional framing devices that are part of the meaning of print publications are not readily interpreted by outsiders: such as the authority of particular journals. But that kind of framing device is already attenuated by electronic publication in many cases: the visual effect of a particular journal's typography is lost in many journals that are published online in HTML conversions, for example, and there is much more of an impetus to focus on individual articles rather than on whole numbers of a journal. Many influential journals in the humanities have had quite specific agendas, but it is comparatively rare for such agendas to be spelled out with any frequency; there is a strong incentive for electronic journals to move in that direction, though, to counteract the leveling tendency of the medium, and greater explicitness about aims helps neophytes.

When the electronic journal is encountered as part of a huge collection of journals, rather than as something with a slant and subject matter that matter for a particular subfield, there is a strong impetus to be explicit not just about overall aims but about individual offerings. Humanities journals have mostly not included abstracts of articles or systematic cross-references to related articles, but we can expect to see more of these because they are the kind of feature that make it clearer what one journal has to offer that's different from what many other equally accessible journals have.

We are starting to see the effects of large-scale electronic publication, and while these may not be as utopian as commonly predicted a decade ago, we can see ways in which they are making life better: not only providing richer resources for scholarly work, but also tending to encourage wider access to the fruits of scholarship and more mixing of the disciplines.

Écriture féminine: Searching for an Indefinable Practice?

MARK OLSEN
University of Chicago
mark@barkov.uchicago.edu

Some 25 years ago, Hélène Cixous provocatively anticipated a distinctly female practice of writing. She declared that *écriture féminine* would be marked by characteristics which challenge the logic of writing within the "phallocentric" tradition, by its focus on the female body, glorying in a femininity too long repressed, and breaking up received truth through laughter. Her argument, based on her view of the poetic, implies a form of "false consciousness" in that not all women would, or even could, produce texts from this alternative practice. Further, she declared that *écriture féminine* cannot be defined or identified outside of itself:

It is impossible to *define* a feminine practice of writing, and this is an impossibility that will remain, for this practice can never be theorized, enclosed, encoded, coded —which doesn't mean that it doesn't exist. But it will always surpass the discourse that regulates the phallocentric system: it does and will take place in areas other than those subordinated to philosophical-theoretical domination. It will be conceived of only by subjects who are breakers of automatisms, by peripheral figures that no authority can ever subjugate.[1]

A generation of French and America feminist critics have addressed Cixous' declaration, pro and con, which is based on two questionable propositions: that *écriture féminine* cannot be defined outside of its own terms and that not all women writing may be said to participate in this unique practice.

It is unfortunate that so important a declaration would be auto-marginalized by positing its own epistemological and social indefinability. While Cixous may be right, that a practice of women's writing would be hard to identify, she carries her attack of "phallocentric systems" of knowledge—rationality and logic—to such an extreme that any attempt to demonstrate the possibility of the existence of, or isolate some of the characteristics of, an *écriture féminine* is invalidated. I would argue, however, that any putative feminine practice of writing should be identifiable as recessive traits in the literary production of women who predate Cixous's declaration and that these traits may be detected using systematic methodologies.

A considerable body of recent work on gender marking in language use shows important, even critical, differences in male and female use of language.[2] While much of this work has been restricted to less formal forms of communication—speech, e-mail, and student essays—there is some evidence that gender is an important discriminant in more formal literary texts. For example, Minna Palander-Collin finds in her study of 17th century private letters that there are marked differences between female and male writing, suggesting that the women's letters are more interactional, personal and "involved" than letters by men, which

are common features of women's communication in Present-Day English.[3] More generally, recent studies by Moshe Koppel, Shlomo Argamon and Anat Shimoni have detected a wide variety of simple lexical and syntactic feature differences in literary texts by men and women in the British National Corpus (BNC). Using machine learning techniques, they are able to infer the gender of an author of an unseen document with approximately 80% accuracy, with moderately better performance for works of fiction than nonfiction.[4] The success of text categorization techniques to identify modern literary texts by gender of author suggests that there are gendered practices of writing and that these gendered traditions are grounded in the history of literary culture and would be an important component of Cixous' prospective *écriture féminine*.

In the early 1990s, I attempted to examine the question of *écriture féminine* using the ARTFL database, only to be confronted by the very significant gender bias of the TLF database as it was then constituted, concluding that the sample of texts by women (3.8% of the titles) was too limited to allow for useful comparisons. This limitation led directly to our ongoing effort to digitize a large collection of French literary texts by women, ARTFL's French Women Writers Project[5] to redress the gender bias of the corpus which was used to compile the TLF dictionary. The gender bias in the data used to compile a massive and "definitive" dictionary is itself an important example of one mechanism of how patriarchal language is propagated and authorized.[6] My initial studies of gender representation in early modern and modern French—based exclusively on male writers describing the feminine —produced some striking examples of long-term shifts in the use and meaning of common gender terms, such as *femme*.[7] Age categorization of women— young/old—becomes one of the most notable patterns only towards the end of the 18th century, reflecting both the rise of the romantic novel and a new politics of desire. Equally important are long-term continuities, such as the collocation of *femme* with possessives, suggesting that the semantic field of the feminine begins with putting her "in her place," being possessed by a male.[8] Cixous' complaint that "woman has always functioned 'within' the discourse of man" simply because women must express themselves in "the language of men and their grammar"[9] is a position that is certainly implied by my initial studies and needs to be taken into account when characterizing earlier examples of feminine writing.

In order to examine possible earlier practices of *écriture féminine*, I am assembling two corpora of about 350 literary texts each by male and female authors from the 17th to the early 20th century, balanced by time period, genre, and subject matter (or collection). The women's texts are drawn from a variety of current holdings at ARTFL, including 110 texts in French Women Writers, 70 from the ARTFL database, 40 from BASILE (Editions Champion), and 130 from various collections produced by Editions Bibliopolis.[10] The comparative male corpus will be selected from the same sources. Comparisons of the two corpora will be based on several distinct types of analysis.

- A comparative overview of the two corpora using the combination of some of the relatively simple lexical and syntactic features used by Koppel and his team, broken down by time period;
- A comparison of collocations of gender terms (*femme, homme, mari, mère*, etc.), loosely based on results from my earlier studies, effectively looking at how men and women represent themselves and the other gender, again broken down by period and genre.
- Examination of the "agency" of male and female actors as represented by the tense and the functional status of selected verbs associated with gender denoting terms.
- And a general assessment of the pragmatics associated with literary writing by males and females, by looking at use of hedges as well as the density of pronouns and adjectives.

Finally, following Pulkkinen's call to consider the history of the concept of "woman" as a political construct, I would like to look at contrasting representations of social groups—inclusive and exclusive use of *nous* and other terms denoting belonging—a technique well known in *lexicologie politique.* [11] Using a combination of quantitative and systematic qualitative approaches to the two corpora, I hope to show changing patterns of gendered writing in French literature over a several century span.

The continued expansion and improvement of electronic text holdings, in terms of the quality of the data as well as coverage of wider ranges of literatures well outside of established national canons, allows us to revisit theoretical and substantive problematics that we could not address previously. Women's writing is surely a case-in-point of this laudable development, facilitating systematic examinations of propositions made by critics and theorists like Hélène Cixous. Identification of the practices and distinguishable characteristics of *écriture féminine* in the centuries predating her "call to the pen" may help situate the traditions of gendered discourse the past as well as their relationship to current feminine writing.

REFERENCES

Hélène Cixous, "The Laugh of the Medusa" in *Signs: Journal of Women in Culture and Society* 1:4 (1976) 883.

See for example, Deborah Tannen, *You just don't understand: Women and men in conversation* (Ballantine, 1990) and *Gender and discourse* (Oxford, 1994).

Minna Palander-Collin, “Male and female styles in 17th century correspondence: I THINK,” in *Language Variation and Change*, 11 (1999), 123–141.

Moshe Koppel, Shlomo Argamon, Jonathan Fine and Amat Shimoni, “Automatically Categorizing Written Texts by Author Gender”, forthcoming in *Literary and Linguistic Computing* (2003) and Moshe Koppel, Shlomo Argamon, Jonathan Fine and Amat Shimoni, “Differences in Writing Style Between Male and Female Authors” (paper submitted for publication).

ARTFL’s French Women Writers Project is one of many projects to digitize neglected literary and non-literary texts by women, most inspired by the Brown Women Writers Project, including the University of Chicago Library’s Italian Women Writers and commercial products such as Alexander Street Press’ *North American Women’s Letters and Diaries*. See **http://www.lib.uchicago.edu/e/ets/efts/Women.html** for a partial list of current projects and products.

“Gender representation and *histoire des mentalités:* Language and Power in the *Trésor de la langue française*,” in *Histoire et measure* VI (1991): 349–73.

“Quantitative Linguistics and *Histoire des mentalités*: Gender Representation in the *Trésor de la langue française*,”, in R. Köhler and B. B. Rieger (eds.), *Contributions to Quantitative Linguistics* (Kluwer, 1993), pp. 361–381.

See also Tuija Pulkkinen, “The History of Gender Concepts: The Concept of Woman”, in *History of Concepts Newsletter*, 5 (2002), 2–5.

Cixous, p. 887.

Links to the ARTFL/PhiloLogic implementations of these collections may be found at **http://www.lib.uchicago.edu/efts/ARTFL/newhome/texts/**.

“Enlightened Nationalism in the Early Revolution” The *nation* in the Language of the *Société de 1789*, in *Canadian Journal of History* (24), 1994, p. 28ff

Chasing DTDs. The Digital Edition of the ‘Repertorium Biblicum Medii Aevi’

SABINE HARWARDT
University of Trier
harwardt@uni-trier.de

STEFAN BÜDENBENDER
University of Trier
bued2101@uni-trier.de

THE REPERTORIUM BIBLICUM MEDII AEVI

The *Repertorium Biblicum Medii Aevi*, edited between 1950 and 1980 by Friedrich Stegmüller and Klaus Reinhardt in 11 volumes, is a major reference work for studies in various medieval disciplines such as theology, history, philology, and philosophy. Within its approximately 12,000 catalogue entries, listing almost 24,000 commentaries, the ‘Repertorium’ includes all commentaries of the Bible that have been written until 1500. The single volumes, published at the ‘Consejo Superior de Investigaciones Cientificas’ (Madrid), have been subdivided into Apocryphes, Commentaries of known and unknown authors, Supplements, and Glossa ordinaria. Thus, the ‘Repertorium’ contains the largest part of the medieval commentary tradition of the Bible.

In July 2002, a team of philologists and computer scientists started to prepare a digitized version of the ‘Repertorium Biblicum Medii Aevi’ taking into account scholarly demands in order to open up new possibilities for tackling this enormous amount of valuable data.

TARGETS, STARTING POSITION

In the printed edition of the ‘Repertorium Biblicum’, the extensive material is structured by the registers of the incipits (beginnings) of a given commentary. However, apart from these incipits, the ‘Repertorium’ also contains short biographies of its authors, bibliographies of research literature, dates, commented works, and the manuscript tradition as well as the tradition of early prints.

The electronic edition of the ‘Repertorium’ will enable all users to search within these and further categories such as specific manuscripts, libraries and archives, commented Bible-books, various types of

commentaries, specific authors, or dates of special importance. Moreover, many supplements and corrections collected after the publication of the printed ‘Repertorium Biblicum’ will be integrated into the digital ‘Repertorium’; scholars will thus have the possibility to work with the updated material on-line.

Currently the data input is taking place; it will be completed in January 2003. The 2nd volume of the ‘Repertorium Biblicum’ has already been digitized and marked up to a certain degree in order to analyze its structure and to develop tagging routines which can be easily adapted to the structure of the other volumes afterwards.

FROM DOCUMENT-ANALYSIS TO ENCODING

The structure of the ‘Repertorium Biblicum’ is very complex. It can be described as an entry catalogue of authors and/or of titles which are given in alphabetical order. In general, an entry has three different parts: 1. work, 2. chronological data and short bibliography, and 3. commentaries of that specific work. An entry can consist of part 1 and part 2 only; if part 3, which can be repeated several times, is given at all, part 2 can be omitted. For inserting the markup, these three parts, with the logical order of an entry, have to be distinguished thoroughly.

Part 3 is the most complex one to be encoded. It consists of two subdivisions, namely of quotations taken from a specific work in the form of incipits or explicits, and of sometimes lengthy lists of manuscripts and editions of the work quoted. Whereas all entries follow the strict alphabetical order and show dictionary patterns (macrostructure), the information on manuscripts given in part 3 rather resembles a repository guide or an archival finding aid (microstructure)—a fact that forces one to combine two different logical structures.

On the basis of a thorough document analysis, we fixed criteria for the markup and encoded the 2nd volume of the ‘Repertorium’ using an internally defined tag-set. The encoding has been carried out using TUSTEP programs. Being designed to cope with mass data, TUSTEP disposes of very powerful and highly specific algorithms which allow to search large amounts of text for recurring patterns. We can than—within TUSTEP—develop our own procedures to automatically copy, modify and sort these patterns according to our individual demands. In the future, the internal tag-set could easily be replaced by a standardized DTD such as TEI or MASTER.

A TESTBED FOR THE SELECTION OF VARIOUS DTDS

Since the development of a new DTD is a very expensive and time-consuming process, we wanted to check whether DTD schemes already in existence could easily be applied to the ‘Repertorium’, taking into account its varying structures. In encoding the ‘Repertorium’, the markup of its dictionary structure is as important as the markup of its archival finding aid patterns. We therefore examined the TEI DTD, which has been successfully applied in various projects at Trier University (cf. Panel Session “Into the Depths of Data. Methods of Subject Specific Content Retrieval” at the ALLC/ACH-Conference 2002 in Tübingen), Docbook, with its rich options for bibliography encoding, and, especially concerning the archival character of the ‘Repertorium’, EAD (Encoded Archival Description) and MASTER.

Having implemented various types of markup, it seems as if Docbook’s possibilities for covering complex technical writing and simple bibliographies were too restricted for encoding the ‘Repertorium Biblicum’. With TEI, the ‘Repertorium’ could be fully encoded by using a great quantity of particularly specified attributes. This could, however, complicate the proper markup and its understanding.

It is not always possible to describe the ‘Repertorium’s structure according to the DTDs especially created for the markup of archival material or manuscript archives; e.g. the ‘Repertorium’ gives the relevant bibliographical data after incipits are quoted etc. While EAD partially covers the criteria of alphabetical order combined with various quotations, its logic does not follow the listing of manuscripts in the third part of a ‘Repertorium’s entry. The definition of MASTER’s <msDescription>-tag, that would fully cover the complex structure of the repository guide part, does not allow one to list several manuscripts after quotations, requesting instead a strict sequence of manuscripts.

At the moment, TEI, MASTER, and EAD are being benchmarked against each other by encoding the 2nd volume according to these standards. MASTER seems to be most appropriate for encoding; however, some questions remain. Our paper will discuss the final decision for MASTER confronting the three DTDs’ advantages.

THE FUTURE OF DIGITIZATION AND RESEARCH

By now, the 2nd volume has been encoded according to an internal markup derived from the document analysis. As mentioned above, this markup could easily be replaced by tags compliant to one of the possible DTDs (MASTER, TEI, EAD, Docbook). At the time of the conference, the majority of the ’Repertorium Biblicum’ will be fully encoded (probably according to MASTER); it will also be partly accessible for varied on-line research. Questions to be discussed when presenting the ‘Repertorium Biblicum’ may focus on problems of the applicability of defined frameworks like TEI to complex documents with an inconsistent and often irregular structure.

REFERENCES

The Cambridge History of the Bible vol. 2: The West from the Fathers to the Reformation. Ed. G. W. H. Lampe (Cambridge 1969).

Klaus Reinhardt - Horacio Santiago Otero, *Biblioteca Bíblica Ibérica Medieval* (Madrid 1986).

Klaus Reinhardt: *La Biblia en la Península Ibérica durante la edad media (siglos XII-XV): el texto y su interpretación* (Coimbra 2001).

Homepage: **http://www.uni-trier.de/repbib**

The Tobacco Documents Corpus: Archiving the Industry

CLAYTON DARWIN
University of Georgia
cdarwin@uga.edu

WILLIAM KRETZSCHMAR
University of Georgia
kretzsch@uga.edu

DONALD RUBIN
University of Georgia
drubin@uga.edu

Our research group has been awarded funding by the National Cancer Institute for a rhetorical analysis of "deception" in the Tobacco Documents (TDs). These documents, which were released by tobacco industry defendants as a result of state and federal litigation and legislative hearings, cover the complete range of corporate operations in the tobacco companies, from memos to research papers to procurement invoices. The documents are stored physically in depositories in Minneapolis (the site of the original trial) and Guildford, England, and large collections of them (more than five million documents) are now available in electronic form on the Web as well. The documents represent a rich source of corporate and technical discourse which had never been subjected to systematic linguistic analysis; indeed, we are not aware that any similar corporate body of documents has ever been available for analysis.

Rather than choosing specific documents for analysis, a method which would leave itself open to attack on grounds of highly selective use of data, the premise of our work is to treat the TDs as a corpus, and to apply accepted methods of corpus and forensic linguistics and rhetorical analysis. Of course, this required that we create sub-corpora for study, since we do not have the resources to include the entire set. Here we will present our experience with the planning and creation of our TD corpora: the sampling strategy, archiving, retrieval, and ultimately, making the corpora available via the Internet for further research.

Our initial goal was to create a series of corpora from the TDs in order to 1) Identify TDs in which rhetorical manipulation ("deception") may have occurred, and to estimate the extent and prevalence of manipulation; and 2) Analyze manipulation we find in order to classify it and develop means to identify similar manipulation in other industrial situations. To do so we have followed a three-part strategy for corpus creation which emphasizes rigorous sampling methods. We first drew a limited sample from the entire body of TDs so that we could determine the best classification of text types and estimate their proportions within the overall body of texts. From those text types which we considered relevant to (i.e. subject to) rhetorical manipulation, we devised quotas for creating a reference corpus of approximately 500,000 words, which we estimated to consist of 808 documents. For this reference corpus, all relevant TDs were sampled whether or not they were thought to contain any manipulation. Finally, we are presently compiling a corpus which includes all texts which we determine to contain any rhetorical manipulation, along with parallel corpora of earlier drafts of the same texts or versions of the same texts prepared for other audiences, so that detailed analysis of rhetorical manipulation can be carried out for itself and by comparison with cross-draft and cross-audience TDs. As it has turned out, the plan has been effective in the first two parts which we have now completed, but we have had to make adjustments at several points in order to take account of our preliminary findings.

Once we began the process of collecting documents we immediately encountered two problems related to archiving and processing the data. The first is that there was no text available. Rather than being stored digitally as the plain ASCII text which we needed for computer-assisted corpus analysis, the tobacco documents are stored as image files, usually as TIF type. This problem was compounded by the fact that the images, although stored as large high-resolution files, are generally too poor in quality for automated text processing such as scanning and OCR. They often have pages that are tipped to one side or, in the case of dot-matrix or fax printing and handwriting, they can be practically illegible. The second problem we encountered was the structural complexity of the documents themselves. For example, just over 50 percent of the documents contained marginalia of some type, such as filing data, distribution lists, stamps of various types, editing, and handwritten comments. Most documents contained large amounts of peripheral data like names, dates, addresses, and distribution lists. Although these features are significant for archiving, they have little or no rhetorical value for our intended research. Other documents often contain or consist of forms, tables, and images that also offered little value for our analysis. To account for these problems we decided to keyboard the documents by hand as plain ASCII text and to code them with XML tags.

When we investigated the existing XML tag sets, TEI in particular, we found that they are particularly well suited for archiving standard texts in standard hierarchies and for naming typesetting conventions. However, we chose to devise a set of tags specific to our project for primarily two reasons. First, we found that many of the documents collected for the corpus had a very non-standard format. In fact, we found no fixed definition for what constitutes a document in the tobacco archives. For example, a document recently coded began in the middle of a paragraph and sentence, proceeded for half of a page, changed to a summary of a court ruling, then to a policy letter, then to a table of denicotinization, and ended with a diagram of a processing facility. The second reason for devising our own tag set is that our primary interest is in archiving rhetorically significant text and events rather than the typesetting conventions used to represent them. Thus although TEI includes a full set of tags to indicate divisions and typographical conventions, use of these tags for our purposes might lead to ambiguities. For example, italics, boldface, and underlining have all been found to denote emphasis in the document set, which is rhetorically significant; however, they have also been found to denote titles, headings, names, quotations, formulas, and standard text, which may be of little value for our analysis. Thus, tagging a word in a document with a tag designed to denote typesetting, such as italics, may not be so useful when the corpus is analyzed linguistically or rhetorically simply because there is no way to know the significance of that particular event. To counter this, we have devised a set of XML tags which accommodates the structural complexity of the original documents and which reflects the purpose of our study.

Data is retrieved from the XML files in a straightforward manner. We have embedded the expat XSLT engine into several Python scripts. This allows us to assemble a text corpus for study from the reference and manipulated-cases corpora according to the needs of the research. That is, with our scripts the XML files are parsed, desired tag content is selected, and the selected content is assembled and written to an ASCII text file. There are, however, two notable differences between the standard Web use of XSLT and that of our project. The first is data permanence. The output of our XSLT processing is ASCII/ANSI text which is written to file for later analysis rather than HTML sent onto the Internet. The other difference is that the XSLT output is not solely determined by the XSL stylesheet. For ease and speed in processing, some general document and tag selection is done by regular expressions in the Python script prior to calling the expat program.

The end result of this initial phase of our project will be a larger general corpus of TDs for use as a reference, and a smaller corpus of “manipulated” TDs for focused analysis. Both will be archived as ASCII text with XML tags, which will allow us to generate tailored sub-corpora for specific studies using XSLT. Although these corpora are being created for our own purposes, our intent is to make them freely available to other researchers over the Internet. What we envision is an integrated Web site that provides access to the corpora in several formats: the TIF and/or PDF images of the original documents, the XML files coded with TEI compliant tags, the XML files coded with our tag set, ASCII text versions of the files, and access to a CGI version of our XSLT scripts for generating task-specific sub-corpora.

Linguistic Corpus Construction and Analysis Before and After the IT Revolution: The Newcastle Electronic Corpus of Tyneside English in the 1960s and Now

HERMANN MOISL
University of Newcastle
hermann.moisl@ncl.ac.uk

URL: **http://www.ncl.ac.uk/necte**

The theme of this conference is the impact of IT generally, and of the Web in particular, on humanities research. The Newcastle Electronic Corpus of Tyneside English (NECTE) project is an ideal case study on this theme. It is, in large part, based on the Tyneside Linguistic Survey (TLS), which, in the decade 1965-1975, attempted to construct an electronic corpus of the distinctive 'Geordie' dialect spoken in north-eastern England, and to analyze it computationally. That attempt failed, but, because of its importance—both historically and to the cultural identity of the region—we have received research council funding to salvage the original TLS materials, amalgamate them with a more recent dialect survey, and produce a state of the art web-based electronic resource of Tyneside English. We are, therefore, in an excellent position to assess the impact of the IT revolution on corpus construction and analysis.

HISTORY

The NECTE project amalgamates two separate corpora of recorded speech, one of them collected in the late 1960s as part of the TLS project (cf. Strang 1968), and the other in 1994 (cf. Milroy et al. 1997). The TLS material is the object of interest here. It originally consisted of audiotaped with 100 informants drawn from a stratified random sample of Gateshead in North-East England. Many, but not all, of the interviews were orthographically and phonetically transcribed. These transcriptions were then electronically encoded, and in that form were the basis for subsequent computational analysis. Although several publications emerged from the corpus in the interim, there was no further work on the corpus until 1995, when it was properly archived. In 2001 we were awarded a substantial research grant to produce an enhanced electronic corpus resource from a combination of the TLS and the 1994 collections which will eventually be made available to the research community in a variety of formats: digitized sound, phonetic transcription, and standard orthographic transcription, all aligned and available on the Web. This process is now well advanced.

ANALYSIS

In the late 1960s the TLS research team pioneered a methodological approach that is still radical today. In contrast to the theory-driven methodology, which was universal in sociolinguistic accounts of the 1960's/1970's—and which still predominates—the TLS proposed a fundamentally empirical approach in which salient factors are extracted from the data itself, and then serve as the basis for model construction. Unlike the Labovian paradigm, therefore, social and linguistic factors were never selected by the analyst on the basis of a predefined model of either language or society.

To this end, the project created an electronic corpus from a subset of the data and applied cluster analysis to it. Interesting preliminary results were published, but nothing was done thereafter: the TLS never completed its research program, and consequently failed to make a significant impact on the research community. This failure is, we feel, primarily due to implementation issues in general and more specifically to the inadequate computational / IT resources available at the time. After 30 or so years the issues which proved so intractable to the TLS research group have been resolved, and it is now possible both to bring the TLS agenda to completion, and also to augment it in important ways.

In what follows, we first identify the computational / IT factors that confronted the TLS project, then describe how these have since been resolved, and finally draw some general conclusions as to how the IT revolution has impacted upon the construction and analysis of linguistic corpora.

PROBLEMS

Hardware

The rudimentary computational hardware of the late 1960s and early 1970s seriously hampered what could be achieved:

- Data input using punched cards was slow and error-prone, and error correction via the same

medium was unreliable. Creation of electronic files from what is, by contemporary standards, a moderate-sized corpus was a major undertaking, and only slightly over half the material was ever digitised as a result.

- Limited memory and low processing speed prevented some tasks being done within a reasonable time, or indeed at all. Cluster analysis of the TLS electronic files had to be done in several stages because the entire data set could not be held in memory, and this had significant consequences for the usability of the results.
- Output was entirely text-oriented; there were no graphics facilities. Direct visualization of analytical results, such as cluster dendograms, was consequently difficult.

Software

- Operating systems of the period handled basic file I/O, but very little application software was available. The TLS project team had to write most of its own processing software.
- The available character set was restricted to the standard upper and lower case letters, numerical digits, and punctuation marks. Fonts for IPA symbols and the elaborate system of diacritics which TLS used for fine phonetic distinctions were unavailable and so these could not be directly visualized; the project had to work instead with numeric codes for these symbols.

Publication

The above hardware and software factors were inconveniences that could be, and to some extent were, overcome by TLS. Publication of the corpus itself and of analytical results was, on the other hand, a genuine impediment:

- Lack of portability: The absence of any generally accepted standards for electronic corpus construction and data encoding combined with the fact that programs for processing the data were project-specific meant that access to the material was not easy for other researchers.
- Lack of a convenient delivery medium: The only way to transfer data from one site to another was physically to carry digital tape from place to place. There is no indication in the project papers of any awareness that it might be possible and, indeed, beneficial from a preservation perspective, for the electronic corpus to be provided to other researchers.

RESOLUTIONS

Thirty or so years on, none of the above constraints apply:

- **Hardware**: Memory size and processing power have progressed enormously since the TLS era, and neither is now an issue for corpus linguists. There is no longer a significant limit on corpus size or on the amount of data that can be simultaneously processed in numerical analysis. Hardware improvements have also made it possible to develop and implement computationally more demanding cluster analysis algorithms, such as self-organizing maps, which could not have run in a reasonable time earlier on. High resolution graphics now make visual display of phonetic symbols and cluster diagrams straightforward.
- **Software**: There is now a very wide range of ready-made software for statistical and cluster analysis as well as for ancillary purposes, so it is rarely necessary to write bespoke software, and there are specialty fonts for the display of phonetic symbols.
- **Publication**: There are now standards for character sets (Unicode) and document structuring (XML) which make corpora that adopt these standards portable and directly usable by other researchers. The connectivity of the Internet, together with the pervasiveness and accessibility of the Web, have opened up new possibilities for corpus publication.

CONCLUSIONS

It is clear that, in attempting to implement its radical research agenda, the TLS was well ahead of its time and bit off more than it could technologically chew. Substantially more powerful hardware and software were a precondition for success, and this is what the past two decades or so have provided: it is now possible to construct and analyze electronic corpora in the manner in which the TLS project intended, and to publish it as a resource for the research community in a way that its members did not, and could not, conceive.

This conclusion does not, however, exhaust the impact of the IT revolution on corpus linguistics, as we have found in our revision of the TLS. There is a second main factor, and it is sociological rather than technological. The TLS worked largely in isolation from the rest of its research community, and feedback on output was determined by publishers' schedules. By contrast, we are part of a global research community that is, in principle, in constant and instantaneous communication via email and the Web. In this community, information can be shared, draft research output informally peer reviewed, and projects monitored as they proceed, all effectively online. Any research group that participates in this community can bring the current state of discipline knowledge to bear on its work, with consequent benefits. For us, this is a major advantage over the originators of the TLS, and one which, we hope, will allow us to do justice to their efforts.

REFERENCES

Jones, V. (1985) 'Tyneside syntax: A presentation of some data from the Tyneside Linguistic Survey', in Viereck, W. (ed.) *Focus on England and Wales*, pp.163–177. Amsterdam: John Benjamins.

Labov, W. (1972) *Sociolinguistic Patterns*. Philadelphia: University of Pennsylvania Press.

Local, J.K., Kelly, J. and Wells, W.H.G. (1986) 'Towards a phonology of conversation turn-taking in Tyneside', *Journal of Linguistics*, 22: 411–437.

Milroy, L., Milroy, J. and Docherty, G. (1997) *Phonological Variation and Change in Contemporary Spoken British English*, Unpublished Final Report to the UK ESRC, grant no. R00234892.

Pellowe, J. et al. (1972) 'A dynamic modelling of linguistic variation: the urban (Tyneside) linguistic survey', *Lingua*, 30: 1–30.

Pellowe, J., and Jones, V. (1978) 'On intonational variety in Tyneside speech', in Trudgill, P. (ed.) *Sociolinguistic Patterns of British English*, pp.101–121. London: Arnold.

Strang, B.M.H. (1968) 'The Tyneside Linguistic Survey'. *Zeitschrift für Mundartforschung*, NF 4 (Verhandlungen des Zweiten Internationalen Dialecktologenkongresses), pp.788–794.Wiesbaden: Franz Steiner Verlag.

Developing Markup Metaschemas to Support Interoperation among Resources

GARY SIMONS
SIL International
gary_simons@sil.org

INTRODUCTION

The work presented in this paper grows out of the EMELD project, a project which seeks to develop "Electronic Metastructures for Endangered Language Data"[1]. One of the major aims stated in the project proposal is the "formulation and promulgation of best practice in linguistic markup of texts and lexicon"[2]. The project is attempting to do this by forging a community consensus through a series of workshops[3].

The first workshop easily reached consensus that the best format for the interchange and archiving of endangered language data is XML-based markup. It just as easily reached consensus that no single system of XML markup could be imposed on all language resources. At the same time, there is consensus that linguists need to be able to perform queries across data sets, even if they do not use the same markup. This paper describes the solution that is being developed to support this kind of interoperability across resources that use different markup systems. Before describing the details of the solution, some definitions and requirements are elaborated.

DEFINITIONS

A markup language, like a natural language, has a lexicon, syntax, and semantics. The following terms are used throughout this paper to refer to the descriptive artifacts that document these three aspects of markup:

- **markup vocabulary**: Enumerates the lexical inventory of markup: i.e., the set of elements and attributes that are used in marking up a resource. (In practice, the vocabulary is enumerated within the markup schema rather than in a separate document.)
- **markup schema**: Specifies the syntax of markup: i.e., a formal grammar defining constraints on where elements and attributes must or may occur with respect to embedding and relative order and on what their values may be. (This is typically realized in an XML DTD or an XML Schema, though other mechanisms are emerging.)
- **markup metaschema**: Specifies the semantics of markup: i.e., a formal mapping from elements and attributes to the linguistic concepts they represent. (This area of markup is not as well developed as the syntactic area, but is beginning to be developed under the impetus of the so-called Semantic Web[4].)

REQUIREMENTS

Given that the markup up of language data will be in XML, what is the nature of the markup vocabulary? The following is the basic requirement on the markup vocabulary, along with consequent features of the

implemented solution:

- Linguists need to be able to do more than just read texts and lexicons in display format; they also need to be able to manipulate the content by selectively accessing individual items of information.
    1. The archival form of electronically encoded resources should not follow a strategy of presentational markup; that is, the markup vocabulary should not be one that simply identifies what the information will look like when displayed.
    2. The archival form of electronically encoded language resources should follow a strategy of descriptive markup; that is, the markup vocabulary should identify what the individual pieces of information are from a linguistic point of view.
    3. The markup vocabulary for a particular text or lexical resource should identify all of the elements of information (not just some of them) that go into the analysis of the text or the description of each lexical item.
    4. Users still need a presentational display of the resource; this should be accomplished by applying a stylesheet to the descriptively marked up resource.

HTML markup, when applied to language resources, is an example of presentational markup. It does not offer linguists the ability to do automated processing of a linguistic nature, for instance, to perform a query like ”What are the part-of-speech categories used in this lexicon?” For this purpose a descriptive markup vocabulary that specifically identifies the linguistic significance of each piece of information is needed. But simply having a markup vocabulary is not enough; for each language resource there is also a grammar that defines how the individual markup elements combine to form a valid text or lexicon.

- The linguist creating a text or lexicon needs for the markup of the resource to be consistent with his or her plan for its content and structure.
    1. A single markup schema that sanctions all common practices in structuring the content of language resources will be too permissive to constrain any single resource to the specific plan of its creator[5].
    2. There is enough convergence of practice that it will be possible to develop one or more specific markup schemas that can be recommended for widespread use while being adequately constraining.
    3. There will always be plans for content and structure that are unique enough to require that a unique markup schema be devised for the resource.

These consequences of requirement 2 mean that there will be multiple markup schemas, even in the context of best practice. In order to achieve interoperability of resources when there are multiple markup schemes it is necessary to introduce a meta-level in the approach to markup:

- Linguists need to be able to query and otherwise manipulate multiple texts or lexicons in a single operation, even though they may individually have different markup vocabularies and schemas.
    1. As a foundation for interoperability, there must be a shared ontology for the kinds of information that are marked up in language resources.
    2. As the bridge to interoperability, each resource must have a metaschema that formally documents how the elements and attributes of its markup schema map onto the concepts of the common ontology.
    3. The metaschema must be separate from the language resource (rather than being an integral part of it) so that multiple resources can share the same metaschema.
    4. It must be possible for a third party to create a metaschema for a resource that lacks one without changing the resource itself. (This implies that the linkage from metaschema to schema to resource is specified in a stand-off manner through metadata.)

IMPLEMENTING METASCHEMAS

The ontology which serves as the foundation for interoperability (3a above) is under development as one of the EMELD project deliverables[6]. The complete paper will present the details of how metaschemas are being implemented. In brief, a metaschema is an XML document that formally expresses the mapping of the elements and attributes in a markup schema to the concepts in the linguistic ontology. The exact meaning of a particular instance of an element or attribute is often dependent on its context in the entire markup structure. In order to specify relevant contexts, the metaschema uses an XPath expression[7].

The metaschema also maps the elements of markup that define structure onto the generic structures of an abstract data model[8]. As a result, a metaschema specifies an equivalent abstract document for any document instance that conforms to its corresponding schema. The elements of the abstract document correspond to generic structures; the specific concept of the ontology that identifies the linguistic meaning of

a particular element is encoded as the value of an attribute of the abstract element.

Interoperability is achieved by means of an XSLT script that compiles the metaschema into an equivalent XSLT script that implements the translation of any document instance into its equivalent abstract document. The complete paper will demonstrate how this transformation process yields documents with comparable markup from documents that originally had different markup schemas.

NOTES

[1.] The project involves five host institutions and is sponsored by a five-year grant from the National Science Foundation. See: **http://saussure.linguistlist.org/cfdocs/emeld/index.cfm**

[2.] Section 3.1 of **http://linguist.emich.edu/%7Eworkshop/E-MELD.html**

[3.] 2001 workshop, "The Need for Standards": **http://saussure.linguistlist.org/cfdocs/emeld/documents/2001docs2.cfm** and 2002 workshop, "Digitizing Lexical Information": **http://saussure.linguistlist.org/cfdocs/emeld/workshop/2002/papers02.html**

[4.] The Semantic Web is an activity of the W3C: **http://www.w3.org/2001/sw/**

[5.] The TEI DTD for dictionaries is an example of a markup schema that is so general as to be too permissive for any one project. For instance, see "Using architectural processing to derive small, problem-specific XML applications from large, widely-used SGML applications," Gary F. Simons, SIL Electronic Working Papers 1998-006, **http://www.sil.org/silewp/1998/006/**.

[6.] Scott Farrar and D. Terence Langendoen, 2002, "GOLD: A General Ontology for Linguistic Description," EMELD working paper, **http://saussure.linguistlist.org/cfdocs/emeld/documents/gold_draft4.doc**. See related papers at **http://emeld.douglass.arizona.edu:8080/group.html**

[7.] XML Path Language (XPath) Version 1.0, W3C Recommendation, 16 November 1999, **http://www.w3.org/TR/xpath**

[8.] Nancy Ide and Laurent Romary, 2001, "Standards for Language Resources," Proceedings of the IRCS Workshop on Linguistic Databases, **http://www.ldc.upenn.edu/annotation/database/papers/Ide_Romary/29.3.pdf**.

Peer Review of Humanities Computing Software

STÉFAN SINCLAIR
University of Alberta
Stefan.Sinclair@UAlberta.ca

JOHN BRADLEY
King's College London
john.bradley@KCL.AC.UK

STEPHEN RAMSAY
University of Georgia
sramsay@hardy.english.uga.edu

GEOFFREY ROCKWELL
McMaster University
grockwel@mcmaster.ca

RAY SIEMENS
Malaspina U-C
siemensr@MALA.BC.CA

The production of peer reviewed scholarship is the single most important activity for professional advancement in academe, including tenure, promotion, and salary increases. The development of software for Humanities Computing has been identified as a crucial need in our field[1], and yet, because of a lack of peer review mechanism for software, computing humanists lack an important incentive to engage fully in programming activities. Why divert valuable time and effort to software development when the payoffs are

generally much greater for the production of articles and books? We believe that this conundrum is a dominant factor in the current dearth of specialised text analysis tools.

This panel will examine the issues surrounding software development in Humanities Computing, and explore possible mechanisms for peer review.

In addition to encouraging software development by computing humanists— those best suited to understand the needs of the community—a peer review process would establish best practices and guidelines that would likely be useful to most developers. A significant number of developers in the Humanities are programming autodidactics who have had little or no formal training in software development and who would especially benefit from guidelines established specifically for the circumstances of Humanities Computing. Guidelines would also be an extremely valuable resource for instructors who wish to teach programming to Liberal Arts students (computer science courses tend to focus on business, scientific and engineering problems using lower-level, strongly-typed languages). Though guidelines for software development could be formulated independently, a peer review mechanism would provide a strong impetus for doing so.

Peer review would also promote a sustained discussion on the software needs of the Humanities Computing community, and future directions for development. Assessing the value of a particular piece of software assumes a reasonable notion of what already exists and what would be most useful to have. A review process would be an opportunity for both reviewers and developers to examine critically the current state of affairs and reflect on where efforts would best be concentrated.

The theoretical and practical challenges to formulating a peer review process for software development are numerous. In order to consider whether or not software should even be considered for peer review—as Humanities Computing scholarship—we need to recognise that software packages can be complex objects consisting of multiple dynamic parts, including:

- code (new, modified, or integrated)
- interface (design, frameworks for various platforms or languages)
- documentation (as comments in code, APIs or instructions for developers, and help for end-users)
- other documents (research statements, technical articles, representative results)

Assessment of each of these components, together or separately, can be further complicated by the circumstances of the development team: project leaders who may not have done any coding, research assistants who may not have contributed to more recent versions, components that may have been developed by external contractors, etc. Software resources tends to be organic entities that resist the fixity to which review of published materials is accustomed, and perhaps dependent.

Meshed with the challenges of defining the nature of the object(s) to be reviewed is the question of who would be qualified and appropriate to do the reviewing of which parts, and based on what general principles and criteria. The relatively small number of developers in Humanities Computing poses a heightened risk of biased reviews, at least for the code component. However, professional programmers (in the public or private sectors) may not be an appropriate alternative, since their concerns and priorities would likely be quite different.

There are useful materials that can provide some guidance as we work through these issues, such as guidelines produced by the MLA[2], and “The Stoa: A Consortium for Electronic Publication in the Humanities” <**http://www.stoa.org/**>.

The panel will begin with a summary (by Sinclair) of the more prominant issues surrounding peer review of software in the humanities. Then each of the participants to the panel will deal individually with one or more of the following questions:

1. Can software tools be considered original contributions to the field comparable to other contributions?
2. Practically speaking, can software be reviewed? Can peers be found who can review software and would they do it? What would they be expected to do when reviewing?
3. What exactly would be reviewed? Functionality, appropriate ease of use for humanists, new algorithms, multimedia...
4. Peer review is usually part of a larger process of publication. How would peer review intergrate into a publication cycle? What outcomes would there be for a community that supported this?

Following each speaker, there will be an opportunity for discussion on specific points raised. After all of the panelists have contributed, there will be a general discussion. Finally, ten minutes will be reserved at the end to formulate a plan of action for future progress.

Despite the challenges involved, the Humanities Computing community is in urgent need of software that is worthy of the sophisticated text encoding schemes that exist for editing and publication. In order to spark a concerted and sustained effort of development, we need to establish mechanisms to recognise institutionally the time and effort that are required, and the valuable contribution to Humanities Computing

scholarship that well developed software represents.

NOTES

[1.] For enlightening details on the history and function of peer review, see “Peer Review and Imprint” in *The Credibility of Electronic Publishing: A Report to the Humanities and Social Sciences Federation of Canada*, Ray Siemens (Project Co-ordinator), **http://web.mala.bc.ca/hssfc/Final/PeerReview.htm**.

[2.] See especially the 2000 MLA report entitled “Guidelines for Evaluating Work with Digital Media in the Modern Languages,” **http://www.mla.org/reports/ccet/ccet_guidelines.htm**.

Figura: A Tool for the Collaborative Editing of Non-nesting Content

RAFAEL ALVARADO
Princeton University
alvarado@princeton.edu

SARAH-JANE MURRAY
Princeton University
sjmurray@princeton.edu

Figura is a web-based database application designed to support the Charrette Project at Princeton University (Uitti 1997). In particular, it supports the work of creating a TEI-compliant critical edition of the Old French manuscript tradition of Chrétien de Troyes's *Le Chevalier de la Charrette*. The application addresses the shortcomings of a traditional, purely document-centric approach to humanities computing applications (described below) through the use of a database management system that acts as a pre-processor for marked up documents. The motivation behind the use of a database has not been to supplant the primary role of XML and the TEI in the development of digital critical editions, but rather to avoid having to employ an "extreme markup" solution that would further complicate an already difficult set of technologies. Instead, Figura targets the black box of traditional humanities computing applications-the indexing engine-and replaces it with something that is more accommodating to the specific needs of scholars, all the while keeping document encoding standards intact.

The traditional approach to humanities computing applications is one in which primary source materials are marked up using a textual encoding standard, such as the TEI (Sperberg-McQueen, Burnard et al. 1994), in order to produce "thick documents" that contain both primary and interpretive content in a single document or collection of documents. These documents are then made available for use by the scholarly public by means of a transformation engine that can generate content in a standard, viewable format, such as HTML, and an indexing engine that will allow the document web to be searched, presumably taking advantage of the rich markup found in the source documents.

A primary task of the Charrette Project has been to encode a set of rhetorical figures—e.g. instances of chiasmus, adnominatio, enjambment, etc.—in the critical edition of the text (Uitti and Foulet 1989). Because of their large number and radically non-nesting character, however, the prospect of directly encoding the figures into the text, using the technique of segmenting and splicing elements, has seemed impractical at best. In addition, the Charrette Project has been collaborative from the outset, involving the work of many editorial assistants, both over time and at any given time. Each editor has been in charge of a figure type, and has been responsible for locating figure instances in the entire document. In the traditional approach, this division of labor would have to be carried out serially, and therefore the length of time required to complete the project would multiply by the number of assistants involved. Each of these problems was solved through the use of a database to store the textual content of the Foulet-Uitti edition.

The problem of encoding non-hierarchical and multiple hierarchies of content objects using markup technologies such as SGML and XML is well documented (Renear, Mylonas et al. 1996; Alvarado 1999; Sperberg-McQueen and Huitfeldt 1999). In areas where this problem cannot be ignored, such as the analysis of qualitative data and discourse, the principle of standoff markup has been developed by several groups (Thompson and McKelvie 1997; Müller and Strube 2001; Glass and Eugenio 2002). Figura employs a methodology similar to that of these examples—which makes sense, given the kinship between discourse analysis and rhetoric—but makes use of a relational database, rather than a collection of documents, to store the links between textual elements and their affiliations in figural elements. The advantage of this approach is that editors are freed from the well-formedness constraint of XML. Once the data is entered, an algorithm may be applied to the join the source document and its non-nesting elements, to produce well-formed XML—or even MECS—using a variety of techniques, such as automated splicing or CSS grouping.

Standoff markup also allows for rich, secondary content to be stored independently of the source document. This enables parallel, collaborative editing, a problem that has not received a great deal of

treatment in the literature, but which posed a threat to the timely completion of the Charrette Project. With Figura, editorial assistants use a web-based markup tool at the same time, without worrying about editing collisions, and with the assurance that their figural data are immediately available to everyone else in the project.

Beyond the immediate gains of such an approach for the Charrette Project, Figura introduces a novel approach to humanities computing applications that is worth consideration and development by others. One of the greatest shortcomings of the traditional approach to humanities computing applications is that it relegates the decisive task of indexing to proprietary technologies, such as DynaText or Oracle Text, which are not only inaccessible to the designers of humanities applications, but are designed for commerce and not scholarship. Because the indexing engine has the pivotal role of mediating between source archive and end-user web, a consistent result is that marked up documents end up being "smarter" than their published variants, leaving a wide gap between a project's collection of marked up texts and the documents that most scholars end up seeing. Put in bare economic terms, indexing engines have been grossly inefficient converters of scholarly labor into intellectual product. McGann once described this situation as a "gulf separating a Unix from a Mac world," and advocated the construction of two separate schemes to support each side of what he characterized as a "double helix" (McGann 1997). Figura seeks to eliminate the gap in the first place by inverting the relationship between index and archive—essentially, by treating the index as the text itself.

In our poster presentation we will present the Figura application and describe its architecture, its place in the Charrette Project, and the design philosophy behind the application. We will also try to focus as much attention on how the application is designed as on how it is used by scholars in actual situations of interpretation. Figura is above all a practical application, created to meet directly the needs of scholars; it was not conceived as a formal solution to an abstract problem. Finally, we will also present what we consider to be the principle shortcomings of the application, covering cases where it is most and least suitable, and compare it to related initiatives.

REFERENCES

Alvarado, R. (1999). "Of Media, Data, Documents: An Argument for the Importance of Relational Technology to the Project of Humanities Computing." Annual ACH-ALLC Conference. Charlottesville, Virginia.

Glass, M. and B. D. Eugenio (2002). "MUP: The UIC Standoff Markup Tool." Third SIGDial Workshop on Discourse and Dialogue (SIGDial-02). Philadelphia.

McGann, J. (1997). "Imagining What You Don't Know: The Theoretical Goals of the Rosetti Archive," Institute for Advanced Technology in the Humanities. 2002.

Müller, C. and M. Strube (2001). "MMAX: A Tool for the Annotation of Multi-modal Corpora." 2nd Workshop IJCAI Workshop on Knowledge and Reasoning in Practical Dialogue Systems. Seattle, SIGdial.

Renear, A., E. Mylonas, et al. (1996). "Refining our Notion of What Text Really Is: The Problem of Overlapping Hierarchies." Research in Humanities Computing. N. Ide and S. Hockey. Oxford, UK, Oxford University Press.

Sperberg-McQueen, C. M., L. Burnard, et al. (1994). "Guidelines for electronic text encoding and interchange." Chicago and Oxford: Text Encoding Initiative.

Sperberg-McQueen, C. M. and C. Huitfeldt (1999). "GODDAG: A Data Structure for Overlapping Hierarchies." Annual ACH-ALLC Conference, Charlottesville, Virginia.

Thompson, H. S. and D. McKelvie (1997). "Hyperlink semantics for standoff markup of read-only documents," Language Technology Group, HCRC, University of Edinburgh. 2002.

Uitti, K. D. (1997). "A Brief History of the 'Charrette Project' and Its Basic Rationale." [**http://www.princeton.edu/~lancelot**]. Last accessed 11/25/2002.

Uitti, K. D. and A. Foulet (1989). Le Chevalier de la Charrette, Classiques Garnier.

On the Content Model for <respStmt>: Newer is Not Necessarily Better

SYD BAUMAN
Brown University Women Writers Project
Syd_Bauman@Brown.edu

The TEI P3 declaration for <respStmt> boils down to

```
<!ELEMENT respStmt  - O  ( (resp & name), (resp | name)* ) >
```

Which means that a <respStmt> must contain one <resp> and one <name> in either order, followed by any number of any combination of either. (Remember that the globally included elements are allowed via an inclusion exception on <text>, so they do not need to be mentioned explicitly in this content model.) So the following order of elements would be valid content.

```
(a) <resp>, <name>, <name>, <name>
(b) <name>, <resp>, <resp>, <resp>, <resp>
```

These make a lot of sense to me. In (a) the <respStmt> is listing all those individuals who shared or had a particular responsibility. In (b) it lists an individual and all the hats she wore. The content model also allows for the following.

```
(c) <name>, <resp>, <name>, <resp>, <name>, <resp>
```

This also makes a lot of sense: three people each with his or her responsibility. This would probably be better encoded as a series of three <respStmt>s, each with one <name> and one <resp>. But the content model also allows for

```
(d) <name>, <resp>, <name>, <name>, <resp>, <resp>, <resp>, <resp>,
<resp> <name>, <resp>, <name>
```

What the heck does that mean, I wonder?

The TEI P4 (XML) declaration for <respStmt> boils down to

```
<!ELEMENT respStmt  (resp | name | %m.Incl;)+ >
```

This content model allows for each of the above sequences of <resp> and <name> (and any other imaginable sequence of those two elements that would have been valid against the P3 content model), but also allows the somewhat bizarre

```
(e) <resp>, <resp>, <resp>
```

which I suppose is how you would indicate tasks in your project that never got done, and for which no one is taking responsibility. Even worse, it allows (f) <name> which I suppose is how you would indicate a freeloader whom your organization pays, but does not actually do anything. And most puzzling, as long as you include at least one element from m.Incl (say, <cb> [1]), it allows you to get away with neither a <resp> nor a <name>:

```
(g)
```

which I suppose is how you indicate that your organization is looking to hire a freeloader who won't do anything, but you haven't found a qualified candidate yet.

Seriously, to me (e), (f), and (g) are clearly errors, and if possible it would be nice to catch them in the XML validation stage. Further, I consider (d) an error, although I will not be surprised if there are those who disagree with me. Lastly, I don't mind giving up the capability to use (c) in the quest to be able to exclude (d)–(g).

Thus, in 1998-12, I came up with the following SGML declaration for the WWP; I have included the comment to explain it a bit:

```
<!--
*** The following declaration for RESPSTMT is similar to the
*** one found in teicore2.dtd in that NAME and RESP are the
```

```
*** only children of RESPSTMT, but different in that we only
*** allow either one NAME or one RESP (and multiples of the
*** other), and we insist that NAMEs come before RESPs. A
*** (purposeful) side effect is that, since we don't use an
*** "&" connector, the content model is valid XML. Note
*** that, in principle, the content model is
***    (  ( name, resp+ ) | ( name+, resp )  )
*** but that is ambiguous (i.e., non-deterministic).
-->
<!ELEMENT %n.respStmt;  - -  (
                               %n.name;,
                               (
                                 (%n.resp;)+
                                 |
                                 ( (%n.name;)+, %n.resp; )
                               )
                             )
                                 >
```

With the parameter entity references resolved and without some of the extra whitespace, this boils down to

```
<!ELEMENT respStmt - - ( name, ( resp+ | ( name+, resp ) ) ) >
```

The content model here requires that a <name> be first and that a <resp> be last; it allows any number (including 0) of <name>s or <resp>s, but not both, in between. Thus it allows

```
(h) <name>, <resp>
(i) <name>, <name>, <name>, <name>, <resp>
(j) <name>, <resp>, <resp>, <resp>
```

but excludes (a), and (c)–(g). Thus the only concession I need to make is that (a) must be written as (j), with the <name> first. Not only does this not bother me, it makes a bit of sense (keeping the encoding of <respStmt>s a bit more consistent).

So now to make this new content model into a P4 content model all that's left to figure out where and how to insert the globally included elements.[2] Luckily the TEI has prepared what could be thought of as a how-to guide on this very subject[3] I will not delve into the logic in that paper here; suffice it to say that I think the following is the correct result of the application of the logic outlined in the aforementioned guide.

```
<!ELEMENT respStmt (
                        ( (%m.Incl;)*, (name, (%m.Incl;)*) ),
                                                             (
                                                               ( resp,
(%m.Incl;)*)+
                                                             |
                                                               (
( name, (%m.Incl;)*)+ ,
( resp, (%m.Incl;)*)
                                                                    )
                                                             )
                                                           ) >
```

This will likely be the WWP replacement for the P4 declaration of <respStmt>, and I am seeking input as to whether or not this should be used in P5.

NOTES

1. "Why is a column break allowed in the <teiHeader>?" you ask. I hear you cry. In P3 elements like <cb>, <lb>, and <pb> are (quite reasonably) inclusion exceptions on <text>. That means they are allowed to occur inside <text> or any descendant of <text>. This allows you to record the fact that there was a page break in the middle of a name in your source text: … And thus

```
<lb/>it was decreed in the Councell at <name
type="place">Nice</name>,
<lb/>that the Byshops should assemble twise
<lb/>every yeare. And in the Councel at <name type="place">Car
```

```
<pb n="225"/>
<lb/>thage</name> it was decreed, that the Bysshops ...
```

(modified from WWP TR00439, John Jewel, "An Apology or Answer in Defence of the Church of England, 1564", Bacon, Ann (Cooke), trans.)

while still, because <pb> is not in the content model of <name>, disallowing a <pb> inside a <name> inside a <respStmt> in the <teiHeader>. Because XML does not have inclusion exceptions, the content models of <name> and many other elements like it which could appear both in the <teiHeader> and in <text> need to include the globally included elements. (There are a couple of rare exceptions, like <titleStmt>, which, although it can be a child of <biblFull>, can appear nowhere else inside <text>, and since a <biblFull> could not have a <pb> anymore than a <fileDesc> could, <titleStmt> does not need to include such things. At least this is my understanding; please correct me if I'm wrong.) Thus in P4 a <cb> could indeed occur in the <teiHeader>.

[2.] You may wonder why bother allowing the globally included elements in <respStmt> at all. In most cases <respStmt> occurs in the <teiHeader>, where such elements are not needed; but <respStmt> can be a child of <bibl> inside the text. Note that I'm presuming that such a <respStmt> is being authored, not transcribed. Since it is allowed as a child of <bibl>, one could easily imagine that this more restrictive content model is moot because of the need to encode (From "I'd Like to Teach the World to Tag", from spoofters Julia Flanders & Syd Bauman) as

```
<bibl rend="pre(\() post(\))">From <title rend="pre(&ldquo;)
post(&rdquo;)>I'd Like to Teach the World to Tag</title>
```

Tag", from spoofters Julia Flanders & Syd Bauman) as

from <respStmt><resp>spoofters</resp> <name>Julia Flanders &</name> <name>Syd Bauman</name></bibl>

Tag", from spoofters Julia Flanders & Syd Bauman) as

However, I do not think of this as a good way to encode such a reference, as evidenced by the fact that there is noplace to put that ampersand (it isn't really part of Julia's name, now, is it?).

Tag", from spoofters Julia Flanders & Syd Bauman) as

[3.] Thanks to then TEI editor C. Michael Sperberg-McQueen; see **http://www.tei-c.org/Vault/ED/edw69.sgm**, or, pre-formatted into HTML at **http://www.tei-c.org/Vault/ED/edw69.htm**.

The Austrian Academy Corpus, an Extensive Corpus of German Literature and Language - The AAC Literary Journals Subcorpora

---

HANNO BIBER
Austrian Academy Corpus
hanno.biber@oeaw.ac.at

EVELYN BREITENEDER
Austrian Academy Corpus
evelyn.breiteneder@oeaw.ac.at

KARLHEINZ MOERTH
Austrian Academy Corpus
karlheinz.moerth@oeaw.ac.at

In this paper we would like to present one special subsection of the Austrian Academy Corpus. The AAC is a newly founded institution organized and planned at the Austrian Academy of Sciences in Vienna, in which large scholarly digital resources are being established. We will describe the specific selection and processing of historical literary journals that are to be integrated into the AAC. The AAC Literary Journals Subcorpora will include a considerable number of influential journals and magazines which will be made available electronically and will be analyzed and digitally interrelated by means of XML annotation. The general

concept and the idea of the AAC, its approach and its potential for the studies of texts of various kinds will also be addressed in this paper.

The AAC is a multifunctional digital text corpus with new research possibilities in the fields of linguistics as well as for textual research, in the fields of literary studies, for discourse studies and the like. The AAC, comprising an abundance of different sources, focuses predominantly on German language texts from the last 150 years, but also includes translations from other languages and corresponding source texts. This large electronic text collection contains a great variety of significant texts, historical texts, literary texts, and texts stemming from various cultural and social domains. Apart from literary texts and literary journals, there will be a wide collection of various text types and other text type carriers incorporated, such as newspapers and newspaper articles, advertisement, posters, speeches and other examples from the media and the entertainment domain, the arts, but also from jurisprudence, religion, politics, philosophy and other domains. The corpus approach being developed at the AAC is determined by our conviction that for specific language related research interests only consistent corpora that provide sufficient context can be useful, corpora that supply users with complete and thoroughly described texts. This is particularly the case in the field of scholarly research which focuses towards literary text studies and historical discourse analysis, a field where literary texts from certain historical periods and their specific features will be the main objects of investigation. Corpus-based linguistic and literary text research will be made possible by means of a variety of descriptive elements integrated into the text collections of the AAC. Reliable corpus tools and sound methods of corpus description and analysis have become indispensable means of research. At the AAC, all texts are digitized and annotated by means of XML, in order to facilitate thorough investigations and research into the textual qualities of the corpus holdings. How modern technology and standards can be integrated in traditionally oriented fields of research such as literary studies will be exemplified through the AAC's digitization projects.

At the core of the AAC and as the starting point for the selection of the texts to be integrated there is the satirical magazine “Die Fackel” which was published by Karl Kraus in Vienna from 1899 to 1936. In his abundant satirical and polemical texts published in ”Die Fackel” consisting of 22.586 pages, Karl Kraus developed satirical textual strategies in which the use of quotation, citation, and commentary of others is of major importance. This unique source of German literature offers a starting point not only for the future incorporation of other texts in our corpus but also for further far-reaching and challenging research into a number of linguistic, discursive and intertextual phenomena to be observed. The magazine published by Karl Kraus offers us a unique opportunity to attract the attention of anyone interested in the German language. And the AAC intends to digitally present a wide selection of different sources of scholarly, journalistic and political texts which were of considerable influence between 1848 and 1989.

In the first phase of the corpus build-up we have started with the digitisation and structured integration of texts, among which are several highly influential and notable literary journals and magazines. Literary journals have been integrated into the AAC for several reasons. One reason is their importance form the point of view of literary history. Literary journals have been neglected in the studies of literary history although they form an integral part of the literary life. Very often they are the place where the first publications of major literary works appear. They are in a mediator position within the literary market and have been so especially in the 19th and 20th centuries not only in the German speaking countries. Secondly, literary journals have been chosen as a considerable subcorpus of the AAC because in most cases literary journals offer a wide variety of interesting text types. Depending on their general and overall political and ideological orientations they offer multiple sources and sources of various kinds which are of particular interest for corpus research both from theoretical and methodological perspectives. The differences and relations between the various literary text types pose challenging questions concerning the description and the development of digital scholarly material.

Among the source materials for the AAC literary journals subcorpora is the well known literary and political journal “Die Weltbuehne” (The World Stage), previously published under the title “Die Schaubuehne” (The Theatrical Stage) since 1905, and two journals that were of major importance for the expressionist movement, “Die Aktion” and “Der Sturm”, all of them published in Berlin in the first decades of the last century. The journal “Der Brenner” was published in Innsbruck in Austria between 1910 and 1954 and is another example from the AAC literary journal subcorpora and was like the ones mentioned above also influenced by “Die Fackel”. In our paper we will briefly describe the interrelations between these examples from the perspective of literary history, and thereby concentrate on the corpus methodology of applying digital techniques for the description of the processes and qualities observed. We will explain the complex digitization processes and the necessary annotation schemes utilized by the AAC working group in order to access the data contained and to be explored in these valuable resources. In addition to that the new editing strategies being developed by the AAC as well as the scholarly commentaries will be presented. The literary journal subcorpora can be regarded as an experimental and exemplary cohesive subsection of the Austrian Academy Corpus.

Teaching literature through the net: an answer to the caos or the construction of the self

LAURA BORRÀS
Universitat Oberta de Catalunya
lborras@uoc.edu

JOAN-ELIES ADELL
Universitat Oberta de Catalunya
jadellp@uoc.edu

ISABEL MOLL
Universitat Oberta de Catalunya
imoll@uoc.edu

INTRODUCTION. UNIVERSITAT OBERTA DE CATALUNYA: THE FIRST VIRTUAL UNIVERSITY IN THE WORLD

In 1995, the Universitat Oberta de Catalunya (UOC) was conceived as the first e-learning university in the world. This new concept of university redefined the idea of teaching, studying and doing research. There has been an evolution within the UOC since new technologies have developed and have become so usual for our everyday life. There has been as well an important feedback between students and lecturers and this has obviously contributed to find our best way to teach Humanities and Philology in this environment, which has been so often considered hostile for these areas of knowledge. In this paper, we would like to focus on the challenging experience of teaching literature through the net and we will present all the questions we (the lecturers) have had to rise in order to present an effective method of teaching literature in the broad sense of both words: teaching and literature.

FROM THE AESTHETICS OF RECEPTION TO THE AESTHETICS OF INTERACTIVITY

The incorporation of new technologies to literature and to the study of literature has allowed to reread the past from a platform in which the literary work is more than ever an open device, with no hierarchical structures, and belonging to an infinite net of hypertexts. We can rethink the literary phenomenon from other textual, critical and hermeneutical perspective. Digital environement clearly shows the inscription of a literary work in an intertextual framework and allows physical connection of texts. Moreover, digital context reformulates higher education, following a process which begins with the invention of writing, and continues wiht the press, and that frees the student from physically attending to the lectures (Landow, 1997: 29-30). In the Universitat Oberta de Catalunya (www.uoc.edu), since 1995, this process is taking place in a concrete place in the net: the virtual campus, that works asincronically, serving to the students and professors comunity. And even in this heterogeneous and disturbing medium, the final purpose is still "teaching". However, when the environment is not the "clasroom" anymore, but the virtuality of a computer screen, the methodological approach is crucial and demands a previous reflection: at the begining of the 21st century, should we continue to offer the same teaching patterns as we have been doing for many decades? In our opinion, this question applies not only to a pioneer university such as the UOC, but also to more traditional universities. Cultural and social changes that Internet has provoked and the information society in which we are living now compels us to reflect on some premises and to consider new ways and different possibilities in tune with the new era. In the context of overinformation surrounding us, professors" duty can not be anymore the transmission of knowledge, but the contribution to the generation of this knowlegde, that is to say, s/he has to assume a role of "guide" and should provide the necessary intellectual tools that will allow students to be able to think critically.

Focusing on literary studies, our mission is basically teaching to read, but giving to this formula all the depth we are able to. In the Humanities and Philology Studies in the UOC, we have chosen an

hermeneutical approach, following Gadamer, whose idea of Hermeneutics is the art of understanding the other"s opinion. For us, this constitutes a redoubtable reflection on the activity of reading, which aspires to participate in the "shared sense". Nevertheless, reading and teaching to read in a medium like Internet and with its natural tool, this is hypertext, means to confront a technological device serving education. Taking as an example the subject "Temes de literatura universal" ("Universal literary topics"), we will try to outline which are the changes that affect literature teaching. We consider that this constitutes an initial step in order to establish which are the real possibilities that allow us to progress, from a theoretical point of view, from the aesthetics of reception to the aesthetics of interactivity. This will provoke a voice game in the virtual classroom not only between professor and students, but also between students themselves, and therefore there will be a collective and dialogical creation of knowledge.

We have organized this subject taking as a starting point a series of texts connected by the fact that they deal with some of the basic topics of universal literature. In the hipertextual material, which combines linear reading with sequential or fragmentary reading inherent to hypertext, as well as video or audio resources, what we offer are interpretations and readings of key literary works through a double intertextual route. That is to say in a genetical and analogical way, and that relates different literary, pictorial or cinematographic texts. During all this process, we have had the intention of analyse the text from a double perspective, both towards the present and the future but, at the same time, it appears within a pardigm formed by other texts, that precede and influence it. Therefore, the texts assums in a explicit or implicit way the cultural past which comes before, even when it has been conceived against this past. The course intends to contribute to a tranversal reading of universal literature in a virtual environment of learning, and at the same time it provides guide lines to students for a practical exercise of compartive literature. It suggests as well reading itineraries crossing periods and literary traditions which are far from each other in time and space. Following intertextual connections (in an analogical and genetical way), that are guiding us inside the hyprtextual corpus, some reading routes are described as random and subjective and they show the broadcast and the arituclation of such a topic in the literary tradition. Students, after familiarizing with the texts presented in the materials and wiht the navigating tools offered by the materials, have to select the topics that are subjectively more relevant, and to build up their own (hyper)textual corpus. Professor R. Pinto, who is one of the lecturers of the subject, asserts that both during the first part of the course, more receptive, and during the second one, more active, there is always a dialogue between the professor and the students.

TEACHING LITERATURE THROUGH THE NET: TO READ AND TO REREAD

"A good reader is a rereader", says categorically Nabokov. The power of this assertion proves that the hypertextual space, in which the reading is fragmentary and non-sequential, the act of rereading is the best methodological strategy for building up a sense. Hypertext allows an exponential rereading of the contents we have designed. Moreover, hypertext is, without any doubts, an essential tool for on-line literary theory and comparative literature teaching. Therefore, it is necessary to seriously consider the fact of teaching in a virtual context, with these digital materials, in a multiple textual environment, without boundaries, or with the sole boundaries of curiosity and desire. And the use of the World Wide Web, that implies getting rid of the acquired habits and changing the comunication techniques of knwoledge discourse. The ways of proving the "validity" of a literary analysis are being modified since we can deveolp discourse following a logic that is not necessarily linear and deductive, but open and relational. In this sense, after the accumulated experience and taking into account the results obtained from our own strategical and creative versions, the conclusion is that the revolution of production patterns, tranmission and inquiry of the texts can be included in hypertextuality as an epistemological mutation.

UNESTABLENESS? CAOS? TOWARDS THE COLLECTIVE CONSTRUCTION OF KNOWLEDGE

We have to add that the extreme fluidity of hypertext oblige us to rethink one of the main preoccupations of who is writing: the possibility of exerting a control over the way of the reader"s reading. Indeed the author-professor creative act requires an interpretative act of the user-student and requires as well a wander around the text. Hypertext shows a new form of "textuality" based on the capacity of "penetration" of a text marked with all the links that open doors to new sense horizons. In hypertexts, any illusion of control vanishes: seduction is the only motivation towards an hypertextual wander. We lose the notion of control because hypertext itself undervalues it. If knowledge spreads itself provoking an infinite virtuality of intertextual connections, that represent the inifite ways of the discoursive configuration of the self (Pinto, 2002: 175), everything is relevant, everything could be interconnected. Our own proposal for this subject, was, therefore, to promote a digital working environment that made evident the literary work in a more physical way, completely textual, and that allowed interaction, interrelation and link. This personal use of

phylology, defined by Pinto as more attentive to the subject who interprets and to the questions that constitute him/herself than to the interpreted text and to its objective historical reality, can be considered as a copernican revolution of literary criticism, promoted by new technologies.

REFERENCES

Landow, Geroge P. (ed.) (1994), Teoría del hipertexto, Paidós Barcelona, 1997

Pinto, Raffaele, "Crítica literaria y construcción del sujeto. Dos modelos de autoanálisis: Sigmund Freud (psicopatología de la vida cotidiana) e Italo Calvino (seis propuestas para el próximo milenio)", en Laura Borràs (ed.) (2002), Deseo, construcción y personaje, Fundación Autor, Madrid, (p.173-193)

Developing a Toolkit for Digital Epigraphy

HUGH CAYLESS
UNC Chapel Hill
hcayless@email.unc.edu

This paper will present the results of work being done at UNC Chapel Hill to support the emerging EpiDoc standard for encoding epigraphic texts. The EpiDoc standard is based on the Text Encoding Initiative guidelines (http://www.tei-c.org). Information on this collaborative effort, which includes scholars and humanities computing experts from North America and Europe, is available at http://www.unc.edu/awmc/epidoc. The mission of the project is to develop "a software and hardware-independent interchange specification for scholarly and educational editions of inscribed and incised texts in Greek, Latin and other languages emanating from the ancient Greek, Roman and nearby civilizations." The Chapel Hill team has focused on building tools to facilitate epigraphers' work on creating EpiDoc texts and managing the storage and presentation of those texts. The paper will begin with a brief outline of the EpiDoc guidelines with examples of their implementation and a discussion of some of the problems inherent in digital epigraphy. Then the various tools developed by the Chapel Hill team will be demonstrated.

The editing and publication of inscribed documents has a rich history, which includes the adoption of widely accepted standards for "marking up" texts. Published Greek and Latin epigraphic texts are almost universally presented according to the "Leiden convention,"developed in 1931 at the 18th International Congress of Orientalists by papyrologists. The Leiden convention employs various symbols to present information about the condition of the inscribed text and editorial supplements and comments. For example, parentheses surround editorial expansions of abbreviated words, so "AVG" on the support would be expanded as "Aug(ustus)" by the editor. In EpiDoc, the same text would be written "Aug<expan<ustus</expan>." Because of the widespread adoption of the Leiden convention, the form of published texts (allowing for some local variations) is remarkably consistent. Because of this consistency, it is possible to write a parser that can translate a digitized inscription into a data structure (such as XML). The Chapel Hill team has developed such a parser, named the Chapel Hill Epigraphic Text Converter (CHET-C). This tool parses texts marked up according to the Leiden convention and outputs EpiDoc XML. Both MS Access and Java versions are being developed.

Digital texts in polytonic Greek pose problems of their own because of the various methods which evolved to encode them prior to the advent of Unicode. One of the earliest of these was Beta Code, which utilizes 7-bit ASCII to represent Greek, Latin, and Aramaic texts, as well as various epigraphical and papyrological sigla. In addition, a number of Greek fonts, each employing its own specialized encoding, were developed over the years to handle the problem of typing and viewing Greek with computers. There exist already various programs and filters which can handle shifting the encoding of texts from one form to another (the Greek texts on http://www.perseus.tufts.edu employ such filters, for example). There was still a need for a more flexible system to handle encoding shifts, however, and so the Chapel Hill team (in collaboration with the Stoa Consortium, http://www.stoa.org) has developed Java-based software that performs this function and which can easily be plugged into the larger toolkit framework. The transcoder classes can perform encoding shifts on text within specific elements of an XML file, and are also designed to be modular and easily extensible to handle languages other than Ancient Greek.

The team has also developed a web-based framework for the presentation of EpiDoc texts. The framework has been built with Apache Cocoon (http://xml.apache.org/cocoon) and is hosted by the Stoa. The "Epidocinator" as it is called, can dynamically transform EpiDoc documents against a variety of XSL

stylesheets. Both documents and stylesheets may be either on site or remote. The framework will also perform validation and error reporting on EpiDoc texts. The transcoder classes will be accessible via a Cocoon Transformer, so that Greek text can automatically be shifted from any supported source format to any supported result format. It will employ the Java version of CHET-C in a custom Cocoon Generator or Transformer to allow users to generate EpiDocs from already-digitized inscriptions. The framework will be packaged as a Web Archive (WAR), so that interested developers can easily deploy their own digital epigraphy frameworks.

Finally, the team is also working to integrate the Java tools with jEdit, a Java-based text editor with a very robust plugin architecture (http://www.jedit.org). The goal is essentially to create an Integrated Development Environment (IDE) for epigraphers which editors will be able to use in the electronic publication of EpiDoc texts. All of the tools and stylesheets developed for use with EpiDoc will be available as jEdit plugins and editors will also have available jEdit's XML and XSLT plugins, which provide useful functions like code-completion, tag creation help (based on a DTD), validation, error reporting, and transformation. The combination of the web framework and the text editor will provide epigraphers with a cross-platform, Open Source solution for creating and publishing inscriptions.

This paper will present the results of work being done at UNC Chapel Hill to support the emerging EpiDoc standard for encoding epigraphic texts. The EpiDoc standard is based on the Text Encoding Initiative guidelines (http://www.tei-c.org). Information on this collaborative effort, which includes scholars and humanities computing experts from North America and Europe, is available at http://www.unc.edu/awmc/epidoc. The mission of the project is to develop "a software and hardware-independent interchange specification for scholarly and educational editions of inscribed and incised texts in Greek, Latin and other languages emanating from the ancient Greek, Roman and nearby civilizations." The Chapel Hill team has focused on building tools to facilitate epigraphers' work on creating EpiDoc texts and managing the storage and presentation of those texts. The paper will begin with a brief outline of the EpiDoc guidelines with examples of their implementation and a discussion of some of the problems inherent in digital epigraphy. Then the various tools developed by the Chapel Hill team will be demonstrated.

The editing and publication of inscribed documents has a rich history, which includes the adoption of widely accepted standards for "marking up" texts. Published Greek and Latin epigraphic texts are almost universally presented according to the "Leiden convention,"developed in 1931 at the 18th International Congress of Orientalists by papyrologists. The Leiden convention employs various symbols to present information about the condition of the inscribed text and editorial supplements and comments. For example, parentheses surround editorial expansions of abbreviated words, so "AVG" on the support would be expanded as "Aug(ustus)" by the editor. In EpiDoc, the same text would be written "Aug<expan>ustus</expan>." Because of the widespread adoption of the Leiden convention, the form of published texts (allowing for some local variations) is remarkably consistent. Because of this consistency, it is possible to write a parser that can translate a digitized inscription into a data structure (such as XML). The Chapel Hill team has developed such a parser, named the Chapel Hill Epigraphic Text Converter (CHET-C). This tool parses texts marked up according to the Leiden convention and outputs EpiDoc XML. Both MS Access and Java versions are being developed.

Digital texts in polytonic Greek pose problems of their own because of the various methods which evolved to encode them prior to the advent of Unicode. One of the earliest of these was Beta Code, which utilizes 7-bit ASCII to represent Greek, Latin, and Aramaic texts, as well as various epigraphical and papyrological sigla. In addition, a number of Greek fonts, each employing its own specialized encoding, were developed over the years to handle the problem of typing and viewing Greek with computers. There exist already various programs and filters which can handle shifting the encoding of texts from one form to another (the Greek texts on http://www.perseus.tufts.edu employ such filters, for example). There was still a need for a more flexible system to handle encoding shifts, however, and so the Chapel Hill team (in collaboration with the Stoa Consortium, http://www.stoa.org) has developed Java-based software that performs this function and which can easily be plugged into the larger toolkit framework. The transcoder classes can perform encoding shifts on text within specific elements of an XML file, and are also designed to be modular and easily extensible to handle languages other than Ancient Greek.

The team has also developed a web-based framework for the presentation of EpiDoc texts. The framework has been built with Apache Cocoon (http://xml.apache.org/cocoon) and is hosted by the Stoa. The "Epidocinator" as it is called, can dynamically transform EpiDoc documents against a variety of XSL stylesheets. Both documents and stylesheets may be either on site or remote. The framework will also perform validation and error reporting on EpiDoc texts. The transcoder classes will be accessible via a Cocoon Transformer, so that Greek text can automatically be shifted from any supported source format to any supported result format. It will employ the Java version of CHET-C in a custom Cocoon Generator or Transformer to allow users to generate EpiDocs from already-digitized inscriptions. The framework will be

packaged as a Web Archive (WAR), so that interested developers can easily deploy their own digital epigraphy frameworks.

Finally, the team is also working to integrate the Java tools with jEdit, a Java-based text editor with a very robust plugin architecture (http://www.jedit.org). The goal is essentially to create an Integrated Development Environment (IDE) for epigraphers which editors will be able to use in the electronic publication of EpiDoc texts. All of the tools and stylesheets developed for use with EpiDoc will be available as jEdit plugins and editors will also have available jEdit's XML and XSLT plugins, which provide useful functions like code-completion, tag creation help (based on a DTD), validation, error reporting, and transformation. The combination of the web framework and the text editor will provide epigraphers with a cross-platform, Open Source solution for creating and publishing inscriptions.

All of the tools described above will be demonstrated briefly, using texts marked up by various projects that have adopted the new standard. Then I will conclude by discussing how the EpiDoc Collaborative plans to proceed in developing the standard and the supporting tools. The project's use of open standards and Open Source development tools provides a useful model for similar types of scholarly publication, one which could be picked up and extended for other purposes without a great deal of effort.

Digital texts in Greek pose problems of their own because of the various methods which evolved to encode them prior to the advent of Unicode. One of the earliest of these was Beta Code, which utilizes 7-bit ASCII to represent Greek, Latin, and Aramaic texts, as well as various epigraphical and papyrological sigla. In addition, a number of Greek fonts, each employing its own specialized encoding, were developed over the years to handle the problem of typing and viewing Greek with computers. There exist already various programs and filters which can handle shifting the encoding of texts from one form to another (the Greek texts on http://www.perseus.tufts.edu employ such filters, for example). There was still a need for a more flexible system to handle encoding shifts, however, and so the Chapel Hill team (in collaboration with the Stoa Consortium, http://www.stoa.org) has developed software in Java that performs this function and which can easily be plugged into the larger toolkit framework. The transcoder classes can perform encoding shifts on text within specific elements of an XML file, and are also designed to be modular and easily extensible to handle languages other than Ancient Greek.

The team has also developed a web-based framework for the presentation of EpiDoc texts. The framework has been built with Apache Cocoon (http://xml.apache.org/cocoon) and is hosted by the Stoa. The "Epidocinator" as it is called, can dynamically transform EpiDoc documents against a variety of XSL stylesheets. Both documents and stylesheets may be either on site or remote. The framework will also perform validation and error reporting on EpiDoc texts. The transcoder classes will be accessible via a Cocoon Transformer, so that Greek text can automatically be shifted from any supported source format to any supported result format. It will employ the Java version of CHET-C in a custom Cocoon Generator or Transformer to allow users to generate EpiDocs from already-digitized inscriptions. The framework will be packaged as a Web Archive (WAR), so that interested developers can easily deploy their own digital epigraphy frameworks.

Finally, the team is also working to integrate the Java tools with jEdit, a Java-based text editor with a very robust plugin architecture (http://www.jedit.org). The goal is essentially to create an Integrated Development Environment (IDE) for epigraphers which editors will be able to use in the electronic publication of EpiDoc texts. All of the tools and stylesheets developed for use with EpiDoc will be available as jEdit plugins and editors will also have available jEdit's XML and XSLT plugins, which provide useful functions like code-completion, tag creation help (based on a DTD), validation, error reporting, and transformation. The combination of the web framework and the text editor will provide epigraphers with a cross-platform, Open Source solution for creating and publishing inscriptions.

All of the tools described above will be demonstrated briefly, using texts marked up by various projects that have adopted the new standard. Then I will conclude by discussing how the EpiDoc Collaborative plans to proceed in developing the standard and the supporting tools. The project's use of open standards and Open Source development tools provides a useful model for similar types of scholarly publication, one which could be picked up and extended for other purposes without a great deal of effort.

Orlando on the Web: From Development System to Web-based Delivery of a Content-Encoded Textbase

PATRICIA CLEMENTS
University of Alberta
Patricia.Clements@ualberta.ca

RENÉE ELIO
University of Alberta
ree@cs.ualberta.ca

SHARON BALAZS
University of Alberta
sbalazs@ualberta.ca

SUSAN BROWN
University of Guelph
sbrown@uoguelph.ca

ISOBEL GRUNDY
University of Alberta, with members of the Orlando Project
Isobel.Grundy@ualberta.ca

INTRODUCTION

The Orlando Project at the Universities of Alberta and Guelph aims to produce the first full scholarly account of women's writing in the British Isles in a mode of literary history designed to take advantage of new technological capabilities. To enable researchers to discover the sophisticated and nuanced interconnections among this complex mass of material, Orlando has produced a custom-designed SGML text encoding system capable of reflecting literary and historical interpretation.

Aspects of this encoding scheme have been discussed in other forums[1, 2] and the ways in which the encoding scheme has, as planned, turned out to be effective in supporting the identification of novel and significant interrelationships in literary history has been presented[3, 4, 5]. During its development, the use of the Orlando textbase and its encoding scheme to explore such relationships has been limited to the immediate research team, using the in-house Orlando development tools. But the vision of the Orlando project is, of course, to present its encoded textbase to the larger research community; i.e. to create a web-based delivery system that researchers world-wide can use to explore Orlando content. This presentation will take a systems view of the Orlando Project, focusing on issues and methods that have arisen in producing the first version of this delivery system.

The three key aspects of this systems view, which will be outlined here, are: (a) the in-house Orlando development environment, its tools, and the decisions made in crafting this environment for the construction of the textbase; (b) the definition and design of a prototype system to achieve certain core, albeit limited, functionalities without precluding the later design and implementation of a more powerful system; (c) general issues in designing an interface that will lead novice users step by step to exploiting at first a manageable selection from an integrated set of literary historical materials and complex underlying encoding scheme, and so by stages to fuller use of the potentialities of such a scheme. A demonstration of the first version delivery system will both clarify and concretise the issues discussed in the presentation.

IN-HOUSE SGML DEVELOPMENT ENVIRONMENT

In brief, the development environment supports the creation and ongoing revision of the three components to the Orlando system. The first of these is a textbase containing SGML documents for the biography and writing career of each individual writer, and for historical topics and issues. The biographical documents contain newly researched material on the lives, backgrounds, and activities of women writers. The writing documents contain newly researched material on literary careers, and the production, textual features, and

reception of texts. The topic documents contain newly researched material under a range of headings deemed crucial for coverage. The second component to the Orlando system is an Oracle database which holds information on historical context: brief accounts of historical events and processes chosen to reflect each period's literary, cultural, and social concerns, anchored to a date or date-range. The goal of this material is to enable a user to produce a chronology—an ordered set of events—relating to any particular time period, writer, word or concept. Sorting by event-type or level of priority is an obvious use of this database. The third component of the Orlando system is another Oracle database, which holds full bibliographic details of all primary texts and secondary sources referred to in the electronic text.

The investment in SGML necessitated using SGML-compatible software such as Oracle in the development environment. The way in which this development environment has in turn affected the development of the delivery system will be outlined in a detailed chart.

DEFINING AND IMPLEMENTING A PROTOTYPE DELIVERY SYSTEM

The first Orlando delivery system had several goals. The primary goal was to provide web-based access to most of the core Orlando materials (writing documents, biographical documents, and events) to create chronologies and to provide the first automated hyperlinks.

The SGML encoding scheme includes a number of tags by which materials in each of these separate databases are related. For example, a <BIBCIT> (bibliographic citation) tag that appears in a writing, biography, topic or event prose refers, by means of unique identifiers of records in the Oracle database, to complete bibliographic material. Also in development is an SGML-tagged name authority list, which not only ensures accurate hyperlinking (via the <NAME> tag) across the various databases, but which also aids in search and retrieval.

The prototype delivery system allows a user to access the Orlando material in three primary ways: through a writer's name, through event chronologies, and through several thematic entry points. Given a writer's name, the writer's biographical and writing documents are retrieved, and a chronology of events mentioning that writer's name is computed and displayed. Certain "core tags" (name, place, date, title, organization name) act as hyperlinks to other areas of the textbase. Users can also access Orlando material through event chronologies generated by means of a freetext search on a given word/phrase, or a search by tagged name, genre, title, place, or organization name. (For these categories users are free to type in a search term or select from a list of possibilities). Users can also specify a given time range, and/or specify particular event types and/or priority levels. Thematic entry points are a means of entering into Orlando by way of material that cuts across the textbase in ways that highlight certain themes. "People" allows users to find people by name, historical period, occupation, or what they wrote"; "Texts" allows users to discover texts by title, subjects, or types of writing"; "Contexts" allows users to search on various topics, organizations, and places in women's literary history; "Networks" allows users to investigate literary, social or family connections, organizational links or intertextual relations; and "Identities and Politics" allows users to investigate cultural and political issues.

The full power of the Orlando system, of course, comes with sophisticated use of its interpretative tags for information exploration, i.e., those tags such as <INTERTEXTUALITY>, <CULTURAL FORMATION>, <RELATIONS WITH PUBLISHER>, and <POLITICS>, which the authors of documents have used to mark up text for literary historical interpretation. However, for the first delivery system, issues taking precedence were those related to working with extant web browsers and XML software, those related to automating the transformation of SGML into XML for web delivery as materials are moved from the development to the delivery environment, and those concerning a first pass at a user interface to what will ultimately be a powerful information exploration system (but that will always need simpler ways in to be available). Orlando has taken a modular approach to the definition, design, and implementation of its delivery system and hence this first version did not fully exploit these interpretative tags.

ON USER INTERFACE ISSUES FOR INFORMATION EXPLORATION

In the first version of a web-based delivery system, we aimed to follow good user interface design principles. This involved, firstly, developing a powerful yet usable interface that would allow both novice and expert users to access Orlando materials. (For example, a novice user may choose to enter directly through a writer's name, whereas an expert user may choose to enter through a more complex chronological search.) Secondly, this involved revealing to the user a portion of the underlying interpretive scheme to allow some exploitation of it. (For example, thematic entry points reveal to the user some of the complexity of the Orlando interpretive scheme.) Thirdly, this involved arriving at design principles and choices to create a coherent display of the material. The paper will outline the ways in which this process served as a foundation for the next stage of

delivery work, currently underway, which is focused on representing the interpretative markup, and ensuring that the interface allows for maximum exploitation of it. By conference time this next phase of delivery work will be completed in its first instantiation and ready for demonstration.

CONCLUSION

This poster presentation will introduce the first version of the Orlando delivery system, discuss in detail the progress that was made in the creation of this system, and outline the ways in which it provided a basis for subsequent delivery development.

REFERENCES

Brown, S., Fisher, S., Clements, P., Binhammer, K., Butler, T., Carter, K., Grundy, I., & Hockey, S. (1998). "SGML and the Orlando Project: Descriptive Markup for an Electronic History of Women's Writing". *Computers and the Humanities*. 31, 271–85.

Butler, T., Fisher, S., Hockey, S., Coulombe, G., Clements, P., Brown, S., Grundy, I., Carter, K., Harvey, K., Wood, J. (2000). "Can a Team Tag Consistently? Experiences on the Orlando Project". *Markup Languages Theory and Practice*, 2, 111–125.

Grundy, I., Clements, P., Brown, S., Butler, T., Cameron, R., Coulombe, G., Fisher, S., & Wood, J. (2000). "Dates and ChronStructs: Dynamic Chronology in the Orlando Project". *Literary and Linguistic Computing*, 15, 265–289.

Brown, S., Grundy, I., et al. (2001). "Text and Intertext in Electronic Documents." *Annual ALLC/ACH Conference*, New York, June 2001

Brown, S., Grundy, I., et al. (2001). Session of three Orlando Project papers: "The Hard and the Soft: Encoding Literary History," "Risking E-Race-Sure/Erasure: Encoding Cultural Formations," and "The Anxiety of Encoding: Intertextuality and Feminist Literary History." *Annual Digital Research in the Humanities Conference*, School of African and Oriental Studies, London University, UK, 9 July 2001

Towards an Electronic Esposizioni: Code as Commentary

CRISTIANA FORDYCE
Brown University
cristiana_fordyce@brown.edu

VIKA ZAFRIN
Brown University
zafrin@brown.edu

In 1374, the city of Florence awarded Giovanni Boccaccio the honor and responsibility of presenting to a civic audience Dante's Divine Comedy, and of commenting on it. The author of the Decameron held a series of lectures on the Comedy at the Church of Santo Stefano. He doubled as commentator and preacher: the lectures were intended as both literary and moral education. Boccaccio was engaged to elucidate all one hundred cantos of the Divine Comedy; unfortunately, he died long before he could complete this task, and the readings are abandoned at Canto 17 of the Inferno.

The text of Boccaccio's lectures has survived to our days as the *Esposizioni sopra la Comedia*. Despite only covering the first seventeen cantos, in print, it is over 700 pages long.[1] The text is comprised of an Accessus (introduction) followed by literal and allegorical exposition on each of the first cantos of the Inferno. (The tenth canto contains no allegorical exposition, as Boccaccio claims that it contains no allegorical import.) The Esposizioni therefore functions both as lecture series and as encyclopedia. From the beginning, Boccaccio keeps a prudent distance from theological and intellectual exposition.

Most interested in revealing the effectiveness and clarity of the Dantean text, Boccaccio strives to cater to the needs of his listeners. He accomplishes this through maintaining the traditional rhetorical form of the expositio, while focusing on the utility of the Comedy for the audience he is to instruct. Boccaccio employs simple and effective means of delivery, in an attempt to evoke in his audience images and exempla from collective memory and individual experience. Despite his awareness of the great responsibility with

which he was charged, Boccaccio seems to have felt indebted neither to the rhetorical tradition nor to moral instruction in his effort to produce the utilitas of comprehension of the magnitude of Dante's work.

Today, we have chosen to offer Dante's text online as commented and presented by one of the most prominent medieval literary critics. We aim at employing the spirit that inspired Boccaccio in his lectures. To this end, we wish to recreate for the modern reader the same utility and efficiency for which Boccaccio's lessons strove with regard to his own audience. The exigencies of our contemporaries when reading a medieval text are not the same as those of the people who gathered in Santo Stefano more than six hundred years ago. The present audience needs to be introduced not only to Dante but to the language and rhetorical structures that Boccaccio employed to benefit the public of his time. To understand the Esposizioni, it is first necessary to understand the audience to whom the lectures were addressed in the first place.

This project is in the beginning stages of development. Our ultimate aim is to map the relationships between our various primary and secondary source texts – the text of the Esposizioni itself, that of the Divine Comedy, and the writing of the many other authors whom Boccaccio quotes. We will be using XML to encode essential information about the people and places mentioned in the Esposizioni, paying particular attention to information that the medieval lecture attendee would have taken for granted, which may not be quite as obvious to the modern reader. In the course of his exposition, Boccaccio frequently alludes to texts by classical and medieval authors (such as Virgil and Livy); we plan to make available the relevant passages from those texts alongside the principal text. For example, it will be useful to provide context and detail about the external texts Boccaccio quotes, as well as historical and biographical information regarding the authors upon whose work he bases his arguments. Where other types of annotation are necessary which will bring our audience closer to Boccaccio's writing, thereby making it more useful in understanding the Comedy, they will also be inserted.

In this paper, we would like to address some issues which have arisen in the process of our thinking about how to approach such a project electronically. The Esposizioni is a fascinating, multilayered medieval text. We treat it as both a commentary and expositio on Dante, and as a stand-alone, authoritative text. We view our electronic publication in the same dichotomous fashion; in addition to presenting the text of the commentary, we provide our own commentary in the form of the code itself. It is essentially a divulgative commentary: our purpose is to bring to our readers enough supplementary information to Boccaccio's text for it to be as useful to them as it was to his medieval audience. Of course, we cannot hope to replicate the experience of the original audience of this text, but with an awareness of the inter-relations between author, speaker, audience, and reader, we hope to create a multivalent resource which will challenge and inform in a similar way to that envisaged by Boccaccio in his lectures.

As our starting point, we embrace Michel Meyer's view of rhetoric as a method of negotiation between two poles – the listener (reader) on one hand, and the writer on the other. We follow Meyer's lead to consider our role in encoding and presenting the Esposizioni as a formal negotiation between intended audiences. The markup of the text, as Lou Burnard has pointed out, rests on the importance of "a single encoding scheme, a unified semiotic system... a single formalism [by using which] we reduce the complexity inherent in representing the interconnectedness of all aspects of our hermeneutic analysis, and thus facilitate a polyvalent analysis" (Burnard 1998). Although Burnard was speaking about mark-up in general, his point is especially valid for a complex electronic medieval text such as the Esposizioni.

What we do by commenting through markup is not new, so why do all this with the aid of a computer? The answer to this is inextricably tied to the principal difference between the potential audience for our project, and Boccaccio's. There are no specific records for those who attended Boccaccio’s expository lectures, but we can assume that a large part of the audience was literate at least in the Florentine vernacular, and possibly in Latin as well. The prestigious lectures would have attracted academic and clerical professionals, scholars and students, as well as interested members of the lay community. There are many differences between Boccaccio’s audience and the modern-day reader of the Esposizioni. First and foremost, we are no longer a memory-based culture, and the modern reader tends not to possess the mnemonic resources which would have enabled Boccaccio’s original audience to recognize his explicit and implied intertextual allusions. The amount of information available to us nowadays is so vast that we are a research-based learning culture, much more heavily dependent on libraries and other information depositories. In addition, the modern reader may have had relatively little exposure to the key texts for the medieval scholar, such as the major classical and medieval authors, and even the Bible. Finally, even if the modern reader has the linguistic ability to engage with an Italian-language text, they may not have the facility to read and understand the many Latin citations in this work.

All of the above problems can be resolved through a sympathetic and accurate annotation of the electronic text. Electronic data storage space is our only choice for the easiest and most coherent presentation of the amount of information we must convey in order to bring our audience closer to Boccaccio's. Furthermore, by putting the Esposizioni online, we offer the option of simplifying the commentary for the electronic user, by offering various different ways into the text. Boccaccio's text often seems chaotic in nature

to the modern reader: it is endlessly self-referential, sometimes self-contradictory, and at other times simply incorrect. (We should not of course forget that the text we have is a series of notes intended for oral delivery and possible further explication.) By a semantic encoding of each of the themes and subjects treated within this dauntingly linear exposition, the user is no longer constrained to follow Boccaccio's digressive reasoning. This opportunity to encode a series of short comments on varied subjects and make them semantically searchable is a great advantage of the medium and will be of great value to future users..

Electronic publication and dissemination of this medieval text does not restore it to its original meaning, but it does restore it to its original purpose: to take a text out of the hands of the elite and bring it closer to all members of the interested public. The electronic medium is thus, paradoxically, a simpler, more straightforward one than print, for our purposes. Paper-based publication of such a tightly interwoven, heavily annotated text could in theory convey all this information, but would do so in a much more awkward fashion. Instead, we will publish it electronically, making it easier both for us to comment, and for our readers to apprehend. The computer is our "wooden key,"; we "wish nothing but to open what is closed," and will use the simplest tool for the task.[2] We will thus de-mystify the Esposizioni for our contemporaries, as Boccaccio in his time de-mystified the Divine Comedy.

NOTES

[1.] In our project, we follow the text as established in the most recent critical edition, edited by Giorgio Padoan, Esposizioni sopra la Comedia (Milan: Mondadori, 1965).

[2.] "It is a noteworthy quality to love the truth in the words, not the words themselves. For what use is a golden key if it cannot unlock what we desire? And what is wrong with a wooden key, if it can unlock what we desire, when we wish nothing but to open what is closed?". Peter Abelard, Prologue to Sic et non, in the online Medieval Sourcebook: **http://www.fordham.edu/halsall/source/Abelard-SicetNon-Prologue.html**

REFERENCES

Abelard, Peter., Prologue to Sic et non. In: Boyer, Blanche B., and Richard McKeon, eds. Sic et Non: a critical edition. Chicago: University of Chicago Press, 1976-1977.

Burnard, Lou. "On the hermeneutic implications of text encoding." In: Fiormonte, Domenico, and Jonathan Usher, eds. New Media and the Humanities: Research and Applications. Proceedings of the first seminar "Computers, literature and philology." Edinburgh, 7-9 September 1998. Oxford: Humanities Computing Unit, University of Oxford, 2001.

Carruthers Mary J. The book of memory: a study of memory in medieval culture. Cambridge: Cambridge University Press, 1990.

Meyer, Michel. Rhetoric, language, and reason. University Park: Pennsylvania State University Press, 1994.

The Development of the Poetry Portal at the Beck Center, Woodruff Library, Emory University

ALICE HICKCOX
Emory University
ahickco@emory.edu

JULIA LEON
Emory University
jleon@emory.edu

Emory University Library's electronic text center, known as the Beck Center for Electronic Collections and Services, holds a number of poetry collections that are served out on the web as separate databases. In order to span these collections of more than 200,000 poems we set out to develop a portal that would allow searching across several different databases. The portal provides students and faculty a tool for retrieving and reading individual poems for personal and classroom teaching and research. The portal was implemented using XML technology.

This is a story of the development of the site, explaining why the project was conceived, and how the library, Information Technology Division and faculty collaborated to design and implement the portal. The web-interface design and technical architecture will also be described.

Since 1995 the Beck Center has built a collection of texts that are tagged in SGML; many of these texts are poetry databases. Both commercially and locally produced texts comprise the electronic poetry holdings of the Beck Center. The texts were served separately out in a number of discrete poetry databases, each with its own search interface.

Emory's Irish Poets collection is one example of a locally produced digital collection. A portion of Emory's literary archives of Ireland's leading poets was digitized. The archives were developed over the course of the last twenty years, and includes worksheets of the poets affiliated with the Belfast Group-including Seamus Heaney, Michael Longley, and Paul Muldoon.

In addition to the substantial body of electronic texts that existed, campus-wide interest in poetry was building, demonstrated by the establishment of a Poetry Council which sponsors periodic poetry readings.

These forces converged to raise questions about how to make the use of on-line poetry resources easier to access and more logical in organization and use. A group of people from the library and the Information Technology support group for faculty came together to plan for a poetry portal. The Beck Center provided data management. ITD provided analysis and programming support. An informal focus group of Emory faculty who teach poetry provided subject matter expertise.

The Beck Center put forth the original vision of the project: to present poetry across multiple collections through a common web portal. A prototype of the portal was built as a proof of concept of the use of metadata and XML technology. It also served as the launching point for further design.

The unifying element of the disparate collections is the metadata, which contains information that was deemed useful for cataloging, indexing, and referencing items.

The portal was built entirely with open-source software, with the one exception of the search engine. The process of building the portal may be divided in to two sub-applications: data preparation and the web interface.

Data preparation involved converting the source collections from SGML to XML. From the XML metadata was created using XSLT and SAX software. The metadata was stored in the Dublin Core format.

The web interface for the poetry portal employs a suite of XML technologies. Specifically, Tomcat and Cocoon, from the Apache Software Foundation, serve up XML source files on the web. XSLT (XML Style Sheet Language Transformation) is used to transform XML to HTML.

The prototype application was demonstrated to English faculty involved in teaching poetry. A blue-sky discussion of how a portal could be useful and interesting in teaching and research formed the basis of the application requirements.

The interface provides browse and search capabilities. The collection may be browsed by author, title, first line, collection, or date. The user may search for poems, either by word used anywhere in the poem, by author, date or title. Either browsing or searching will take the user to a poem window or a series of poem windows.

From a poem window the user may also search for other poems by the same author, other poems of the same date, other poems with the same title or other poems in the same volume or collection. The user may also execute a simple search in the Oxford English Dictionary from the poem window, and bring up results in another window.

Additional features that we hope to implement include user-defined "poetry notebooks," which allow users to save links to particular poems or searches, a search history for registered users, and the ability to save searches.

Some associated features that were not part of the XML search-retrieval programming also emerged as special projects. One such project would be to collect manuscript and published versions of certain poems of poets that were in their collections in various forms, and to present them for study. Another specialized application is to have poems accompanied by audio files so that students can hear the poems read aloud. For some of the poems in their Special Collections, tapes were available of poets reading their own work. Other possibilities exist for audio performance as well.

Solving the Legacy-Encoding Debacle with On-line Transliteration

JOHN PAOLILLO
Indiana University
paolillo@indiana.edu

Non-roman scripts have always faced severe challenges in computer applications. Early challenges concerned the lack of non-roman support in ASCII. Today, Unicode provides or promises to provide support for almost all non-roman scripts. But Unicode support is not widespread in many languages, for example in the languages of South and Southeast Asia. In the last decade, the international expansion of the World-Wide Web caused demand for non-roman text encodings to rise faster than Unicode development could proceed. The consequent void was filled in many cases by ad-hoc 8-bit font encodings. While these encodings lack many of Unicode's advantages, they allowed many South Asian websites, especially newspaper companies, to establish a native-language web presence. Now they are firmly established in use: hundreds of web sites use special 8-bit encodings, with new material being added every day. These encodings are thus likely to be widely used for some time, even if Unicode support grows. In addition, many materials encoded in these forms are now legacy materials, and continued access to them is required for historical and other studies [Baker, et al., 2000].

These 8-bit encodings have many well-known problems, the most salient of which is the large number of alternative encoding schemes. Languages such as Sinhala or Tamil have three or four widely-used encodings, and Hindi has six or more. Few of these encodings reflect local, regional or international standardization efforts. Numerous fonts are required for display, and words are not likely to match across any two texts, seriously hampering efforts to search or index the documents. For example, when different newspapers use different font-based encodings to post stories on the same current events, the keywords for those stories will not match. Users searching for those stories must search separately for pages in each of the relevant encodings, or else fail to find them entirely. Many opportunities for humanistic, literary and linguistic research are encumbered by this situation. Unicode alone cannot solve these problems: conversion methods are needed to reconcile the variant text encodings.

This poster presents a general solution for these problems in the from of a protocol for transliterating variant text encodings of a language. The design of the protocol employs conversion tables for each supported encoding written in XML. Websites presenting materials in 8-bit encodings would publish such a conversion table where it can be readily retrieved by browsers and search engines. On the application side, the conversion tables are compiled by a general transliteration program into finite state transducers. When the encoded text is encountered, the transliterator is invoked to convert the encoding into one that can be used for indexing or display. Font detection is supported so that multilingual pages are handled appropriately, converting only that portion of the text in a document whose encoding requires it.

The transliterator may be used in any context where it is required. Search engines can select a conversion target for all web pages in a given language, e.g. Unicode, which would then be correctly indexed alongside other languages. Similarly, browsers could convert the encoding of a web page into one for which there is an available font, or convert a romanized query input into a suitable representation to be matched by a search engine, etc. Application software equipped to handle this protocol, whether web browsers, search engines, text editors or anything else, need only know the location of the conversion table for each encoding.

The transliteration program is based on the notion of a graphic template, which permits even highly complex alpha-syllabic Brahmi-based South Asian scripts to be efficiently transliterated. A graphic template is used instead of, e.g. phonetic representations, because it simplifies the many-to-many relations among script elements and their phonetic representation. Additionally, non-programmers and non-specialists can modify a conversion table more readily if its elements refer to graphic elements, rather than to phones or phonemes, which require specialized linguistic knowledge to appreciate. Graphic templates are finitely bounded, and hence can be parsed efficiently using finite state transducers, which are readily written in rule form [Antworth, 1990], hence, the conversion table is a list of finite state rules. The graphic template functions as an interlingual representation for the transliterator, meaning that bi-directional conversions between any two variant encodings can be realized by providing a suitable conversion table for each encoding. Roman transliterations are easily incorporated into the same framework, meaning that the protocol can used in additional ways for research purposes.

A working prototype transliterator written in SWI-Prolog will be demonstrated in several deployment scenarios: as a reading interface for several South Asian Language websites, in an interactive editor for text, and in an index and a query interface for a search engine. These scenarios illustrate the diversity of applications of the transliterator, as well as its ease-of-use and efficiency. It is hoped that widespread adoption of this system will facilitate the use of non-roman text for scholars and non-scholars alike.

REFERENCES

Antworth, Evan. 1990. *PC-KIMMO: A Two-Level Processor for Morphological Analysis*. Summer Institute of Linguistics, Dallas, Texas.

Baker, Paul; Tony McEnery, Mark Leisher, Hamish Cunningham, Robert Gaizauskas. 2000. *Mapping multiple South Asian 8-bit character sets to the Unicode standard*. Institute for Research on Cognitive Science, University of Pennsylvania, Philadelphia, Pennsylvania. Linguistic Exploration: Workshop on Web-Based Language Documentation and Description.

Unicode Consortium. 2002. *The Unicode Standard, Version 3.0*. Addison-Wesley Longman, New York, New York.

TAPoR Tools: Portal text analysis tools and other primitives

GEOFFREY ROCKWELL
McMaster University
grockwel@mcmaster.ca

LIAN YAN
McMaster University
lyan@mcmaster.ca

STÉFAN SINCLAIR
University of Alberta
Stefan.Sinclair@ualberta.ca

This poster will demonstrate a collection of text processing tools designed to work through a portal over the Web. The tools are designed to work on plain text, html or xml encoded e-texts. They are easily used to search electronic texts without the need to install software, preprocess the texts, or master complex tools.

HOW DO TAPOR.TOOLS (T.TOOLS) WORK?

T.tools are written in Ruby, an object-oriented scripting language like Perl and Python.[1] The T.tools are written so that they can be run on the command line or as CGI programs off our portal. This means that users of the tools need not install or maintain them, but if they wish, advanced users can download and adapt them.

Using Web forms as an interface to the tools gives T.tools the capacity to be easily adapted to hide complexity or to provide local adaptations. Simple search and concordance forms can be created that can utilize xml markup without having to change the tools. Small scale publishers of electronic texts can provide T.tool Web forms that process their Web accessible e-texts without installing the software. The forms simply pass the URL for the text in a hidden field to the appropriate tool residing on our portal for processing. (We will demonstrate the adaptation of these tools to support the Hyperliste project, a collection of French medieval poetry online.)

WHAT ARE PORTAL TOOLS?

A portal is an entry point into a field.[2] In this case the T.tools are written to provide simple text processing tools for TAPoR (Text Analysis Portal for Research), a multi-institutional project which has Canada Foundation for Innovation funding to create a portal for text analysis.[3] They are designed to provide a suite of simple text transformations that will eventually be managed by a portal environment that additionally provides user and interface customization tools. At present they can do the following:

- List and count words in a text.
- List and count elements in an xml text.

- List attributes and values in an xml text.
- Extract elements from an xml text.
- Find patterns (words or phrases) in a text.
- Find patterns in specific elements in a text.
- Create a concordance of found patterns or elements.
- Output results in either html for reading or xml for further processing.

This combination of functions allows the user to query an xml text to find words in specific parts of text or to extract selected parts by element name and attribute value. Users can also, should they not know the structure of the text, get a list of the elements or a list of words to search for. Users have a choice of output from simple output in html to output in xml that could be saved and processed locally.

WHAT TYPES OF USERS ARE ENVISAGED?

T.tools are designed to be used in three ways by users of differing levels of expertise.

Introductory Users. A portal should be place one can learn through playful discovery about a field like computer assisted text analysis. T.tools are designed so that new users can try basic operations on electronic texts without having to install software or texts. As the T.tools do not preprocess texts they can be run on any text a new user can find on the Web. This allows a new user to experiment with text analysis on texts they know without much training. The tools are also designed so that they can be explained and documented in different ways to make them accessible to different communities.

Small E-text Publishers. While large e-text projects have access to programmers and systems that allow them to adapt text processing tools to their texts, many small projects cannot afford to do more than make available their scholarly texts on the Web in html or xml/css form. T.tools provides tools run on our server which can be passed a text (actually a URL) for processing from a Web form set up by the publisher. Thus small projects can adapt our forms to their needs and integrate them into their sites.

Advanced Users. As the code is made available as "open source" according to the definition at the Open Source Initiative, advanced users can download it and adapt it to their research needs.[4] T.tools have been built around a library of object classes commonly used in text processing whose methods can be called in new Ruby scripts, from the command line, or through IRB (Interactive Ruby). Thus the advanced user can use T.tools as a text processing language and then build new scripts to do things unanticipated by the developers.

WHAT ARE THE PROBLEMS WITH THIS MODEL?

The major drawback to this model is that T.tools are slow by comparison to other Web text tools like TACTWeb because they do not work with preprocessed indexes.[5] T.tools work best on chapter to book length texts not on larger corpora. The processing capabilities of the server used for the portal is also important. TAPoR has been funded to install high-end servers for the portal that will partially compensate for the cost of processing, but there is no substitution for more efficient tools when dealing with large texts. This is an issue which deserves further study. Further, T.tools is based on a "pipe-and-flow" model common in the Unix world which may not scale to certain types of interactive processing needed in the humanities, for example, situations where texts are being enriched and studied simultaneously.

WHY A POSTER?

This project is presented as a poster as that will provide the most convenient way to demonstrate the tools to potential users and redevelopers. Through the venue of a poster session we can demonstrate how T.tools work on the texts of the poster visitors. It will also allow us to engage humanists with small text projects that might benefit from such tools. With individual visitors we can adapt Web forms to show the usefulness of T.tools to their projects. CD-ROMs of the code and documentation will be distributed to advanced users.

REFERENCES

Ruby is available for Macintosh, Unix and Windows at the "Ruby Home Page", URL: **http://www.ruby-lang.org/en/**, Accessed Nov. 20, 2002.

Katz, Richard N. and Associates. *Web Portals and Higher Education; Technologies to Make IT Personal*. San Francisco: Jossey-Bass, 2002.

Text Analysis Portal for Research, URL: **http://www.tapor.ca**, Accessed Nov. 21, 2002. It should be noted that in this poster presentation we will not be presenting on the portal as a whole, just a specific set of tools designed to work through (or not) the portal that is being implemented.

"The Open Source Definition", URL: **http://www.opensource.org/docs/definition_plain.php**, Accessed Nov. 20, 2002.

For more on TACTWeb see URL: **http://tactweb.humanities.mcmaster.ca/**, Accessed Nov. 20, 2002. TACTWeb is built on TACT which has a preindexing program MAKEBASE which prepares the Text DataBase file (TDB) which is then used by TACT and TACTWeb to quickly process queries.

Web Prompts the Increase of Chinese Non-English Majors' Speaking, Writing and Translating Abilities

YAN TIAN
School of Foreign Languages, Shanghai Jiao Tong University
maryyantian@yahoo.com

English is taught as a foreign language in China, which means there is no English environment for students to get immersed in after class. However, English is a compulsory course for Chinese college students. According to “The English Syllabus for Non-English Majors”, students should be quite skilled at speaking, writing and translating abilities. Yet non-English majors only have four periods (180 minutes) of English classes every week from freshman year to sophomore year. Teachers of English are supposed to be responsible for increasing students’ productive skills (speaking, writing, and translating) as well as their receptive skills (listening and reading). However, compared with their receptive skills, their productive skills are rather poor despite the fact that they can get high scores in various English tests. Therefore, teachers are always at a loss as to how to increase students’ productive skills in and out of class. Furthermore, there are usually 30~40 students in an English class, which makes it almost impossible to carry on in-class activities. In view of the above mentioned, the researcher turns to the Web for help.

Based on a research project, this paper explores a new possibility of increasing students’ productive skills. In previous teaching practices, students were only interested in the designed activities aimed at increasing their productive skills at the beginning of the semester. Then no matter how hard the teacher had tried, they would have lost interest gradually as time went on. This is partly due to their limited language proficiency and partly due to the low efficiency of in-class activities. As a result, the students’ receptive skills are always far better than their productive ones. In this project, the researcher takes the advantage of the Web to arouse students’ interest in increasing their productive skills.

Web has a charm that almost no one can resist. Many students spend hours surfing on the Web searching for information, sending emails, chatting with friends or playing online games. Some even have Web sites of their own before they entered the university. What’s more, a few have had the experience of establishing Web sites for business companies.

At Shanghai Jiao Tong University, all students live on campus and the computers in their dormitories are all connected to the World Wide Web. Under this circumstance, the researcher encouraged the students to establish their own English learning Web site to demonstrate their progresses in their productive skills.

The subjects are 165 freshmen from Shanghai Jiao Tong University who enrolled in the fall semester of 2002. They were assigned to four English classes before the semester begun. In this project, students in each class were required to sign up for at least one productive skill group, e.g. speaking group, writing group or translating group, according to their own interest. Then each group was asked to select one director taking charge of the group activities and one assistant-director responsible for coordinating the group activities. They were also asked to select five “computer engineers”. Those engineers were responsible for the designing and maintaining of the Web site. Three of them were also coordinators, contacting the speaking groups, the writing groups and the translating groups of the four classes respectively. In order to guarantee enough materials for the Web, each productive group was required to meet and practice at least once a week and to report in class once a week about their group activities.

The students are very skilled at establishing World Wide Web. The main tools used are Dreamweaver and Frontpage. The School of Foreign Languages at Shanghai Jiao Tong University offered a server for this project. In one month, they established their English learning Web, called “Cool English”. Since all the dormitories are connected to the World Wide Web, students can easily visit their Web site as well as the World Wide Web. At the beginning, there were only three columns, namely Speaking, Writing and Translating on it. Each week, directors of each group send, by emails, the “achievements” of their groups to the computer engineers. After proof reading and checking technically, the computer engineers uploaded those materials to their Web site#“Cool English”. They were very excited at the beginning, but after a couple of weeks, the students were not satisfied with the three columns at all. So they searched on the World Wide Web for interesting materials and expanded the contents to include such columns as “English News”, “English Idioms”, “English Jokes”, “Top Students’ Speeches on English Study”, to name only a few. Besides, they also offered on their Web site some other English learning Web sites for the students’ convenience.

Although the Web site “Cool English” is still at its infancy, yet it has attracted the attention of many

students, including those from other classes. Virtually in order to feed their Web site, the students have to surf on the Internet, searching for useful and interesting information to enrich their Web site constantly. In this process, their receptive skills, reading skill in particular, have also been increased dramatically.

From this project, the researcher comes to the conclusion that the Web has served as a bridge between in-class activities and out-of-class activities and has connected the receptive skills with the productive ones. Furthermore, it is a faster, cheaper and fascinating “publishing house” for students. Their sense of achievement has been greatly increased by working on the Internet.

Primarily History: Historians' Search for Primary Resource Materials

---

HELEN TIBBO
UNC-CH
tibbo@ils.unc.edu

OVERVIEW.

This paper will present findings from the Primarily History project, concerning how historians are locating primary sources in the early years of the digital age. Primarily History, an international study housed at the University of North Carolina at Chapel Hill and the University of Glasgow, also explores how historians teach their students to find such research materials and how archivists and librarians can facilitate this education and resource discovery and use. Data provided here reflect information-seeking behaviors of historians working in the United States.

NEW PATHWAYS TO PRIMARY RESOURCES.

Historical research has long been a detective game. While ascertaining the veracity and authenticity of primary resources and the critical interpretation of documents within the context of historical and cultural understanding lie at the heart of historical scholarship, the not so trivial task of locating these materials that serve as the grist of history, must precede any high level analysis. Throughout the 20th century, historians sought the materials they needed to shed light on the past in a number of time-tested ways. Studies conducted in the 1970s and 1980s found that historians most often located relevant primary source materials by following references in published histories, talking with colleagues, and searching likely repositories. In a print paradigm these were all appropriate and efficient methodologies. Recent technologies, however, present the historian with many new possibilities for locating research materials that may prove more efficient and even more effective.

The road to online finding aids and other digitized resources was not an easy one for archival repositories. For the past twenty years archivists have expended a good deal of time, money, intellectual effort, and angst to produce electronic access tools for the collections in their repositories. The first efforts focused on creating MARC/AMC (Machine Readable Cataloging/Archives and Manuscripts Collections) records for online catalogs and bibliographic utilities such as OCLC and RLIN. Many archivists doubted the efficacy of MARC, a library-based standard, for archival description. In the early 1990’s, some pioneering archivists mounted machine-readable text finding aids on gopher sites on the Internet. By the mid-1990s, a small group of archivists were developing the Encoded Archival Description SGML DTD. By the turn of the millennium, most special collection repositories, at least those in larger units such as academic libraries, had some sort of website, many of which contained HTML encoded finding aids. Today, a small but steadily growing number of repositories have EAD finding aids at their websites. These networked tools not only facilitate information discovery, they can also prepare a scholar for a very productive in-person visit to a repository.

Today’s historians can study digitized collections of materials online as well as search bibliographic databases with descriptions of both secondary and primary materials. Perhaps most significantly, historians can read, download, or print entire finding aids for collections in a large number of archival and manuscript repositories world wide. Now that there is a significant corpus of finding aids online it is time to explore how researchers such as historians use them and how archivists might make them more useful. While spending a good deal to create electronic access, to date, archivists have conducted few user studies to judge its effectiveness or ascertain the need for enhanced user education.

PRIMARILY HISTORY PROJECT.

Funded by the Gladys Kriebel Delmas Foundation, The Primarily History project, a collaboration of the School of Information and Library Science at the University of North Carolina at Chapel Hill (UNC-CH) and the Humanities Advanced Technology and Information Institute (HATII) at the University of Glasgow, Scotland, is the first international, comparative project to explore historians' information-seeking behaviors in today's web-based, networked environments. Perhaps most importantly, we are examining how historians are preparing the next generation of scholars, specifically, what they are teaching their graduate students about information seeking in the digital library environment and how the students are learning to use retrieval tools. This project is also surveying how special collections libraries and archives provide access to these materials and is seeking enhanced models for outreach and user education that will facilitate historians and their students in locating and using primary resources.

Through surveys and interviews we are exploring how historians are employing these new tools and techniques. Dr. Tibbo from UNC-CH has surveyed 700 historians located at leading U.S. history programs; Dr. Ian Anderson from Glasgow surveyed close to 800 historians working at universities in the United Kingdom. Both investigators followed the surveys with in-depth interviews with a subset of these populations.

Questions in the initial stage of this research focused on scholarly information-seeking in archives, manuscript repositories, and special collections libraries. We asked historians how they found the materials upon which they based their research. Questions included: Do they use library catalogs or the OCLC or RLIN databases that contain MARC records of finding aids? Do they use electronic indexes such as Archives USA? Do they search the Web for finding aids using keywords? Do they go directly to the websites of likely repositories and search for materials? The interviews asked historians to describe how they had located primary resource materials for a recent project and what they taught their graduate students regarding information seeking methodologies.

A SAMPLING OF FINDINGS.

For many historians working in the United States, the traditional methodologies for locating primary materials remain the most utilized. Ninety-eight percent indicated that they found materials by following leads and citations in printed sources; 77% searched printed bibliographies; 57% consulted printed documentary editions; 73% searched printed finding aids; 73% searched printed repository guides; 51% used newspaper files to find other materials; 50% used government documents in this way; and 38% used the now out-of-date printed NUCMC volumes. The most noticeable difference in behaviors across ranks is that only 24% of assistant professors used the printed NUCMC volumes with 40% of all full professors, including distinguished faculty and deans, using this now out-of-date tool. Interestingly, however, only 18% of those individuals who had taught history for 40 years or more searched the printed NUCMC.

Use of traditional resources, however, does not preclude the use of digital technologies developed within the last decade and populated with research materials (e.g., a critical mass of electronic finding aids and some digitized documents) only within the last three to four years. Sixty-nine percent of the history scholars used their own institution's online (library) public access catalog (OPAC) [while digital, and now probably a web-based technology, not really a new approach]; 64% searched other institutions' OPACs via the Internet; 59% used bibliographic utilities such as RLIN and OCLC; 58% said they looked for information directly on repository websites; 40% indicated that they searched the Web for primary materials using a search engine such as Alta Vista or Google. Twelve percent of respondents indicated they used the Archives USA database, while 14% searched NUCMC online from the Library of Congress website with only 8 individuals (3% of respondents) exhibiting both behaviors.

At the most superficial level, we can see that historians are using traditional finding tools more heavily than newer approaches but we need qualitative interview data to understand why. It may well be that many individuals are most comfortable employing methodologies they have found useful throughout their careers and that these tools remain sufficient. It may also be that they have not found searching the web or bibliographic utilities very useful. After all, most archival websites have mounted electronic finding aids only since 1996 with many institutions still working on making a significant portion of their inventories web accessible. When we look at the 220 projects with end dates of 1998 or later, 135 or 61% of these researchers visited repository websites within the course of their work and 98 (45%) used web search engines to find relevant collections.

Significantly, the historians in this study averaged over twenty years of teaching experience each. This is significant because the initial round of interviews indicated that many of the veteran historians had completed the largest part of their information seeking years ago, having located large stores of records and papers to analyze for many years. These individuals appear to invest little time in learning new information-seeking behaviors. Other individuals who more dramatically shifted topics across projects, might well seek out primary resources during their entire careers.

PRELIMINARY CONCLUSIONS.

While much work remains in this study, it is clear that archivists and special collection curators must both maintain traditional print descriptive tools and create an increasing number of electronic finding aids and digitized documents as historians are employing a wide range of information seeking behaviors. Evidence to date indicates that younger scholars are turning to digital resources more often than their older counterparts but such behaviors may also linked be linked to research topic. Additional interviews, especially with current Ph.D. students will clarify emerging historical research trends in the early years of the Digital Age. Increasing use of the Web for information seeking and virtual trips to repositories demands that archivists create highly accessible websites, study user behavior so as to improve discovery tools, and provide increased user instruction for those who will only visit the collection from afar.

Index of Authors