# Variation of Style: Diachronic Aspect

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

## Introduction

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

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

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

## Material

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

## Characteristics

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

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

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

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

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

## Method

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

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

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

## Results

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

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

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

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

**Period 1**

But when within thy waves she looks –

Which glistens then, and trembles –

Why, then, the prettiest of brooks

Her worshipper resembles –

For in my heart – as in thy streem –

Her image deeply lies <...>

*(To the River)*

**Period 2**

You know the most enormous flower –

That rose – <...>

I tore it from its pride of place

And shook it into pieces <...>

*(Fairy Land)*

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

## Bibliography

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

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

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

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

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

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

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

# Exploring Historical Image Collections with Collaborative Faceted Classification

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

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

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

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

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

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

## Introduction

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

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

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

## Related Research

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

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

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

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

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

## Architecture and Prototype Implementation

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

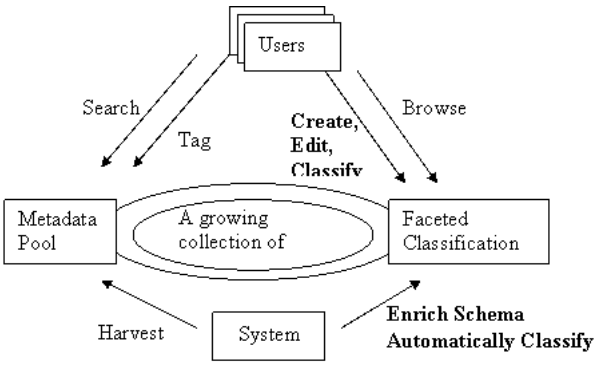



*Figure 1. System Architecture*

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

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

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

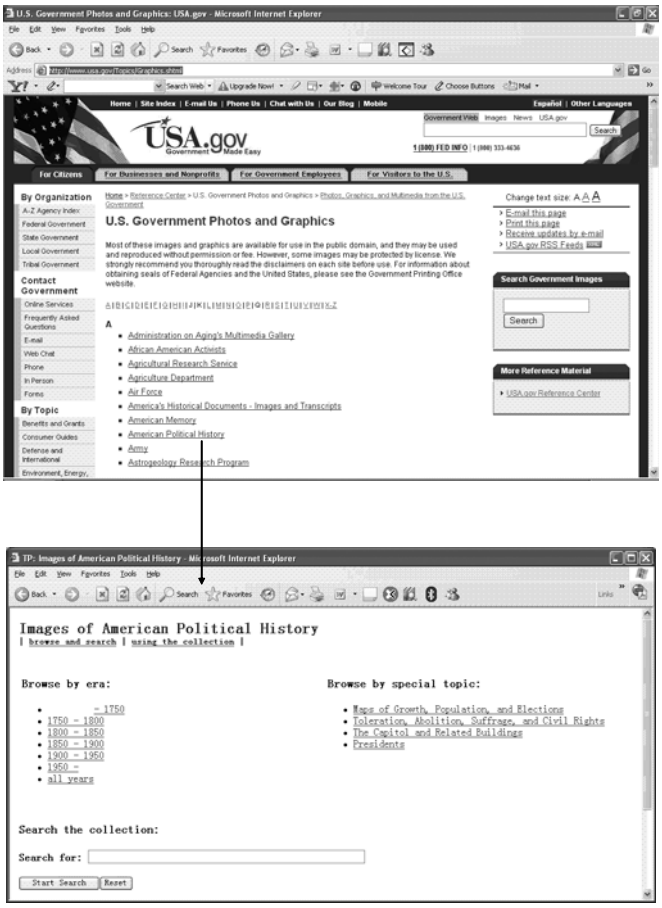
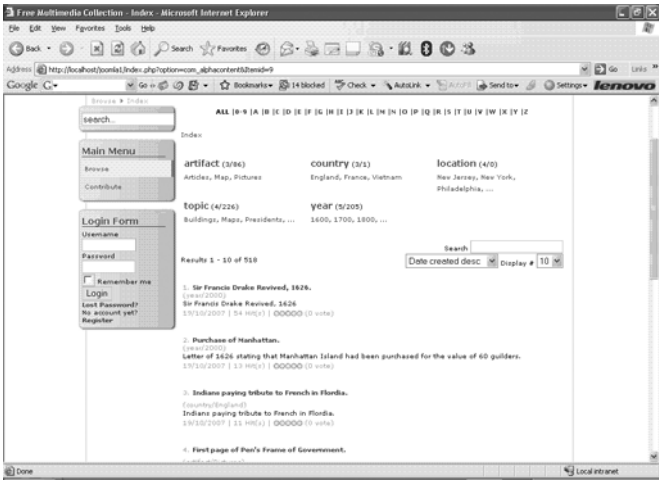



*Figure 2. American Political History Collection at usa.gov*

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

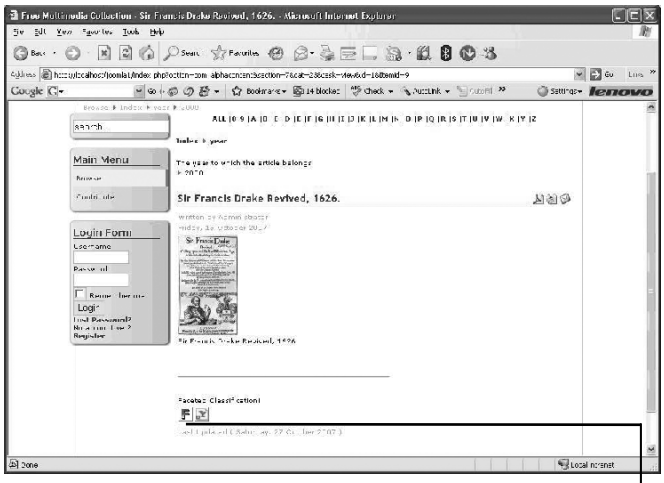

*The system integrated with a Flamenco-like Interface*



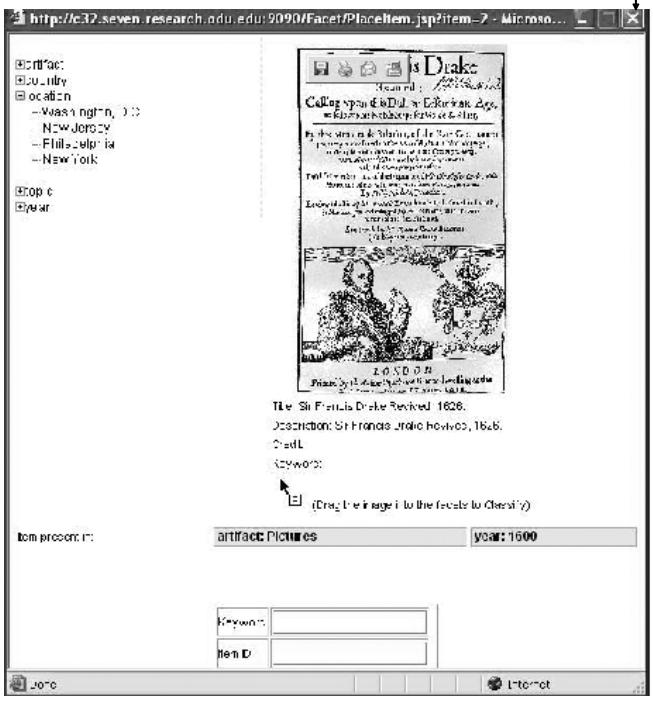

*The system integrated with Joomla!, a popular content management system*

*Figure 3. Multi-facet Browsing*

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




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




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

*Figure 4. Classification Interface*

## Evaluation and Future Steps

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

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

## Acknowledgements

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

## References

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

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

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

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

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

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

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

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

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

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

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

## Introduction

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

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

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

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

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

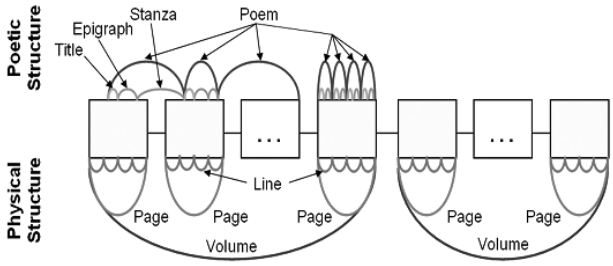



*Figure 1: A simplified diagram showing an edition of collected poems (in two volumes) represented as an annotated facsimile edition.*

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

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

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

## Annotated Facsimile Editions

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

## The Image Stream

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

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

*Table 1: Information represented by an annotation.*

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

## Annotations

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

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

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

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

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

## Discussion

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

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

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

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

## Acknowledgements

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

## References

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

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

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

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

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

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

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

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

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

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

# CritSpace: Using Spatial Hypertext to Model Visually Complex Documents

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

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

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

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

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

## Introduction

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

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

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

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

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

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

## System Design

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

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

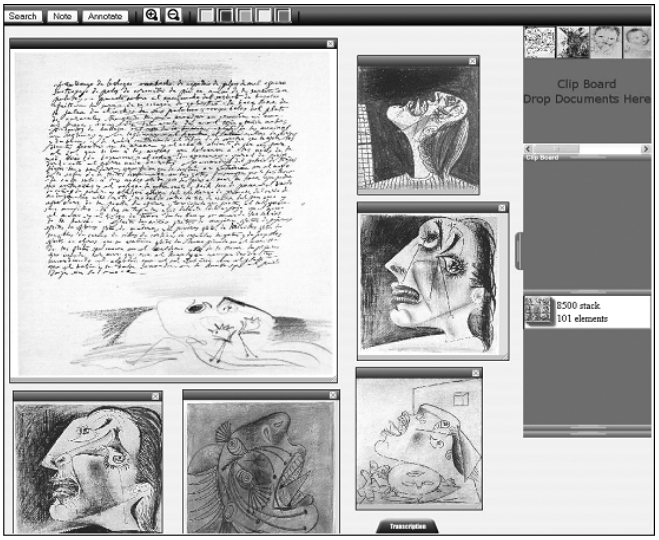



*Figure 1: Screenshot of CritSpace showing a document along with several related preparatory sketches Picasso made for Guernica about the same time.*

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

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

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

## Future Work

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

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

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

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

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

## Discussion

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

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

## Acknowledgements

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

## References

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

# Glimpses though the clouds: collocates in a new light

***David Beavan***

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

*University of Glasgow, UK*

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

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

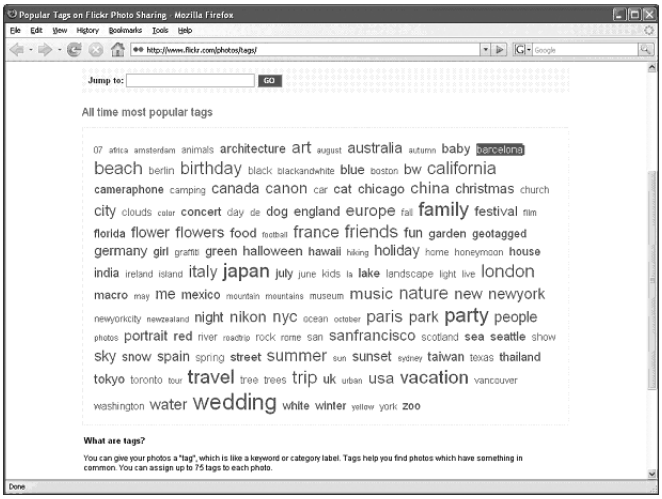



*Figure 1. Flickr tag cloud showing 125 of the most popular photograph keywords*

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

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

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

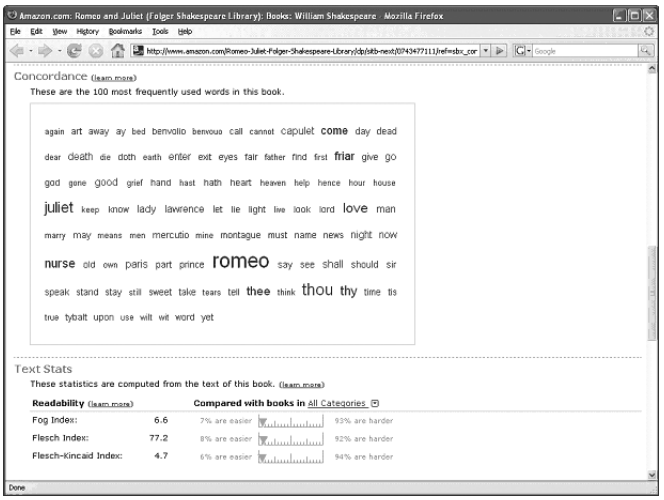



*Figure 2. Amazon.com's 'Concordance' displaying the 100 most frequent words in Romeo and Juliet*

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

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

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

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

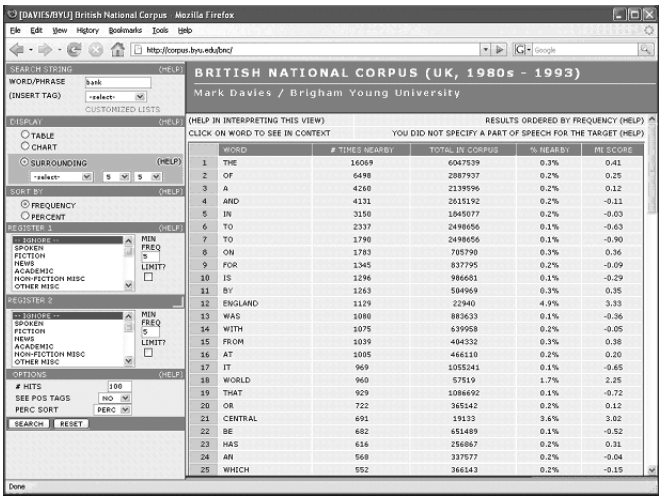



*Figure 3. British National Corpus through interface developed by Mark Davies, searching for 'bank', showing collocates*

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

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

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

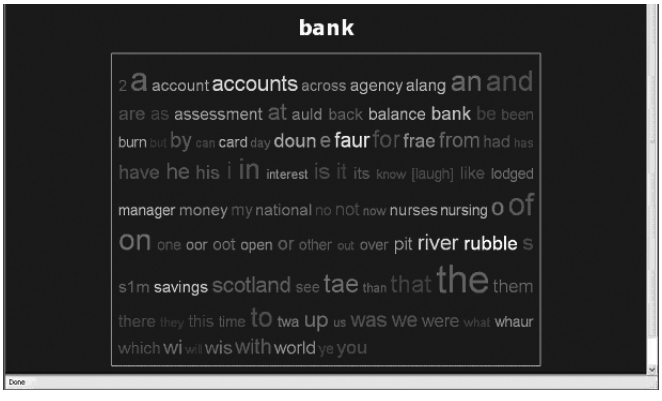



*Figure 4. Demonstration of collocate cloud, showing node word 'bank'*

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

## Notes

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

## Introduction

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

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

## AAC-FACKEL

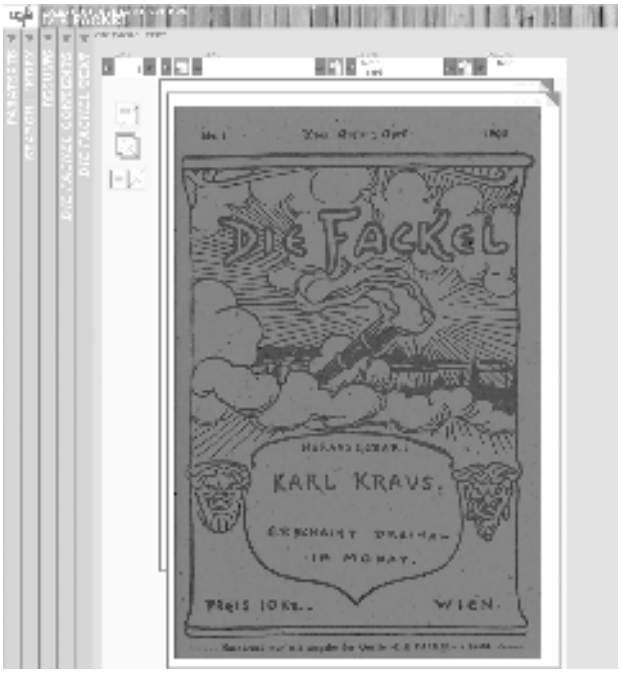



*Figure 1. AAC-FACKEL Interface*

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

## BRENNER ONLINE


*Figure 2. BRENNER ONLINE Interface*

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

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

## Conclusion

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

## References

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

## References

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

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

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

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

# An OWL-based index of emblem metaphors

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

## Intro

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

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

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

## Metaphor model

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

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

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

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

## Application

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

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

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

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

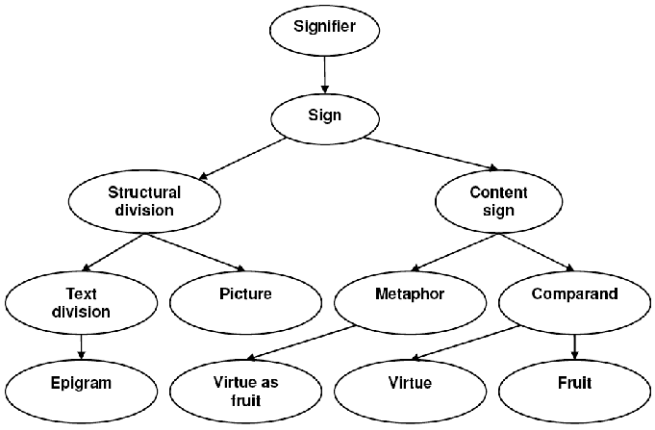



*Figure 1 Part of the classes that make up the metaphor ontology. Arrows point to subclasses. The classes at the bottom level are just examples; many other could have been shown if more space were available. For simplicity, this diagram ignores class properties*

## Discussion

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

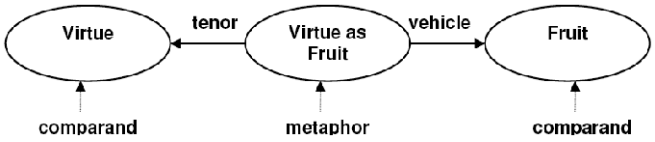



*Figure 2 A metaphor and the properties relating it to the comparands*


*Figure 3 Objects can be queried by their relations*

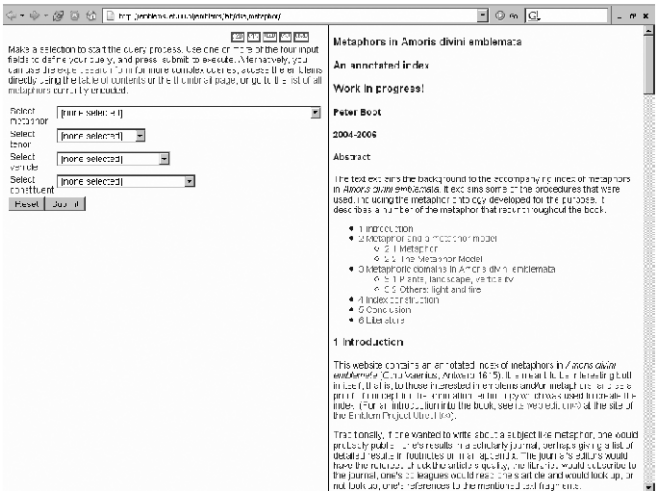



*Figure 4 Overview perspective*

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

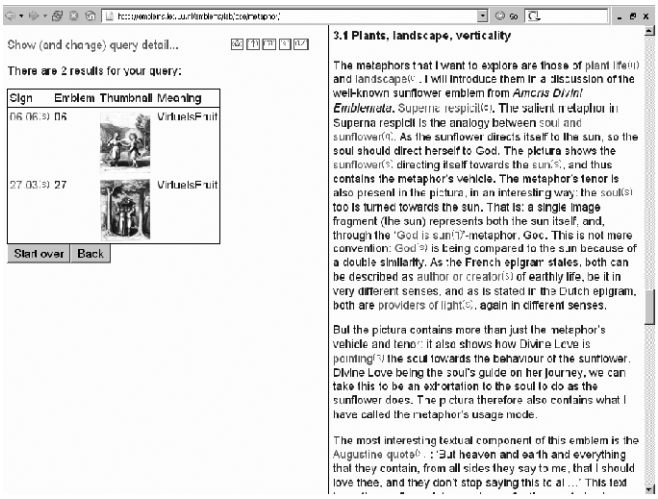



*Figure 5 Clicking the hyperlink 'plant life' (top right) executes a query with hits shown in the left panel*

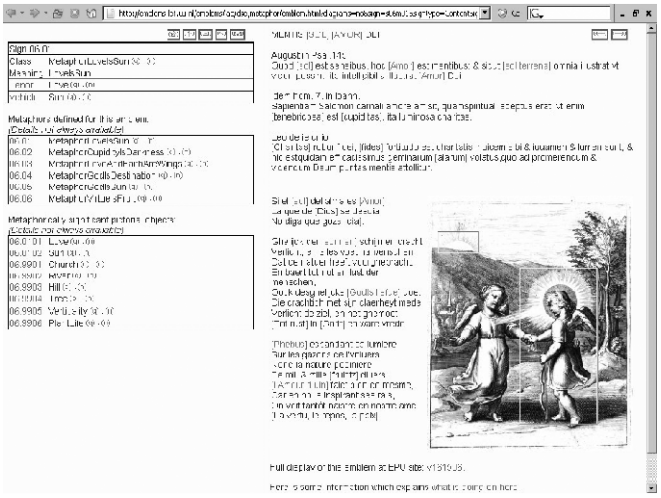


*Figure 6 Emblem perspective with one metaphor highlighted in picture and text*

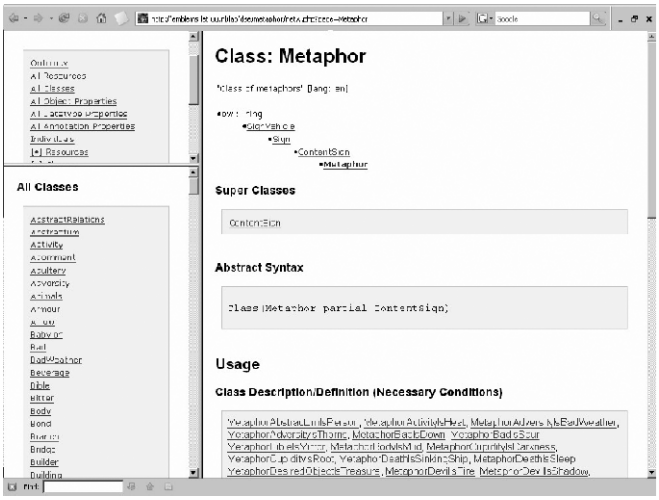



*Figure 7 Ontology perspective , with display of class metaphor*

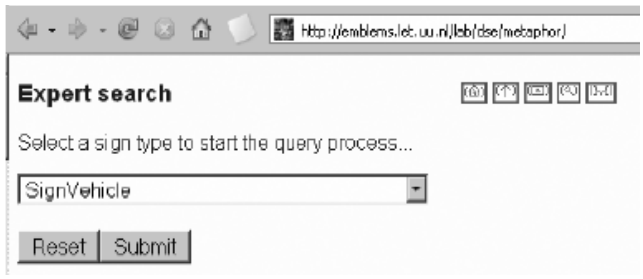



*Figure 8 Expert search, start building query*

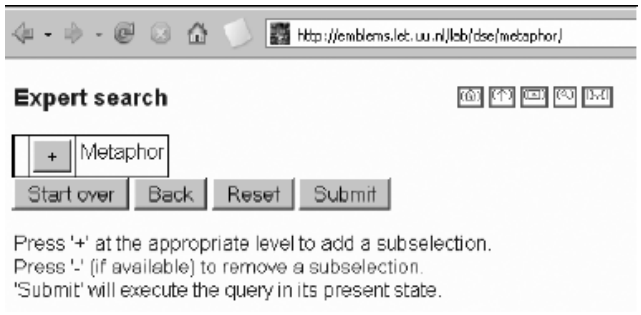



*Figure 9 Expert search. Click '+' to create more criteria*

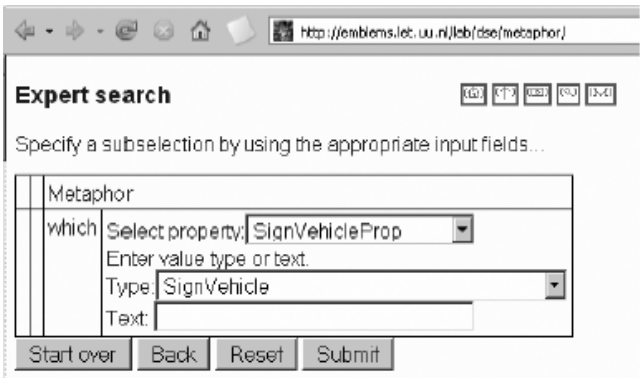



*Figure 10 Expert search. Select the desired criterion*

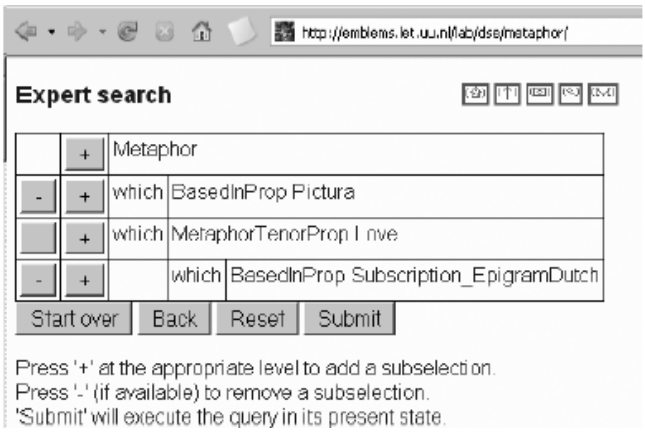



*Figure 11 Expert search. Final state of query*

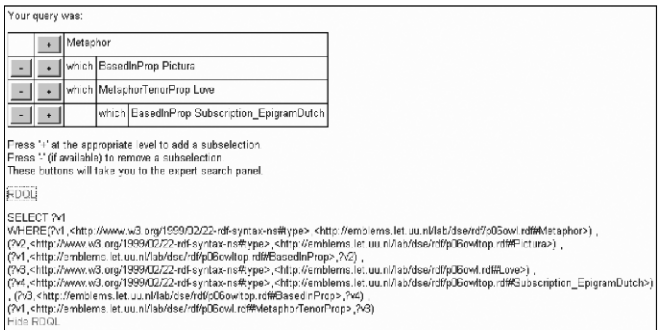



*Figure 12 Expert search. Display of executed query and generated RDQL in results panel*

## References

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

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

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

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

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

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

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

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

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

# Collaborative tool-building with Pliny: a progress report

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

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

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

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

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

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

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

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

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

Pliny, then, is about two issues:

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

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

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

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

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

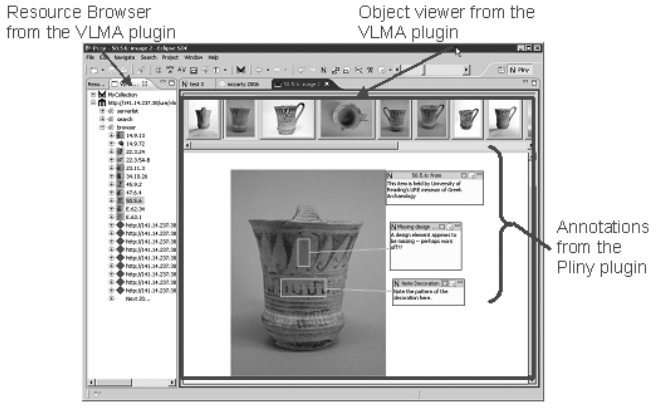



*Figure I: Pliny annotations in a VLMA viewer*

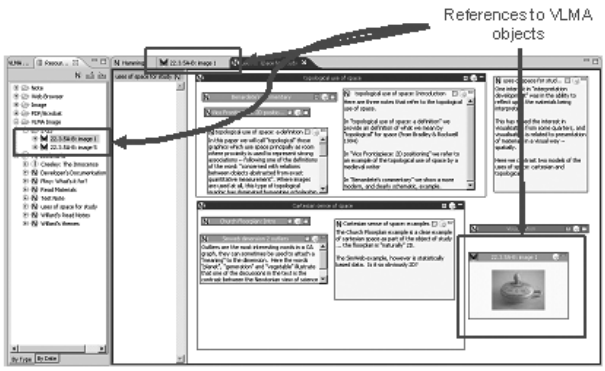



*Figure II:VLMA objects in a Pliny context*

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

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

## References


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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

## Introduction

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

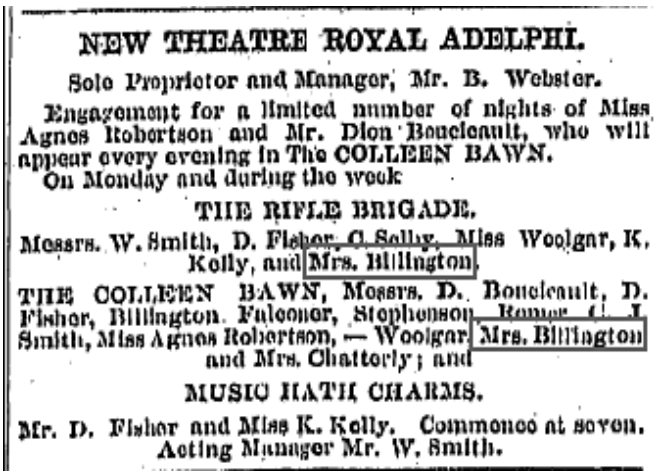
The application of text mining techniques (named entity extraction, text categorisation, etc.) allowed names, institutions and places etc. to be extracted as well as individual articles to be classified according to events, subjects and genres. Users of the digital edition can search for very specific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

## The problem

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

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

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



NEW THEATRE ROYAL ADELPHI.

Sole Proprietor and Manager, Mr. B. Webster.

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

On Monday and during the week

THE RIFLE BRIGADE.

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

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

MUSIC HATH CHARMS.

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

Acting Manager Mr. W. Smith.

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

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

## The algorithm

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

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

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

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

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

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

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

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

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

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

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

## Bibliography

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

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

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

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

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

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

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

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

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

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

# Degrees of Connection: the close interlinkages of Orlando

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

## References

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

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

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

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

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

# The impact of digital interfaces on virtual gender images

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

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

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

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

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

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

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

## Introduction

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

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

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

## Research Perspectives & Research Questions

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

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

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

Our investigations and reflections are guided by the following questions:

- Which gender images do exist in virtual spaces?

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

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

## Objects of Investigation

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


*Image 1 LM: Interface*

*Image 1 SL: Interface*

## Methodology

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

Body presence

- modes of existence or being there

Personality / individuality

- forms of personal performance

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

- modes of emotional expressions

Patterns of gender

- categories or models of gender

Actions and spatial Movements

- modes of behavior and actions

- modes of spatial orientation and navigation

Patterns of social behavior

- modes of communication and interaction

- community standards, behavioral guidelines and rules

## Research Results

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

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

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


*Image 2.1 LM: Rename command*

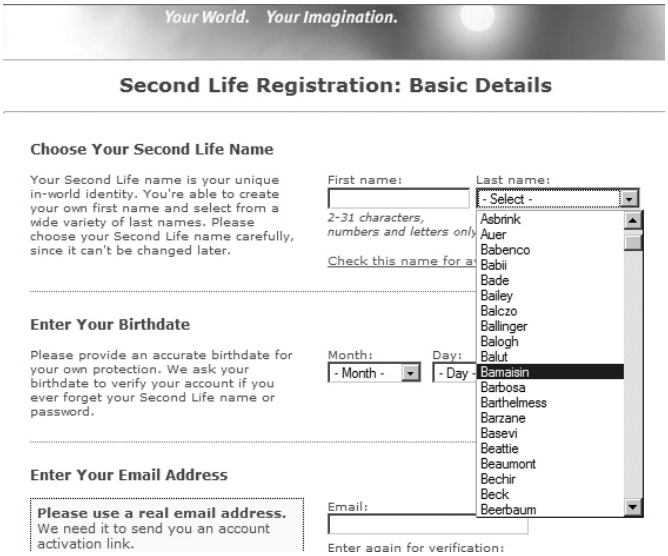



*Image 2.1 SL: Choose surname*


*Image 2.2 LM: Nickname Magenta_Guest*

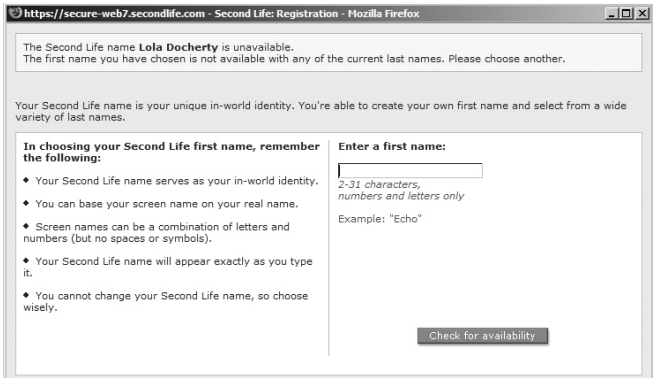



*Image 2.2 SL: Name check*

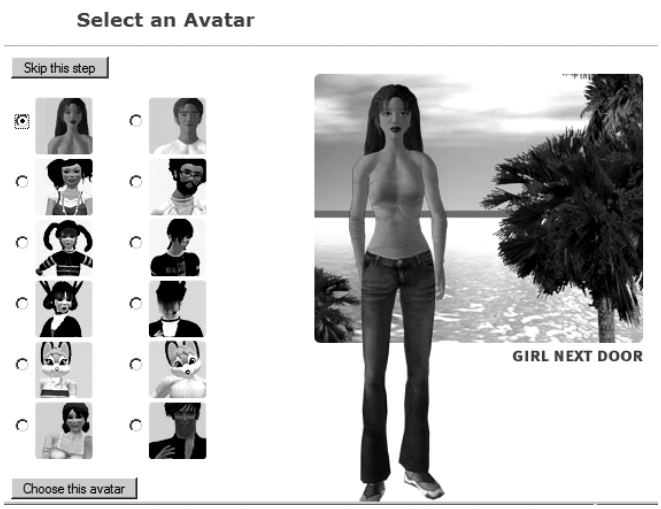



*Image 3 SL: First set of avatars*


*Image 4.1 LM: Gender categories*

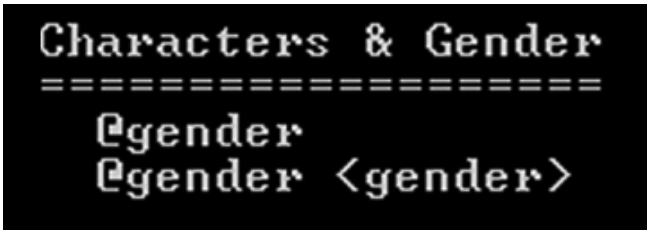



*Image 4.2 LM: Re-gender*

*Image 5.1 SL: Male avatar without genitals*

*Image 5.2 SL: Edit masculinity - small*

*Image 5.3 SL: Edit masculinity - big*

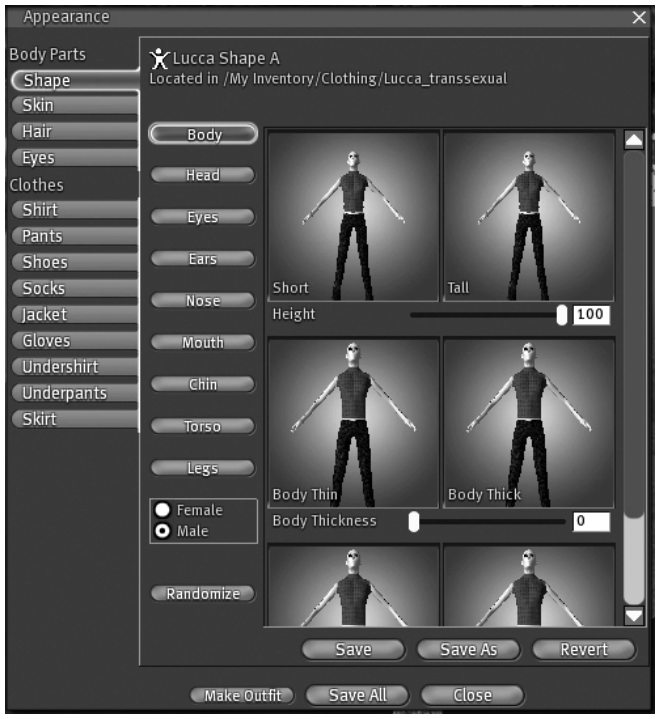



*Image 6 SL:Appearance editor*

*Image 7.1 SL:Transgender*


*Image 7.2 SL: Gender swap from female to male*

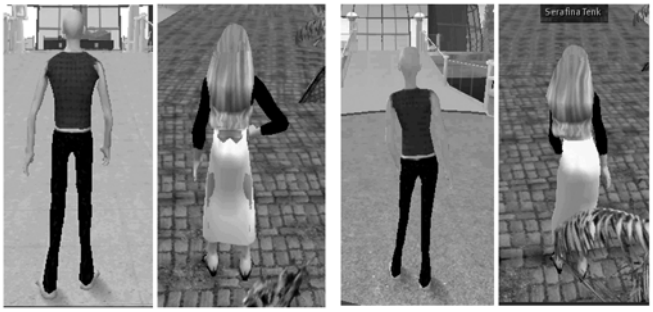



*Image 8.1 SL:Walking avatars*

*Image 8.2 SL: Sitting bowlegged/close legged*

*Image 9.1 SL: Common gestures - female avatar*

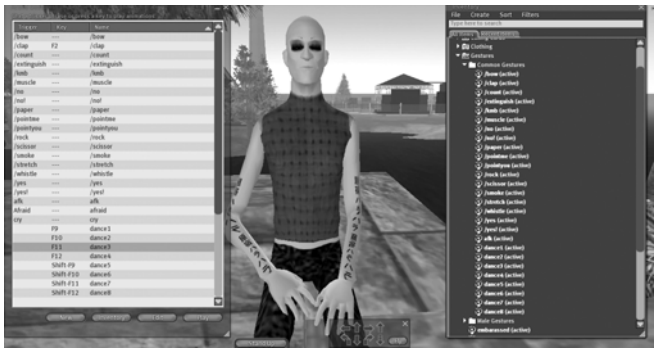


*Image 9.2 SL: Common gestures - male avatar*

*Image 10.1 SL: Female gestures - female avatar*

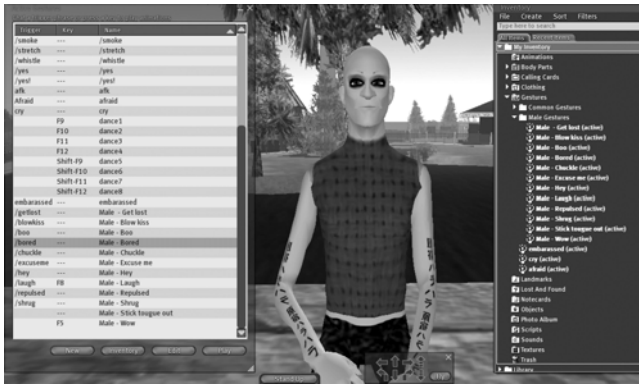


*Image 10.2 SL: Male gestures - male avatar*

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

*Image 11.1 SL: Female gestures - male avatar*

*Image 11.2 SL: Male gestures - female avatar*

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

*Image 12 SL: Female to male swap - unfitting clothes*

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

## Outlook

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

## References


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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Rex, Felis alias Richards, Rob: http://www.LambdaMOO.info/

Rommes, E. (2000): Gendered User Representations. In Balka, E. and Smith, R. (ed.): *Women, Work and Computerization. Charting a Course to the Future*. Dodrecht, Boston: Kluwer Academic Pub, p. 137-145.

Second Life Blog: April 2007 Key Metrics Released http://blog.secondlife.com/2007/05/10/april-2007-key-metrics-released/, http://s3.amazonaws.com/static-secondlife-com/economy/stats_200705.xls

Teubner, U. and A. Wetterer (1999): Soziale Konstruktion transparent gemacht. In Lober, J. (Editor): *Gender Paradoxien*, Leske & Budrich: Opladen, p. 9-29.

Thomas, Sue (2003): www.barcelonareview.com/35/e_st.htm, issue 35, March - April

Trettin, K. (1997): Probleme des Geschlechterkonstruktivismus. In: G. Völger, Editor: Sie und Er, Rautenstrauch-Joest-Museum, Köln

Turkle, Sherry (1986): *Die Wunschmaschine. Der Computer als zweites Ich*, Reinbek bei Hamburg

Turkle, Sherry (1999): *Leben im Netz. Identität in Zeiten des Internet*, Reinbek bei Hamburg

West, C., Zimmermann, D.H. (1991): Doing Gender. In Lorber, J. and Farell, S. A. (eds): *The Social Construction of Gender,* Newbury Park/ London, p. 13 – 37

West, C., Zimmerman, D.H. (1995): Doing Difference. *Gender & Society*, 9, p. 8-37.

# Towards a model for dynamic text editions

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

***Malte Rehbein***
*malte.rehbein@nuigalway.ie*
*National University of Ireland, Galway, Ireland*

Creating digital editions so far is devoted for the most part to visualisation of the text. The move from mental to machine processing, as envisaged in the Semantic Web initiative, has not yet become a priority for the editorial practice in a digital environment. This drawback seems to reside in the almost exclusive attention paid until now to markup at the expense of textual data models. The move from "the database as edition" [Thaller, 1991: 156-59] to the "edition as a database" [Buzzetti et al., 1992] seems to survive only in a few examples. As a way forward we might regard digital editions to care more about processing textual information rather than just being satisfied with its visualisation.

Here we shall concentrate on a recent case study [Rehbein, forthcoming], trying to focus on the kind of logical relationship that is established there between the markup and a database managing contextual and procedural information about the text. The relationship between the markup and a data model for textual information seems to constitute the clue to the representation of textual mobility. From an analysis of this kind of relationship we shall tentatively try to elicit a dynamic model to represent textual phenomena such as variation and interpretation.

## I.

The case study uses the digital edition of a manuscript containing legal texts from the late medieval town Göttingen. The text shows that this law was everything else but unchangeable. With it, the city council reacted permanently on economical, political or social changes, thus adopting the law to a changing environment. The text is consequently characterised by its many revisions made by the scribes either by changing existing text or creating new versions of it. What has come to us is, thus, a multi-layered text, reflecting the evolutionary development of the law.

In order to visualise and to process the text and its changes, not only the textual expression but, what is more, its context has to be regarded and described: when was the law changed, what was the motivation for this and what were the consequences? Answers to these questions are in fact required in order to reconstruct the different layers of the text and thereby the evolution of the law. Regarding the text nowadays, it is however not always obvious how to date the alterations. It is sometimes even not clear to reveal their chronological order.

A simple example shall prove this assumption. Consider the sentence which is taken from the Göttingen bylaws about beer brewing

*we ock vorschote ~~100~~ marck, de darf 3 warve bruwen*

together with 150 as a replacement for 100 and 2 as a replacement for 3. (The meaning of the sentence in Low Middle German is: one, who pays 100 (150) marks as taxes, is allowed to brew beer 3 (2) times a year.) Without additional information, the four following readings are allowed, all representing different stages of the textual development:

*R1: we ock vorschote 100 marck, de darf 3 warve bruwen*

*R2: we ock vorschote 100 marck, de darf 2 warve bruwen*

*R3: we ock vorschote 150 marck, de darf 3 warve bruwen*

*R4: we ock vorschote 150 marck, de darf 2 warve bruwen*

With some more information (mainly palaeographical) but still limited knowledge, three facts become clear: firstly, that R1 is the oldest version of the text, secondly that R4 is its most recent and thirdly that either R2 or R3 had existed as text layers or none of them but not both. But what was, however, the development of this sentence? Was it the path directly from R1 to R4? Or do we have to consider R1 > R2 > R4 or R1 > R3 > R4? In order to answer these questions we need to know about the context of the text, something that can not be found in the text itself. It is the external, procedural and contextual knowledge that has to be linked to the textual expression in order to fully analyse and edit the text.

Textual mobility in this example means that, to a certain extent, the textual expression itself, its sequence of graphemes, can be regarded as invariant and objective, the external knowledge about its context cannot. It is essential in our case study not only to distinguish between the expression and the context of the text but what is more to allow flexibility in the definition and reading of (possible) text layers. It became soon clear, that for both, visualising and processing a dynamic text, a new understanding of an edition is needed, and, as a consequence, the mark-up strategy has to be reconsidered. This new understanding would "promote" the reader of an edition to its user, by making him part of it in a way that his external knowledge, his contextual setting would have influence on the representation of the text. Or in other words: dynamic text requires dynamic representation.

The way chosen in this study is to regard textual expression and context (external knowledge) separately. The expression is represented by mark-up, encoding the information about the text itself. Regarding this stand-alone, the different units of the text (its oldest version, its later alterations or annotations) could indeed be visualised but not be brought into a meaningful relationship to each other. The latter is realised by a database providing structured external information about the text, mainly what specific "role" a certain part of the text "plays" in the context of interest. Only managing and processing both, markup and database, will allow to reconstruct the different stages of the text and consequently to represent the town law in its evolutionary development.

Using the linkage mechanism between mark-up and database, the whole set of information is processable. In order to create a scholarly edition of the text, we can automatically produce a document that fulfils TEI conformity to allow the use of the widely available tools for transformation, further processing and possibly interchange.

## II.

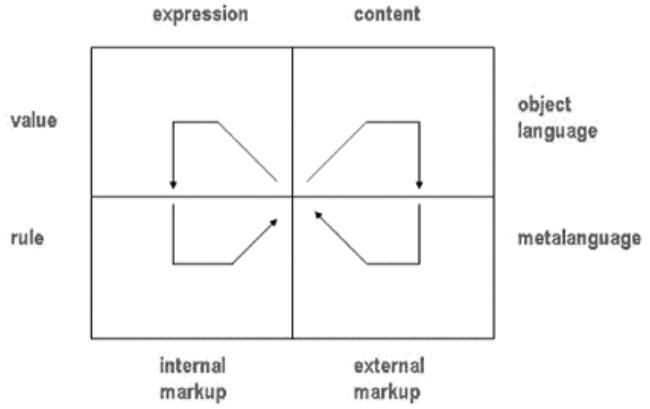
The case study just examined shows that in order to render an edition processable we have to relate the management system of the relevant data model to the markup embedded in the text. But we cannot provide a complete declarative model of the mapping of syntactic markup structures onto semantic content structures. The markup cannot contain a complete content model, just as a content model cannot contain a complete and totally definite expression of the text. To prove this fact we have to show that a markup description is equivalent to a second-order object language self-reflexive description and recall that a second-order logical theory cannot be complete. So the mapping cannot be complete, but for the same reason it can be categorical; in other words, all the models of the text could be isomorphic. So we can look for general laws, but they can provide only a dynamic procedural model.

Let us briefly outline the steps that lead to this result. In a significant contribution to the understanding of "the meaning of the markup in a document," [Sperberg-McQueen, Huitfeldt, and Renear, 2000: 231] expound it as "being constituted," and "not only described," by "the set of inferences about the document which are licensed by the markup." This view has inspired the BECHAMEL Markup Semantics Project, a ground breaking attempt to specify mechanisms "for bridging [...] syntactic relationships [...] with the distinctive semantic relationships that they represent" [Dubin and Birnbaum, 2004], and to investigate in a systematic way the "mapping [of] syntactic markup structures [on]to instances of objects, properties, and relations" [Dubin, 2003] that could be processed trough an appropriate data model. Following [Dubin and Birnbaum, 2004], "that markup can communicate the same meaning in different ways using very different syntax", we must conclude that "there are many ways of expressing the same content, just as there are many ways of assigning a content to the same expression" [Buzzetti, forthcoming].

The relationship between expression and content is then an open undetermined relationship that can be formalized by taking into account the "performative mood" of markup [Renear, 2001: 419]. For, a markup element, or any textual mark for that matter, is ambivalent: it can be seen as part of the

text, or as a metalinguistic description/ indication of a certain textual feature. Linguistically, markup behaves as punctuation, or as any other diacritical mark, i.e. as the expression of a reflexive metalinguistic feature of the text. Formally, markup behaves just as Spencer-Brown's symbols do in his formal calculus of indications [1969]: a symbol in that calculus can express both an operation and its value [Varela, 1979: 110-111].

Markup adds structure to the text, but it is ambivalent. It can be seen as the result of a restructuring operation on the expression of the text (as a textual variant) or as an instruction to a restructuring operation on the content of the text (as an interpretational variant). By way of its ambivalence it can act as a conversion mechanism between textual and interpretational variants [Buzzetti and McGann, 2006: 66] [Buzzetti, forthcoming].




Markup originates a loop connecting the structure of the text's expression to the structure of the text's content. An implementation of the *markup loop* would considerably enhance the functionality of text representation and processing in a digital edition. To achieve implementation, markup information could be integrated into the object (or datatype) 'string' on which an application system operates. Extended strings, as a datatype introduced by Manfred Thaller [1996, 2006], look as a suitable candidate for the implementation of the markup loop.

Markup originates a loop connecting the structure of the text's expression to the structure of the text's content. An implementation of the markup loop would considerably enhance the functionality of text representation and processing in a digital edition. To achieve implementation, markup information could be integrated into the object (or datatype) 'string' on which an application system operates. Extended strings, as a datatype introduced by Manfred Thaller [1996, 2006], look as a suitable candidate for the implementation of the markup loop.

## Bibliography

[Buzzetti, 1992] Buzzetti, Dino, Paolo Pari e Andrea Tabarroni. 'Libri e maestri a Bologna nel xiv secolo: Un'edizione come database,' *Schede umanistiche*, n.s., 6:2 (1992), 163-169.

[Buzzetti, 2002] Buzzetti, Dino. 'Digital Representation and the Text Model,' *New Literary History*, 33:1 (2002), 61-87.

[Buzzetti, 2004] Buzzetti, Dino. 'Diacritical Ambiguity and Markup,' in D. Buzzetti, G. Pancaldi, and H. Short (eds.), *Augmenting Comprehension: Digital Tools and the History of Ideas*, London-Oxford, Office for Humanities Communication, 2004, pp. 175-188: URL = <http://137.204.176.111/dbuzzetti/pubblicazioni/ambiguity.pdf>

[Buzzetti and McGann, 2006] Buzzetti, Dino, and Jerome McGann. 'Critical Editing in a Digital Horizon,' in *Electronic Textual Editing*, ed. Lou Burnard, Katherine O'Brien O'Keeffe, and John Unsworth, New York, The Modern Language Association of America, 2006, pp. 51-71.

[Buzzetti, forthcoming] Buzzetti, Dino. 'Digital Editions and Text Processing,' in *Text Editing in a Digital Environment*, Proceedings of the AHRC ICT Methods Network Expert Seminar (London, King's College, 24 March 2006), ed. Marilyn Deegan and Kathryn Sutherland (Digital Research in the Arts and Humanities Series), Aldershot, Ashgate, forthcoming.

[Dubin, 2003] Dubin, David. 'Object mapping for markup semantics,' *Proceedings of Extreme Markup Languages 2003*, Montréal, Québec, 2003: URL = <http://www.idealliance.org/papers/extreme/proceedings//html/2003/Dubin01/EML2003Dubin01.html>

[Dubin and Birnbaum, 2004] Dubin, David, and David J. Birnbaum. 'Interpretation Beyond Markup,' *Proceedings of Extreme Markup Languages 2004*, Montréal, Québec, 2004: URL = <http://www.idealliance.org/papers/extreme/proceedings//html/2004/Dubin01/EML2004Dubin01.html>

[McGann, 1991] McGann, Jerome. *The Textual Condition*, Princeton, NJ, Princeton University Press, 1991.

[McGann, 1999] McGann, Jerome. 'What Is Text? Position statement,' *ACH-ALLC'99 Conference Proceedings*, Charlottesville, VA, University of Virginia, 1999: URL = <http://www.iath.virginia.edu/ach-allc.99/proceedings/hockey-renear2.html>

[Rehbein, forthcoming] Rehbein, Malte. Reconstructing the textual evolution of a medieval manuscript.

[Rehbein, unpublished] Rehbein, Malte. Göttinger Burspraken im 15. Jahrhundert. Entstehung – Entwicklung – Edition. PhD thesis, Univ. Göttingen.

[Renear, 2001] Renear, Allen. 'The descriptive/procedural distinction is flawed,' *Markup Languages: Theory and Practice*, 2:4 (2001), 411–420.

[Sperberg-McQueen, Huitfeldt and Renear, 2000] Sperberg-McQueen, C. M., Claus Huitfeldt, and Allen H. Renear. 'Meaning and Interpretation of Markup,' *Markup Languages: Theory and Practice*, 2:3 (2000), 215-234.

[Thaller, 1991 ] Thaller, Manfred. 'The Historical Workstation Project,' *Computers and the Humanities*, 25 (1991), 149-62.

[Thaller, 1996] Thaller, Manfred. 'Text as a Data Type,' *ALLC-ACH'96: Conference Abstracts,* Bergen, University of Bergen, 1996.

[Thaller, 2006] Thaller, Manfred. 'Strings, Texts and Meaning.' *Digital Humanities 2006: Conference Abstracts*, Paris, CATI - Université Paris-Sorbonne, 2006, 212-214.

# Reflecting on a Dual Publication: Henry III Fine Rolls Print and Web

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

***Tamara Lopez***
*tamara.lopez@kcl.ac.uk*
*King's College London, UK*

A collaboration between the National Archives in the UK, the History and Centre for Computing in the Humanities departments at King's College London, the Henry III Fine Rolls project (http://www.frh3.org.uk) has produced both a digital and a print edition (the latter in collaboration with publisher Boydell & Brewer) [1] of the primary sources known as the Fine Rolls. This dual undertaking has raised questions about the different presentational formats of the two resources and presented challenges for the historians and digital humanities researchers involved in the project, and, to a certain extent, for the publisher too.

This paper will examine how the two resources evolved: the areas in which common presentational choices served both media, and areas in which different presentational choices and production methodologies were necessary. In so doing, this paper aims to build a solid foundation for further research into the reading practices and integrated usage of hybrid scholarly editions like the Henry III Fine Rolls.

## Presentation as interpretation

In Material Culture studies and, in particular, in studies of the book, the presentational format of text is considered to be of fundamental importance for the study of production, social reading and use. Therefore, description of and speculation about the physical organisation of the text is essential of understanding the meaning of the artefact that bears that text. Similarly, in Human Computer Interaction studies and in the Digital Humanities, the presentation of a text is considered to be an integral outgrowth of the data modelling process; a representation of the text but also to some degree an actualisation of the interpretative statements about the text. Indeed, to the eyes of the reader, the presentational features of both a printed book and a digital written object will not only reveal the assumptions and beliefs of its creators, but affect future analysis of the work.

## Dual publication: digital and print

On the practical side, within the Henry III Fine Rolls project, different solutions of formatting for the two media have been negotiated and implemented.

The print edition mainly represents a careful pruning of the digital material, especially as pertains to the complex structure of the indexes.

The indexes were built upon the marked-up fine rolls texts and generated from an ontological framework (Ciula, Spence, Vieira, & Poupeau; 2007). The latter, developed through careful analysis by scholars and digital humanities researchers, constitutes a sort of an *a posteriori* system that represents familial networks, professional relationships, geo-political structures, thematic clusters of subjects, and in general various types of associations between the 13th century documents (the so called Fine Rolls) and the the roles played by places and people in connection with them.

The ontology is used to produce a series of pre-coordinate axes (the indices) that the reader can follow to explore the texts. The flexibility of the ontology allows the texts to be fairly exhaustively indexed, just as the presentational capabilities of the digital medium allow for the display and navigation of indexes that are correspondingly large.

By contrast, the print edition had to follow the refined conventions of a well established scholarly tradition in publishing editions in general and calendar [2] editions in particular, both in terms of formatting and, more importantly for us, in terms of content selection/creation and modelling.

Though the indices within the printed edition are also pre-coordinate axes along which to explore the text, the way in which they are produced is perceived to be a nuanced and intuitive aspect of the scholarship and one that revealed itself to be less tolerant to change. This, coupled with the presentational constraints of the printed medium result in indices that present information succinctly and with a minimum of conceptual repetition. Similarly, the first print volume of around 560 pages gives absolute prominence -something that can be stated much more strongly in a linear publication than in a digital one- to a long and detailed historical introduction, followed by a section on the adopted editorial strategies.

However, the two artefacts of the project also share many points in common, either because the digital medium had to mirror the tradition of its more authoritative predecessor, or for more practical -nevertheless not to be dismissed- reasons of work-flow and foreseen usage. An interesting example of the latter is the adopted layout of footnotes, where the print format was modelled on the base of the digital layout and, although it was a completely unusual arrangement, was accepted as suitable by the publisher.

On the base of the work done so far and on the feedback on the use of the book and the website, the presentational format will be refined further for future print volumes to come and for the additional material to be included in the digital edition before the end of the project.

## One reading process

On the methodological side, we believe that further research into the usage and reading process of these parallel publications could lead towards a better understanding of scholarly needs and therefore a better modelling of such a dual product that is becoming a more and more common deliverable in digital humanities projects.

As this paper will exemplify, the presentation of data needs to be tailored to take into account the more or less fine conventions of two different media which have different traditions, different life cycles, different patterns of use and, possibly, different users.

However, although very different in nature, these two publications are not necessarily perceived and – more importantly- used as separate resources with rigid boundaries between them. For a scholar interested in the fine rolls, the reading of the edition and the seeking of information related to it (persons, places, subjects and any other interesting clue to its historical study in a broader sense) is a global process that does not stop when the book is closed or the the browser shut. We believe that, when supported by a deep interest in the material, the connection between the two publications is created in a rather fluid manner.

The reality of the reading process and information seeking, as it happens, is influenced by the products it investigates, but ultimately has a form of its own that is different from the objects of analysis. It is dynamic and heterogeneous, it leaves on the integration between different types of evidence, no matter what their format is, including other kind of external sources. Indeed, the library or archive is the most likely environment where a scholar of the fine rolls would find herself browsing the print or digital edition, eventually the original primary sources or their digital images, plus any range of secondary sources.

## Studying the integration of print and digital

The data behind the two publications are drawn from the same informational substrate, but are separated to create two presentational artefacts. As established, *reading* is expected to be the primary activity performed using both and a stated design goal for the project is that the two artefacts will form a rich body of materials with which to conduct historical research. The heterogeneity of the materials, however, suggests that working with texts will of necessity also involve periods of *information seeking*: moments while reading that give rise to questions which the material at hand cannot answer and the subsequent process embarked upon in order to answer them. Our working hypothesis is that to fill these information gaps (Wilson, 1999), the reader will turn to particular texts in the alternative medium to find answers, moving between the website and the books, fluctuating between states of reading and seeking.

Thus, the analytical stream in this paper will move from the practices of creating two types of resources to establishing an analytical framework for evaluating their use. Situating the project materials and domain experts within the literature of information behaviour research, we will identify and develop a model for evaluating how well the features of the website and the book support information seeking activities that bridge (Wilson, 1999) reading within the individual media.

## Conclusions

Based on our experience in creating a hybrid edition for the Henry III Fine Rolls project, the challenges and adopted solutions for the two types of published resources are a starting point from which to reflect on the integrated production of a dual object. At the same time, continuing work begun elsewhere in the digital humanities (Buchanan, Cunningham, Blandford, Rimmer, & Warwick; 2006) to adapt methodologies used in Information Science and Book Studies, a rationale and method for the design of an analysis of their use and, in particular, of the interaction between scholars and the website/books can be outlined.

## Notes

[1] The first volume was published in September 2007 (Dryburgh et al. 2007).

[2] Calendar stays here for an an English summary of the Latin records.

## Bibliography

Buzetti, Dino and Jerome McGann (2005) "Critical Editing in a Digital Horizon". In Burnard, O'Keeffe, and Unsworth eds. *Electronic Textual Editing*.

<http://www.tei-c.org/Activities/ETE/Preview/mcgann.xml>.

Buchanan, G.; Cunningham, S.; Blandford, A.; Rimmer, J. & Warwick, C. (2005), 'Information seeking by humanities scholars', *Research and Advanced Technology for Digital Libraries, 9th European Conference, ECDL*, 18--23.

Ciula, A. Spence, P., Vieira, J.M., Poupeau, G. (2007) *Expressing Complex Associations in Medieval Historical Documents: The Henry III Fine Rolls Project*. Paper presented at Digital Humanities 2007, Urbana-Champaign, 4-8 June, 2007. <http://www.digitalhumanities.org/dh2007/abstracts/xhtml.xq?id=196>

Dervin, B. (1983) "An overview of sense-making research: Concepts, methods, and results to date", *International Communication Association Annual Meeting, Dallas, TX*, 1983.

Drucker, Johanna (2003). "The Virtual Codex from Page Space to E-space." *The Book Arts Web* <http://www.philobiblon.com/drucker/>

Dryburgh, Paul and Beth Hartland eds. Arianna Ciula and José Miguel Vieira tech. Eds. (2007) *Calendar of the Fine Rolls of the Reign of Henry III [1216-1248]*, vol. I: 1216-1224, Woodbridge: Boydell & Brewer.

Lavagnino, John (2007). *Being Digital, or Analogue, or Neither*. Paper presented at Digital Humanities 2007, Urbana-Champaign, 4-8 June, 2007. <http://www.digitalhumanities.org/dh2007/abstracts/xhtml.xq?id=219>

McGann, Jerome. (1997) "The Rational of Hypertext." In Kathryn Sutherland ed. *Electronic text: investigations in method and theory*. Clarendon Press: Oxford: 19-46.

Siemens, Ray, Elaine Toms, Stéfan Sinclair, Geoffrey Rockwell, and Lynne Siemens. "The Humanities Scholar in the Twenty-first Century: How Research is Done and What Support is Needed." *ALLC/ACH 2004 Conference Abstracts*. Göteborg: Göteborg University, 2004. <http://www.hum.gu.se/allcach2004/AP/html/prop139.html>

Wilson, T. (1999) "Models in information behaviour research." *Journal of Documentation* 55(3), 249-270.

# Performance as digital text: capturing signals and secret messages in a media-rich experience

***Jama S. Coartney***
*jcoartney@virginia.edu*
*University of Virginia Library, USA*

***Susan L. Wiesner***
*slywiesner@virginia.edu*
*University of Virginia Library, USA*

## Argument

As libraries increasingly undertake digitisation projects, it behooves us to consider the collection/capture, organisation, preservation, and dissemination of all forms of documentation. By implication, then, these forms of documentation go beyond written text, long considered the staple of library collections. While several libraries have funded projects which acknowledge the need to digitise other forms of text, in graphic and audio formats, very few have extended the digital projects to include film, much less performed texts. As more performing arts incorporate born-digital elements, use digital tools to create media-rich performance experiences, and look to the possibility for digital preservation of the performance text, the collection, organisation, preservation, and dissemination of the performance event and its artefacts must be considered. The ARTeFACT project, underway at the Digital Media Lab at the University of Virginia Library, strives to provide a basis for the modeling of a collection of performance texts. As the collected texts document the creative process both prior to and during the performance experience, and, further, as an integral component of the performance text includes the streaming of data signals to generate audio/visual media elements, this paper problematises the capture and preservation of those data signals as artefacts contained in the collection of the media-rich performance event.

## Premise

In a report developed by a working group at the New York Public Library, the following participant thoughts are included:

> Although digital technologies can incorporate filmic ways of perceiving [the performing arts], that is the tip of the iceberg. It is important for us to anticipate that there are other forms we can use for documentation rather than limiting ourselves to the tradition of a camera in front of the stage. Documentation within a digital environment far exceeds the filmic way of looking at a performance' (Ashley 2005 NYPL Working Group 4, p.5)

> How can new technology both support the information we think is valuable, but also put it in a format that the next generation is going to understand and make use of?' (Mitoma 2005 NYPL Working Group 4, p.6)

These quotes, and many others, serve to point out current issues with the inclusion of the documentation of movement-based activities in the library repository. Two important library tasks, those of the organization and dissemination of text, require the development of standards for metadata. This requirement speaks towards the need for enabling content-based searching and dissemination of moving-image collections. However, the work being performed to provide metadata schemes of benefit to moving image collections most often refers to (a) a filmed dramatic event, e.g. a movie, and/or (b) metadata describing the film itself. Very little research has been completed in which the moving image goes beyond a cinematic film, much less is considered as one text within a multi-modal narrative.

In an attempt to address these issues, the authors developed the ARTeFACT project in hopes of creating a proof-of-concept in the University of Virginia Library. Not content, however, to study the description of extant, filmic and written texts, the project authors chose to begin with describing during the process of the creation of the media-rich, digital collection, including the description of a live performance event. The decision to document a performance event begged another set of answers to questions of issues involved with the collection of texts in a multiplicity of media formats and the preservation of the artefacts created through a performance event.

Adding to this layer of complexity was the additional decision to create not just a multi-media performance, but to create one in which a portion of the media was born-digital during the event itself. Created from signals transmitted from sensor devices (developed and worn by students), the born-digital elements attain a heightened significance in the description of the performance texts. After all, how does one capture the data stream for inclusion in the media-rich digital collection?

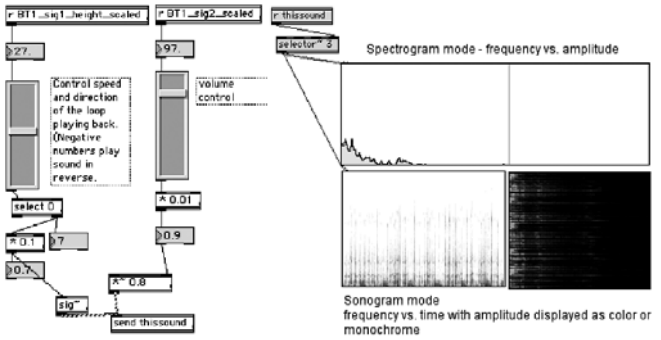
## Methodology

The ARTeFACT project Alpha includes six teams of students in an Introductory Engineering Class. Each team was asked to design and build an orthotic device that, when worn, causes the wearer to emulate the challenges of walking with a physical disability (included are stroke 'drop foot' and paralysis, CP hypertonia, Ricketts, etc.) During the course of the semester, the student teams captured pre-event process in a variety of digital formats: still and video images of the prototypes from cameras and cell phones, PDF files, CAD drawings, PowerPoint files and video documentation of those presentations. The digital files were collected in a local SAKAI implementation.

In addition to these six teams, two other teams were assigned the task of developing wireless measurement devices which, when attached to each orthotic device, measures the impact

of the orthotic device on the gait of the wearer. The sensor then transmits the measurement data to a computer that feeds the data into a Cycling74's software application: Max/MSP/Jitter. Jitter, a program designed to take advantage of data streams, then creates real-time data visualization as output. The resultant audio visual montage plays as the backdrop to a multi-media event that includes student 'dancers' performing while wearing the orthotic devices.

The sensors generate the data signals from Bluetooth devices (each capable of generating up to eight independent signals) as well as an EMG (electro-myogram) wireless system. At any given time there may be as many as seven signalling devices sending as many as 50 unique data streams into the computer for processing. As the sensor data from the performers arrives into Max/MSP/Jitter, it is routed to various audio and video instruments within the application, processed to generate the data visualization, then sent out of the computer via external monitor and sound ports. The screenshot below displays visual representations of both the input (top spectrogram) and output (bottom: sonogram) of a real-time data stream. The data can be visualized in multiple ways; however, the data stream as we wish to capture it is not a static image, but rather a series of samples of data over time.




There are several options for capturing these signals, two of which are: writing the data directly to disk as the performance progresses and/or the use of external audio and video mixing boards that in the course of capturing can display the mix.

Adding to the complexity, the performance draws on a wide variety of supporting, media rich, source material created during the course of the semester. A subset of this material is extrapolated for use in the final performance. These elements are combined, processed, morphed, and reformatted to fit the genre of the presentation and although they may bear some similarity to the original material, they are not direct derivatives of the source and thus become unique elements in the production. Further, in addition to the capture of data streams, the totality of the performance event must be collected. For this, traditional means of capturing the performance event have been determined to be the simplest of the challenges faced. Therefore, a video camera and a microphone pointed at the stage will suffice to fill this minimum requirement for recording the event.

## Conclusion

The complexity of capturing a performance in which many of the performance elements themselves are created in real time, processed, and used to generate audio visual feedback is challenging. The inclusion of data elements in the artefact collection begs questions with regard to the means of capturing the data without impacting the performance. So, too, does it require that we question what data to include: Does it make sense to capture the entire data stream or only the elements used at a specific instance in time to generate the performance; what are the implications of developing a sub-system within the main performance that captures this information? When creating a collection based on a performance as digital text, and before any work may be done to validate metadata schemes, we must answer these questions. We must consider how we capture the signals and interpret the secret messages generated as part of the media-rich experience.

## Sample bibliography

Adshead-Lansdale, J. (ed.) 1999, *Dancing Texts: intertextuality and interpretation* London: Dance Books.

Goellner, E. W. & Murphy, J. S. 1995, *Bodies of the Text* New Brunswick, NJ: Rutgers University Press.

Kholief, M., Maly, K. & Shen, S. 2003, 'Event-Based Retrieval from a Digital Library containing Medical Streams' in *Proceedings of the 2003 Joint Conference on Digital Libraries* (Online) Available at http://ieeexplore.ieee.org/Xplore/login.jsp?url=/iel5/8569/27127/01204867.pdf

New York Public Library for the Performing Arts, Jerome Robbins Dance Division 2005, Report from Working Group 4 *Dance Documentation Needs Analysis* Meeting 2, NYPL: New York.

Reichert, L. 2007, 'Intelligent Library System Goes Live as Satellite Data Streams in Real-Time' (Online) Available at http://www.lockheedmartin.com/news/press_releases/2001/IntelligentLibrarySystemGoesLiveAsS.html

# Function word analysis and questions of interpretation in early modern tragedy

***Louisa Connors***
*Louisa.Connors@newcastle.edu.au*
*University of Newcastle, Australia*

The use of computational methods of stylistic analysis to consider issues of categorization and authorship is now a widely accepted practice. The use of computational techniques to analyze style has been less well accepted by traditional humanists. The assumption that most humanists make about stylistics in general, and about computational stylistics in particular, is that it is "concerned with the formal and linguistic properties of the text as an isolated item in the work" (Clark 2005). There are, however, two other points of emphasis that are brought to bear on a text through a cognitive approach. These are: "that which refers to the points of contact between a text, other texts and their readers/listeners", and "that which positions the text and the consideration of its formal and psychological elements within a socio-cultural and historical context" (Clark 2005). Urszula Clark (2005) argues that these apparently independent strands of analysis or interpretive practice are an "integrated, indissolvable package" (Clark 2005).

Computational stylistics of the kind undertaken in this study attempts to link statistical findings with this integrated indissolvable package; it highlights general trends and features that can be used for comparative purposes and also provides us with evidence of the peculiarities and creative adaptations of an individual user. In this case, the individual user is Elizabeth Cary, the author of the earliest extant original play in English by a woman, *The Tragedy of Mariam, The Fair Queen of Jewry* (1613). As well as *Mariam*, the set of texts in the sample includes the other 11 closet tragedies associated with the "Sidney Circle", and 48 tragedies written for the public stage. All plays in the study were written between 1580 and 1640.[1] The only other female authored text in the group, Mary Sidney's *Antonius*, is a translation, as is Thomas Kyd's *Cornelia*[2]. Alexander Witherspoon (1924), describes the Sidnean closet tragedies as "strikingly alike, and strikingly unlike any other dramas in English" (179). He attributes this to the extent to which the closet writers draw on the work of French playwright Robert Garnier as a model for their own writing. In contrast to other plays of the period closet tragedies have not attracted much in the way of favourable critical attention. They are, as Jonas Barish (1993) suggests, "odd creatures" (19), and Mariam is described as one of oddest.

*Mariam*, as it turns out, is the closet play that is *most* like a play written for the public stage, in terms of the use of function words. But this isn't immediately obvious. Some textual preparation was carried out prior to the analysis. Homographs were not tagged, but contracted forms throughout the texts were expanded so that their constituents appeared as separate words. The plays were divided into 2,000 word segments and tagged texts were then run through a frequency count using *Intelligent Archive* (IA). A total of 563 two-thousand word segments were analysed, 104 of which were from closet plays, and 459 from plays written for the public stage. A discriminant analysis on the basis of the frequency scores of function words demonstrates that there are significant differences between the two groups of plays. Table 1 shows the classification results for a discriminant analysis using the full set of function words. In this test, 561 of the 563 segments were classified correctly. One segment from each group was misclassified. Thus 99.6% of cross-validated grouped cases were correctly classified on the basis of function words alone. The test also showed that only 38 of the 241 function word variables were needed to successfully discriminate between the groups.

*Table 1*

*Classification results for discriminant analysis using the full set of function words in 60 tragedies (1580-1640) closet/non-closet value correctly assigned*

| | | Closet/Stage | Predicted Group Membership | | Total |
|---|---|---|---|---|---|
| | | | Closet | Public | |
| Original | Count | Closet | 103 | 1 | 104 |
| | | Public | 1 | 458 | 459 |
| | % | Closet | 99.0 | 1.0 | 100.0 |
| | | Public | .2 | 99.8 | 100.0 |
| Cross-validated (a) | Count | Closet | 103 | 1 | 104 |
| | | Public | 1 | 458 | 459 |
| | % | Closet | | 1.0 | 100.0 |
| | | Public | | 99.8 | 100.0 |

*a Cross validation is done only for those cases in the analysis. In cross validation, each case is classified by the functions derived from all cases other than that case.*

*b 99.6% of original grouped cases correctly classified.*

*c 99.6% of cross-validated grouped cases correctly classified.*

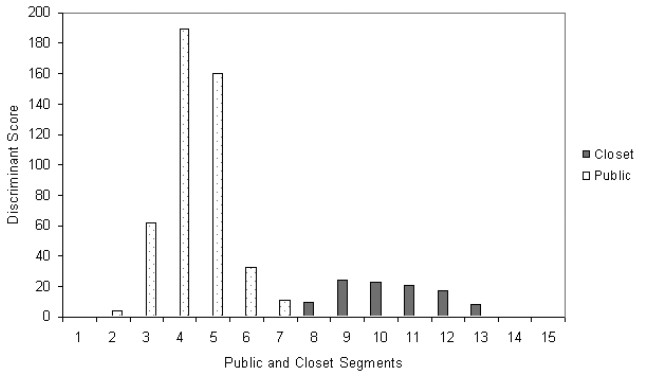



*Figure 1. Discriminant scores for correctly identified public and closet play segments from 60 tragedies (1580-1640) in 2000 word segments on the basis of 38 most discriminating function words*

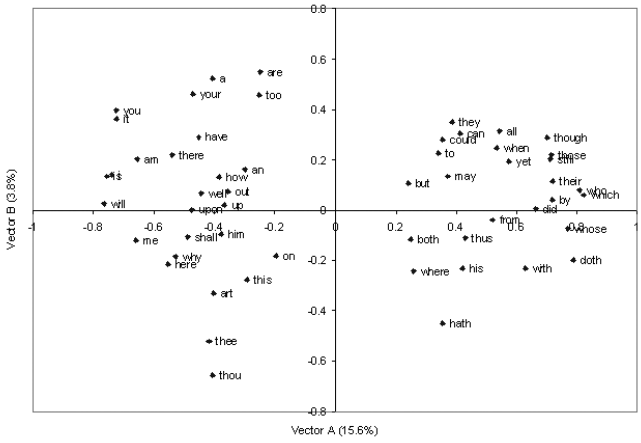



*Figure 2. Principal component analysis for 60 tragedies (1580-1640) in 4000 word segments for 54 most discriminating function words selected from 100 most frequently occurring function words - word plot for first two eigenvectors*

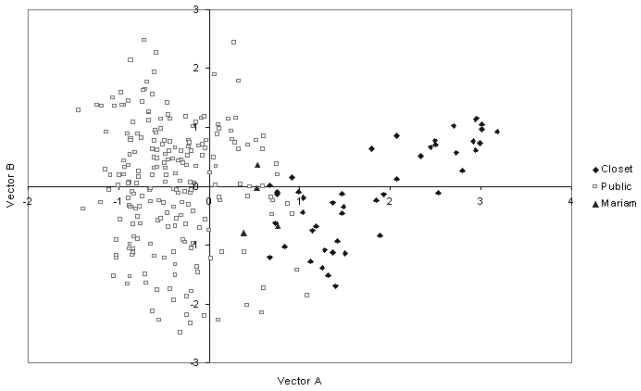



*Figure 3. Principal component analysis for 60 tragedies (1580-1640) in 4000 word segments for 54 most discriminating function words selected from 100 most frequently occurring function words*

A principal component analysis (PCA) gives additional information about the differences and similarities between the two sets. An Independent-samples T-test was used to identify the variables most responsible for the differences. This process picked out 54 variables that were significant at the level of 0.0001 and these 54 variables were used for the PCA. PCA looks to define factors that can be described as most responsible for differences between groups. For this text, the texts were broken into 4000 word segments to ensure that frequencies remained high enough for reliable analysis. This produced 268 segments in total (50 closet segments and 218 public play segments). Figure 2 plots the 54 most discriminating words-types for the first two eigenvectors (based on factor loadings for each variable) and shows which words behave most like or unlike each other in the sample.

In Figure 3 the eigenvalues from the component matrix have been multiplied through the standardised frequencies for each of the 4,000 word segments to show which segments behave most like or unlike each other on the first two principal components. The scores that produce Figure 3 are "the sum of the variable counts for each text segment, after each count is multiplied by the appropriate coefficient" (Burrows and Craig 1994 68). High counts on word variables at the western end of Figure 2 and low counts on word variables at the eastern end bring the text segments at the western end of Figure 3 to their positions on the graph. The reverse is true for text segments on the eastern side of Figure 3. We can see that the far western section of Figure 3 is populated predominantly with public play segments, and that the eastern side of the *y*-axis is populated exclusively with segments from stage plays. It is clear that in the case of these most discriminating variables, the segments from *Mariam* are the closet segments most intermingled with the segments written for the public stage.

Looking at Figure 2 there is evidence of an "emphasis on direct personal exchange" in western section of the graph. In the opposite section of the graph there is evidence of a more disquisitory style of language that is "less personal", with "markers of plurality, past time, and a more connected syntax" (Burrows and Craig 1994 70). It may be that the results reflect the kinds of observations that critics have long made about early modern closet tragedy and tragedy written for the public stage, suggesting that "word-counts serve as crude but explicit markers of the subtle stylistic patterns to which we respond when we read well" (Burrows and Craig 1994 70). It may also be the case, however, that we can link these statistical results with more interpretive work.

Returning to *Mariam*, the function words which most distinguish the *Mariam* segments from the rest of the segments of both closet and public plays, are the auxiliary verbs *did* (8.4) and *had* (6.9). In the case of *did* over 500 of the segments have scores of between -1 and 2. Six of the eight of the *Mariam* segments are extremely high (the two middle segments of *Mariam* have fairly average z-scores for *did*). The lowest score occurs in segment 4, when Mariam is conspicuously absent from the action. A very similar pattern emerges for *had*. In conventional grammars *do* and other auxiliaries including *be* and *have* are viewed as meaningless morphemes that serve a grammatical purpose. Langacker argues that serving a specifiable grammatical function is not inherently incompatible with being a meaningful element (1987 30).

Cognitive linguistics suggests that function word schemas interact with each other to produce what Talmy calls a "dotting" of semantic space (1983 226), and that they "play a basic conceptual structuring role" (Talmy 88 51). In this framework, auxiliaries are viewed as profiling a process – they determine which entity is profiled by a clause and impose a particular construal. Langacker argues, for example, that *do* always conveys some notion of activity or some kind of volitionality or control on the part of the subject. *Have* designates a subjectively construed relation of anteriority and current relevance to a temporal reference point (Langacker 1991 239). Talmy argues that *have* can be understood in terms of force dynamics patterns; it "expresses indirect causation either without an intermediate volitional entity…or… with

such an entity" (1988 65). Talmy goes further to suggest that the "concepts" of force dynamics are "extended by languages to their semantic treatment of *psychological* elements and interactions" (1988 69).

Bringing the tools of cognitive linguistics to bear on the results of computational analysis of texts can provide a framework that validates the counting of morphemes like did and had. The same framework may also shed light on questions of interpretation. This approach appears to provide a disciplined way of identifying and analyzing the linguistic features that are foregrounded in a text, while supporting their interpretation as part of an integrated, indissolvable package.

## Notes

1 Thomas Kyd wrote for both the public and the private stage. Kyd's Cornelia and The Spanish Tragedy are included in the study.

2 John Burrows (2002) explores some of the issues around translation and whether it can be "assumed that poets stamp their stylistic signatures as firmly on translation as their original work" (679). Burrows found that Drydan was able to "conceal his hand", but in other cases it appeared that a "stylisitic signature", even in the case of a translation, remained detectable (696).

## References

Barish, J. (1993). Language for the Study: Language for the Stage. In A. L. Magnusson & C. E. McGee (Eds.), *The ElizabethanTheatre XII* (pp. 19-43). Toronto: P.D. Meany.

Burrows, J. (2002). The Englishing of Jevenal: Computational stylistics and translated Texts. Style, 36, 677-750.

Burrows, J. F., & Craig, D. H. (1994). Lyrical Drama and the "Turbid Mountebanks": Styles of Dialogue in Romantic and Renaissance Tragedy. *Computers and the Humanities, 28*, 63-86.

Cooper, M. M. (1998). Implicature and *The Taming of the Shrew.* In J. Culpeper, M. H. Short & P. Verdonk (Eds.), *Exploring the Language of Drama: From Text to Context* (pp. 54-66). London: Routledge.

Clark, U. (2005). *Social cognition and the future of stylistics, or "What is cognitive stylistics and why are people saying such good things about it?!*" Paper presented at the PALA 25: Stylistics and Social Cognition.

Connors, L. (2006). An Unregulated Woman: A computational stylistic analysis of Elizabeth Cary's *The Tragedy of Mariam, The Faire Queene of Jewry. Literary and Linguistic Computing, 21* (Supplementary Issue), 55-66.

Sweetser, E. (1990). *From Etymology to Pragmatics: Metaphorical and cultural aspects of semantic change.* Cambridge: Cambridge University Press.

Langacker, R. W. (1999). Losing control: grammaticization, subjectification, and transparency. In A. Blank & P. Koch (Eds.), *Historical Semantics and Cognition* (Vol. 13, pp. 147-175). Berlin: Mouton de Gruyter.

Talmy, L. (1983). How language structures space. In H. Pick & L. Acredolo (Eds.), *Spatial orientation: Theory, research, and application.* New York: Plenum Press.

Talmy, L. (1988). Force Dynamics in Language and Cognition. *Cognitive Science, 12.* 49-100.

Witherspoon, A. M. (1924: rpt. 1968). *The Influence of Robert Garnier on Elizabethan Drama.*

# Deconstructing Machine Learning: A Challenge for Digital Humanities

***Charles Cooney***
*cmcooney@diderot.uchicago.edu*
*University of Chicago, USA*

***Russell Horton***
*russ@diderot.uchicago.edu*
*University of Chicago, USA*

***Mark Olsen***
*mark@barkov.uchicago.edu*
*University of Chicago, USA*

***Glenn Roe***
*glenn@diderot.uchicago.edu*
*University of Chicago, USA*

***Robert Voyer***
*rlvoyer@diderot.uchicago.edu*
*University of Chicago, USA*

Machine learning tools, including document classification and clustering techniques, are particularly promising for digital humanities because they offer the potential of using machines to discover meaningful patterns in large repositories of text. Given the rapidly increasing size and availability of digital libraries, it is clear that machine learning systems will, of necessity, become widely deployed in a variety of specific tasks that aim to make these vast collections intelligible. While effective and powerful, machine learning algorithms and techniques are not a panacea to be adopted and employed uncritically for humanistic research. As we build and begin to use them, we in the digital humanities must focus not only on the performance of specific algorithms and applications, but also on the theoretical and methodological underpinnings of these systems.

At the heart of most machine learning tools is some variety of classifier. Classification, however, is a fraught endeavor in our poststructuralist world. Grouping things or ideas into categories is crucial and important, but doing so without understanding and being aware of one's own assumptions is a dubious activity, not necessarily for moral, but for intellectual reasons. After all, hierarchies and orders of knowledge have been shown to be both historically contingent and reflections of prevailing power structures. Bowker and Star state that "classification systems in general... reflect the conflicting, contradictory motives of the sociotechnical situations that gave rise to them" (64). As machine learning techniques become more widely applied to all forms of electronic text, from the WWW to the emerging global digital library, an awareness of the politics of classification and the ordering of knowledge will become ever more important. We would therefore like to present a paper outlining our concerns about these techniques and their underlying technical/intellectual assumptions based on our experience using them for experimental research.

In many ways, machine learning relies on approaches that seem antithetical to humanistic text analysis and reading and to more general poststructuralist sensibilities. The most powerful and effective techniques rely on the abilities of systems to classify documents and parts of documents, often in binary oppositions (spam/not spam, male/female, etc). Features of documents employed in machine learning applications tend to be restricted to small subsets of available words, expressions or other textual attributes. Clustering of documents based on relatively small feature sets into a small and often arbitrary number of groups similarly tends to focus on broad patterns. Lost in all of these operations are the marginal and exceptional, rendered hidden and invisible as it were, in classification schemes and feature selection.

Feature set selection is the first necessary step in many text mining tasks. Ian Witten notes that in "many practical situations there are far too many attributes for learning schemes to handle, and some of them -- perhaps the overwhelming majority -- are clearly irrelevant or redundant" (286-7). In our work, we routinely reduce the number of features (words, lemmas, bigrams, etc) using a variety of techniques, most frequently by filtering out features which occur in a small subset of documents or instances. This selection process is further required to avoid "overfitting" a learner to the training data. One could build an effective classifier and train it using features that are unique to particular documents, but doing so would limit the general applicability of the tool. Attempting to classify French novels by gender of author while retaining the names of characters (as in Sand's novel, *Conseulo*) or other distinctive elements is very effective, but says little about gendered writing in 19th century France (Argamon et. al., *Discourse*). Indeed, many classification tasks may be successful using a tiny subset of all of the words in a corpus. In examining American and non-American Black Drama, we achieved over 90% accuracy in classifying over nearly 700 plays using a feature set of only 60 surface words (Argamon et. al., *Gender, Race*). Using a vector space similarity function to detect articles in the *Encyclopédie* which borrow significantly from the *Dictionnaire de Trévoux,* we routinely get impressive performance by selecting fewer than 1,000 of the 400,000 unique forms in the two documents (Allen et. al.). The requirement of greatly reductive feature set selection for practical text mining and the ability of the systems to perform effective classifications based on even smaller subsets suggests that there is a significant distance from the texts at which machine learning must operate in order to be effective.

Given the reductive nature of the features used in text mining tasks, even the most successful classification task tends to highlight the lowest common denominators, which at best may be of little textual interest and at worst extremely misleading, encouraging stereotypical conclusions. Using a decision tree to classify modern and ancient geography articles in the *Encyclopédie*, we found "selon" (according to) to be the primary distinction, reflecting citation of ancient

sources ("selon Pline"). Classification of Black Drama by gender of author and gender of speaker can be very effective (80% or more accuracy), but the features identified by the classifiers may privilege particular stereotypes. The unhappy relationship of Black American men with the criminal justice system or the importance of family matters to women are both certainly themes raised in these plays. Of course, men talk more of wives than women and only women tend to call other women "hussies," so it is hardly surprising that male and female authors/characters speak of different things in somewhat different ways. However, the operation of classifiers is predicated on detecting patterns of word usage which most distinguish groups and may bring to the forefront literary and linguistic elements which play a relatively minor role in the texts themselves. We have found similar results in other classification tasks, including gender mining in French literary works and *Encyclopédie* classifications.

Machine learning systems are best, in terms of various measures of accuracy, at binomial classification tasks, the dreaded "binary oppositions" of male/female, black/white and so forth, which have been the focus of much critical discussion in the humanities. Given the ability of statistical learners to find very thin slices of difference, it may be that any operation of any binary opposition may be tested and confirmed. If we ask for gender classification, the systems will do just that, return gender classifications. This suggests that certain types of hypothesis testing, particularly in regard to binary classifications, may show a successful result simply based on the framing of the question. It is furthermore unclear as to just what a successful classification means. If we identify gender or race of authors or characters, for example, at a better than 80% rate and generate a list of features most associated with both sides of the opposition, what does this tell us about the failed 20%? Are these errors to be corrected, presumably by improving classifiers or clustering models or should we further investigate these as interesting marginal instances? What may be considered a failure in computer science could be an interesting anomaly in the humanities.

Machine learning offers great promise to humanistic textual scholarship and the development of digital libraries. Using systems to sift through the ever increasing amounts of electronic texts to detect meaningful patterns offers the ability to frame new kinds of questions. But these technologies bring with them a set of assumptions and operations that should be subject to careful critical scrutiny. We in the digital humanities must do this critical work, relying on our understanding of epistemology and our technical skills to open the black box and shine light on what is inside. Deconstruction in the digital library should be a reading strategy not only for the texts found therein, but also of the systems being developed to manage, control and make the contents of electronic resources accessible and intelligible.

## Bibliography

Allen, Timothy, Stéphane Douard, Charles Cooney, Russell Horton, Robert Morrissey, Mark Olsen, Glenn Roe, and Robert Voyer. "Plundering Philosophers: Identifying Sources of the *Encyclopédie* using the Vector Space Model" in preparation for *Text Technology.*

Argamon, Shlomo, Russell Horton, Mark Olsen, and Sterling Stuart Stein. "Gender, Race, and Nationality in Black Drama, 1850-2000: Mining Differences in Language Use in Authors and their Characters." *DH07*, Urbana-Champaign, Illinois. June 4-8, 2007.

Argamon, Shlomo, Jean-Baptiste Goulain, Russell Horton, and Mark Olsen. "Discourse, Power, and *Ecriture Féminine*: Text Mining Gender Difference in 18th and 19th Century French Literature." *DH07*, Urbana-Champaign, Illinois. June 4-8, 2007.

Bowker, Geoffrey C. and Susan Leigh Star. *Sorting Things Out: Classification and its Consequences*. Cambridge, MA: MIT Press, 1999.

Introna, L. and H. Nissenbaum. "Shaping the Web: Why the Politics of Search Engines Matters." *The Information Society*, 16(3): 1-17, 2000.

Witten, Ian H. and Eibe Frank. *Data Mining: Practical Machine Learning Tools and Techniques,* 2nd ed. San Francisco, CA: Morgan Kaufmann, 2005.

# Feature Creep: Evaluating Feature Sets for Text Mining Literary Corpora.

***Charles Cooney***
*cmcooney@diderot.uchicago.edu*
*University of Chicago, USA*

***Russell Horton***
*russ@diderot.uchicago.edu*
*University of Chicago, USA*

***Mark Olsen***
*mark@barkov.uchicago.edu*
*University of Chicago, USA*

***Glenn Roe***
*glenn@diderot.uchicago.edu*
*University of Chicago, USA*

***Robert Voyer***
*rlvoyer@diderot.uchicago.edu*
*University of Chicago, USA*

Machine learning offers the tantalizing possibility of discovering meaningful patterns across large corpora of literary texts. While classifiers can generate potentially provocative hints which may lead to new critical interpretations, the features sets identified by these approaches tend not to be able to represent the complexity of an entire group of texts and often are too reductive to be intellectually satisfying. This paper describes our current work exploring ways to balance performance of machine learning systems and critical interpretation. Careful consideration of feature set design and selection may provide literary critics the cues that provoke readings of divergent classes of texts. In particular, we are looking at lexical groupings of feature sets that hold the promise to reveal the predominant "idea chunklets" of a corpus.

Text data mining and machine learning applications are dependent on the design and selection of feature sets. Feature sets in text mining may be described as structured data extracted or otherwise computed from running or unstructured text in documents which serves as the raw data for a particular classification task. Such features typically include words, lemmas, n-grams, parts of speech, phrases, named entities, or other actionable information computed from the contents of documents and/or associated metadata. Feature set selection is generally required in text mining in order to reduce the dimensionality of the "native feature space", which can range in the tens or hundreds thousands of words in even modest size corpora [Yang]. Not only do many widely used machine learning approaches become inefficient when using such high dimensional feature spaces, but such extensive feature sets may actually reduce performance of a classifier. [Witten 2005: 286-7] Li and Sun report that "many *irrelevant* terms have a detrimental effect on categorization accuracy due to *overfitting*" as well as tasks which "have many relevant but *redundant* features... also hurt categorization accuracy". [Li 2007]

Feature set design and selection in computer or information science is evaluated primarily in terms of classifier performance or measures of recall and precision in search tasks. But the features identified as most salient to a classification task may be of more interest than performance rates to textual scholars in the humanities as well as other disciplines. In a recent study of American Supreme Court documents, McIntosh [2007] refers to this approach as "Comparative Categorical Feature Analysis". In our previous work, we have used a variety of classifiers and functions implemented in PhiloMine [1] to examine the most distinctive words in comparisons of wide range of classes, such as author and character genders, time periods, author race and ethnicity, in a number of different text corpora [Argamon et. al. 2007a and 2007b]. This work suggests that using different kinds of features -- surface form words compared to bilemmas and bigrams -- allows for similar classifier performance on a selected task, while identifying intellectually distinct types of differences between the selected groups of documents.

Argamon, Horton et al. [Argamon 2007a] demonstrated that learners run on a corpus of plays by Black authors successfully classified texts by nationality of author, either American or non-American, at rates ranging between 85% and 92%. The features tend to describe gross literary distinctions between the two discourses reasonably well. Examination of the top 30 features is instructive.

**American**: ya', momma, gon', jones, sho, mississippi, dude, hallway, nothin, georgia, yo', naw, alabama, git, outta, y', downtown, colored, lawd, mon, punk, whiskey, county, tryin', runnin', jive, buddy, gal, gonna, funky

**Non-American**: na, learnt, don, goat, rubbish, eh, chief, elders, compound, custom, rude, blasted, quarrel, chop, wives, professor, goats, pat, corruption, cattle, hmm, priest, hunger, palace, forbid, warriors, princess, gods, abroad, politicians

Compared side by side, these lists of terms have a direct intuitive appeal. The American terms suggest a body of plays that deal with the Deep South (the state names), perhaps the migration of African-Americans to northern cities (hallway and downtown), and also contain idiomatic and slang speech (ya', gon', git, jive) and the language of racial distinction (colored). The non-American terms reveal, as one might expect, a completely different universe of traditional societies (chief, elders, custom) and life under colonial rule (professor, corruption, politicians). Yet a drawback to these features is that they have a stereotypical feel. Moreover, these lists of single terms reduce the many linguistically complex and varied works in a corpus to a distilled series of terms. While a group of words, in the form of a concordance, can show something quite concrete about a particular author's oeuvre or an individual play, it is difficult to come to a nuanced understanding of an entire

corpus through such a list, no matter how long. Intellectually, lists of single terms do not scale up to provide an adequate abstract picture of the concerns and ideas represented in a body of works.

Performing the same classification task using bilemmas (bigrams of word lemmas with function words removed) reveals both slightly better performance than surface words (89.6% cross validated) and a rather more specific set of highly ranked features. Running one's eye down this list is revealing:

**American**: yo_mama, white_folk, black_folk, ole_lady, st_louis, uncle_tom, rise_cross, color_folk,front_porch, jim_crow, sing_blue black_male, new_orleans, black_boy, cross_door, black_community, james_brown,

**Non-American:** palm_wine, market_place, dip_hand, cannot_afford, high_priest, piece_land, join_hand,bring_ water, cock_crow, voice_people, hope_nothing, pour_ libation, own_country, people_land, return_home

**American** (not present in non-American): color_boy, color_girl, jive_ass, folk_live

Here, we see many specific instances of African-American experience, community, and locations. Using bigrams instead of bilemmas delivers almost exactly the same classifier performance. However, not all works classified correctly using bilemmas are classified correctly using bigrams. Langston Hughes, *The Black Nativity,* for example, is correctly identified as American when using bigrams but incorrectly classified when using bilemmas. The most salient bigrams in the classification task are comparable, but not the same as bilemmas. The lemmas of "civil rights" and "human rights" do not appear in the top 200 bilemmas for either American or non-American features, but appear in bigrams, with "civil rights" as the 124th most predictive American feature and "human rights" as 111th among non-American features.

As the example of *The Black Nativity* illustrates, we have found that different feature sets give different results because, of course, using different feature sets means fundamentally changing the lexically based standards the classifier relies on to make its decision. Our tests have shown us that, for the scholar interested in examining feature sets, there is therefore no single, definitive feature set that provides a "best view" of the texts or the ideas in them. We will continue exploring feature set selection on a range of corpora representing different genres and eras, including Black Drama, French Women Writers, and a collection of American poetry. Keeping in mind the need to balance performance and intelligibility, we would like to see which combinations of features work best on poetry, for example, compared to dramatic writing. Text classifiers will always be judged primarily on how well they group similar text objects. Nevertheless, we think they can also be useful as discovery tools, allowing critics to find sets of ideas that are common to particular classes of texts.

## Notes

1. See http://philologic.uchicago.edu/philomine/. We have largely completed work on PhiloMine2, which allows the user to perform classification tasks using a variety of features and filtering on entire documents or parts of documents. The features include words, lemmas, bigrams, bilemmas, trigrams and trilemmas which can be used in various combinations.

## References

[Argamon 2007a] Argamon, Shlomo, Russell Horton, Mark Olsen, and Sterling Stuart Stein. "Gender, Race, and Nationality in Black Drama, 1850-2000: Mining Differences in Language Use in Authors and their Characters." *DH07*, Urbana-Champaign, Illinois. June 4-8, 2007.

[Argamon 2007b] Argamon, Shlomo, Jean-Baptiste Goulain, Russell Horton, and Mark Olsen. "Discourse, Power, and *Ecriture Féminine*: Text Mining Gender Difference in 18th and 19th Century French Literature." *DH07*, Urbana-Champaign, Illinois. June 4-8, 2007.

[Li 2007] Li, Jingyang and Maosong Sun, "Scalable Term Selection for Text Categorization", *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning* (EMNLP-CoNLL), pp. 774-782.

[McIntosh 2007] McIntosh, Wayne, "The Digital Docket: Categorical Feature Analysis and Legal Meme Tracking in the Supreme Court Corpus", *Chicago Colloquium on Digital Humanities and Computer Science,* Northwestern University, October 21-22, 2007

[Witten 2005] Witten, Ian H. and Eibe Frank. *Data Mining: Practical Machine Learning Tools and Techniques*, 2nd ed. San Francisco, CA: Morgan Kaufmann, 2005.

Yang, Yiming and Jan Pedersen, "A Comparative Study on Feature Selection in Text Categorization" In D. H. Fisher (ed.), *Proceedings of ICML-97, 14th International Conference on Machine Learning,* pp. 412-420, Nashville, US. Morgan Kaufmann Publishers, San Francisco, US.