Final Report

HD-50320-08

Encoding Names for Contextual Exploration in Digital Thematic Research Collections

Project Director: Julia Flanders
Brown University

April 30, 2010

Introduction

In recent years, increased interest in what is now termed "linked data"—and more generally in the contextualization of digital primary sources—has led to significant progress in the development of standards for representing and sharing information about named entities. A number of digital projects have also begun experimenting in this area, notably the Henry III Fine Rolls Project at Kings College London (http://www.frh3.org.uk/index.html). Technologies like RDF, the emergence of web service models for publishing authority data, and the inclusion of extended prosopographic encoding in the most recent release of the TEI Guidelines have opened up new arenas of development and experimentation.

With these possibilities in mind, the Brown University Women Writers Project (WWP) has developed the first phase of a detailed prosopography of the people named in our collection of pre-Victorian women's writing in English, Women Writers Online (WWO, http://www.wwp.brown.edu). Known informally as a "personography," it offers both an abstract information model designed to bring consistency to the representation of basic biographical data, and a practical solution to the problem of storing information about large numbers of people. For many years, the WWP has encoded its primary texts following the Text Encoding Initiative (TEI) Guidelines, and for this reason the WWP's personography also uses the TEI's mechanisms for encoding contextual information about people.[1] The TEI's approach to text encoding, and particularly its emphasis on standard methods for customizing the XML representation of textual information, gives us the ability to represent information about the people named in our texts in a highly consistent way while also affording us considerable latitude in tailoring our personography to the WWP's specific needs.

Over the course of this project (funded by the NEH through a Level II startup grant of $49,992 for a project running from July 2008 through December 2009), we have collected information on several thousand people named in 33 texts from Women Writers Online (roughly 10% of our total textbase). We have also built several prototype tools that generate visualization data from the resulting personography; these tools represent a few of the numerous possibilities for exposing this kind of information to readers alongside more traditional reading interfaces, and for using the visualization of contextual information to develop new research questions and methodologies.

The white paper submitted to the Office of Digital Humanities at the conclusion of the grant provides more detail on all of the activities described below.

---

[1] Information about the Text Encoding Initiative is available at http://www.tei-c.org/. The TEI's most recent guidelines significantly expand the range of available mechanisms for representing data associated with people and places. See Chapter 13, "Names, Dates, People, and Places," for a full description of the TEI's prosopographic encoding (http://www.tei-c.org/release/doc/tei-p5-doc/en/html/ND.html).

Project Activities and Accomplishments

The primary activities we undertook during the course of this project were as follows:

- We developed extensive personography of people (both historical, cultural, and fictional; both diegetic and production) referenced in the Women Writers Online collection.
- We developed internal specification for representing biographical information, using TEI guidelines; added a few things to the basic TEI approach: the idea of indicating the sphere of reference for each person (historical, fictional, mythological, folklore, classical, scriptural, etc.); their sphere of geographical reference; and their role in the textbase (as a producer or as a referent)
- We developed specifications and documentation for dealing with difficult cases: ambiguous reference, multiple reference, metaphoric reference.
- We developed a TEI schema customization that represents and constrains these dimensions of our data: to ensure consistency and to express our intentions in formal terms that can be shared and published.
- We developed an initial set of tool prototypes which are visible at the WWP sandbox: http://golf.services.brown.edu/sandbox/. These represent a set of initial possibilities for exposing our name data for exploration.
- We began building these same tools into the interface for the WWP exhibits and other public-facing materials at the WWP site.

Audiences

There are three important audiences for the results of this project:

- Readers of Women Writers Online: these are faculty and students at institutions that subscribe to WWO (or individual subscribers), who will encounter the personographic information and new tools we have developed as part of their interactions with WWO.
- Readers of public WWP materials, including exhibits, public visualizations and exploration tools to be made freely available at the WWP site. This is potentially a broader group since these materials will be freely available to the public as part of the next update to Women Writers Online. They are intended as a public-facing view of the WWO collection.
- Developers of digital humanities projects involving personographic (and other contextual) data in TEI. For this audience, probably the most important outcomes of the project are the white paper submitted to ODH (and also published from the WWP site) and the prototypes showing how our personographic data can be harnessed in visualizations using publicly accessible tools.

Evaluation

In the context of a prototype like this, there are a few different types of evaluation that are appropriate, some of which lie ahead of us:

- Formal data constraint: part of our work for this project was to take an initial set of name and biographical data that had been collected over a span of years, and bring it into clearer formal shape, by refining and expanding our ideas about what data we should be capturing, and then by expressing those ideas in formal terms through a schema. That schema constituted an ongoing evaluation tool through which we could gauge both the completeness and the internal consistency of the data set as it developed. For example, early on we imposed a validation constraint on the format of dates (e.g. birth, death, and floruit dates) and used the results to eliminate multiple formats for date ranges and questionable dates, bringing the data into an ISO-standard date format that can be used successfully in timelines and other date-oriented visualizations. Since in some cases the non-standard date format concealed either ambiguity or missing information, feedback from this evaluation step in many cases necessitated further research.
- User evaluation of the value of the visualizations we developed: although a full-scale evaluation of this kind will have to wait a few more months until our personographic data and visualizations are incorporated into the published version of Women Writers Online, we have been able to glean some informal feedback from the initial exposure these tools have had during seminars and conferences and to internal users. It is clear even at this early stage that these visualizations provide a much more immediate form of interaction with the data, and permit a kind of instant grasp of the relationship between time, space, individual human entities, and the textual landscape than was ever possible for us before. To take a simple example, one of our tools displays the names of both publishers and authors, arranged as a network of relationships. For each publisher, we can see which WWO authors he or she published; for each author, we can see which publishers she worked with. We can thus get an instant visual sense of which publishers were most widely used in the authoring community we are dealing with, and of what the communities of affinity were. By adding a temporal dimension (something we can do in the future) we could see further which publishers over time tended to support the community of women authors.

Long-term Impact

The long-term impact of this project will be felt in at least two ways. Perhaps the most concrete is the effect it will have on the reader community of the Women Writers Project: both subscribers to Women Writers Online and users of our publicly accessible materials. The presence of this added data about persons and their textual traces will help to embed these texts more firmly in cultural context

and make some aspects of textual interconnectedness more visible and more tractable to exploration. For students, these connections will help to diminish the distance between "expert" and "non-expert" readers and readings, reducing the effect of oracular isolation that sometimes accompanies unfamiliar texts from the remote past. For more advanced researchers, these connections may suggest contexts or interpretations that are unexpected: for instance, the fact that a given publisher had contact with a significant number of women writers may suggest a political inflection; the ability to visualize the interplay of classical and scriptural references may lead to further insight into how genre and female writerly authority are structured. Although there has been considerable discussion in the digital humanities world concerning new modes of reading, in fact the digital resources that are most firmly embedded in the humanities research landscape—resources like EEBO, ECCO, JSTOR—are all cast in the model of "search/retrieve/read". We hope with the work arising from this pilot project to demonstrate a much more responsive kind of resource.

Beyond our own application of this work, we hope that there will also be a long-term effect for other projects. As a next step, we plan two new initiatives: first, a larger contributory personography developed in partnership with other projects; and second, a workshop on visualization tools for humanists that would cover the kinds of basic APIs and data formats used in tools like Simile widgets, Google maps, and other visualization toolkits. The workshop would address a need we have identified in our NEH-funded advanced workshops on contextual information, which have brought us into contact with many nascent digital humanities projects that are developing personographic data but have no clear idea of how to harness it in their user interface. The shared personography would also support projects like these, who are eager to develop their own personographic data at scale but would like to avoid duplication of labor and of data.

Grant Products

The products for this grant are primarily infrastructural and documentary. Our white paper (submitted to the ODH) provides full detail on the outcome of our research on the encoding of names, personographic data, and name references in TEI, and also models for using it in a full-text resource like Women Writers Online. At the infrastructural level, we have created an initial personography of nearly 7500 names referenced in 33 texts from the WWO collection, and developed the basic architecture necessary to access these records via URI. Within a few months, as part of our next update of the WWO interface, we will begin exposing this data to WWO readers and the general public in several ways:

- Through visualizations that permit readers to examine the WWO "cast of characters" using maps, timelines, and networks, and also to explore relationships between people (for instance, the connections between authors and publishers)

- Through contextual notes that provide biographical information about individuals named in the texts (for instance, as a pop-up or marginal note)
- Through visualizations that form part of the interface for our publicly accessible exhibits: scholarly articles and explorations that introduce, comment on, and contextualize the WWO texts.
- Through visualizations that permit non-subscribers to explore the WWO textbase in non-textual ways: for instance, to discover how many 17th-century widows are represented among the authors, or what languages the authors could read.

Prototypes of these visualizations are now visible at the WWP sandbox: http://golf.services.brown.edu/sandbox/ and even in their current simple form they permit readers to get an overall view of the WWO collection from various perspectives: for instance, a view that shows all WWO authors plotted on a map by birthplace, with timeline sliders and other facets by which the reader can expand or contract the total set being viewed. Within a year we will have much more fully-featured versions of these tools built into WWO and also into the public interface.