Not Just One of Your Holiday Games:
Names and Name Encoding in the Women Writers Project Textbase

White Paper to the NEH Office of Digital Humanities
Level 2 Digital Humanities Start-Up Grant
January 2010

John Melson and Julia Flanders
Women Writers Project, Brown University
WWP@brown.edu

The Brown University Women Writers Project received Level II startup funding ($49,992) for a project running from July 2008 through December 2009. The grant supported the exploration of challenges in the detailed encoding of names and personographic data using the TEI Guidelines, with special attention to issues of identification, disambiguation, metaphoric reference, and other issues arising from a wide-ranging collection of literary texts. It also supported the development of a set of test data and experimental applications of visualization tools. The test data and tools are publicly visible at the WWP site (http://golf.services.brown.edu/sandbox/), and the specifications and documentation resulting from this work have been incorporated into the WWP's NEH-funded advanced seminar series on representing contextual information.

The Naming of Cats is a difficult matter,
It isn't just one of your holiday games…
— T. S. Eliot

Introduction

The appearance of personal names is a feature of literary texts that often goes unremarked. For scholars and casual readers alike, the names of people mentioned in a given text may contribute to the overall texture and richness of the work but rarely do they constitute the primary object of study. Instead they recede into the background, giving way to the formal and thematic features thought to be either more interesting or more instrumental in directing the text's social and cultural work. There is nothing particularly literary, after all, about lists of the kings of England, or the French generals at the Battle of Waterloo, or even, for that matter, the stock names so common to English pastoral verse in the seventeenth and eighteenth centuries. Yet personal names also offer a consistent method for accessing many aspects of literary texts that do interest scholars. The people a text names—whether fictional characters, historical figures, or Biblical prophets—provide valuable information about the cultural and imaginative spaces it occupies. How an author imagines herself in relation to other writers, departs from established generic patterns, or rewrites shared literary and cultural histories are all features that can be better understood through close attention to patterns of naming within and across texts. Once discoverable by readers in a systematic way, information about names can enable new approaches in numerous areas, such as the comparative study of literary genres, the social production of written texts, and patterns of literary influence and imitation.

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

---

[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).

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 in 2008-2009), 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.

This paper describes several aspects of the WWP's approach to gathering and representing contextual information, focusing in particular on our customization of the TEI's personography encoding, the challenges of representing personal information from the early modern period, and the tools we have begun to develop for interacting with this sort of data.

The Value of Personal Names

In any literary text, personal names play a number of crucial roles. Most obviously, names are the means by which readers identify and distinguish the people whose actions and thoughts influence the real or imagined events around which literary and historical narratives are constructed. This is true whether the people involved are fictional characters (as in novels or plays), historical figures (as in biographies, literary histories, and philosophical or scientific treatises), scriptural or mythological persons (often the case in sermons, religious prophecy, works of Biblical exegesis, or even pastoral and neoclassical verse), or possibly hybrids combining two or more of these qualities (historical figures treated fictionally, for instance). Projected outward into the social sphere of textual circulation, names also provide more complex information about the social and cultural production of the texts in which they appear. The people named in a text may reveal patterns of literary patronage or expose historical connections between the authors, printers, publishers, booksellers, subscribers, and readers who played a role in producing, distributing, and consuming these texts. Viewed in this broader cultural and historical context, names provide a crucial medium for exploring intertextual and paratextual relationships as well as the evolution of literary traditions.

Consider, for instance, the character of Titania mentioned in Penina Moise's *Fancy's Sketch Book*, a collection of poetry published in Charleston, South Carolina, in 1833. Moise's Titania bears no real resemblance to the Shakespearean character of the same name; she appears simply as consort to the fairy king when, in Moise's poem "The Fairy's Album," he is "seized with a mania / […] to compose" sonnets. But Titania's name still resonates culturally with her Shakespearean counterpart, establishing a subtle but nonetheless highly suggestive link between the two works. The fact that this Titania shares a name with Shakespeare's queen of the fairies marks Moise's own status within a literary realm where knowledge of Shakespearean drama confers cultural legitimacy. At the same time, the indirect reference also functions as a more self-conscious effort to appropriate and reimagine portions of a shared cultural heritage—a gesture of autonomy, so to speak, in which Moise, a Jewish woman living in the early nineteenth-century American South, declares a small measure of literary independence.

As this brief example suggests, names offer a rich and multi-layered context for exploring a wide range of issues that matter to contemporary literary and cultural studies. Careful observation of personal names at a larger scale—across dozens, hundreds, or possibly even thousands of texts—

promises us additional informational leverage on the texts involved, to say nothing of larger cultural narratives those texts have helped construct. Demarcating a variety of personal, professional, and cultural relations, names approached in this spirit can tell us a good deal about the socio-cultural contexts in which early moderns women's writing took place.

Treating Names as Data

Within the context of a digital collection, the intellectual motivations outlined above require realization in formal terms. To work with names as cultural and textual signifiers we must first identify them within the text flow, then express them (and any related information) as information objects. Only once this has been done can we build tools and systems that make use of those objects in meaningful ways. In an XML representation like TEI, the identification is accomplished in the markup of the text itself:

```
Printed by <persName>J. Orme</persName>
```

Additional information associated with this name (including the identity of the person named, and facts about their life, their role in the text, etc.) is represented as a separate information structure located outside the text:

```
<person xml:id="jorme.ysd" sex="1">
      <persName>
            <forename>James</forename>
            <surname>Orme</surname>
      </persName>
      <birth when="1661"/>
      <death when="1708"/>
</person>
```

In TEI, these two pieces of information—the name itself, and the information about the person it identifies—are then linked explicitly, with the `@ref` attribute of the `<persName>` element pointing to the unique identifier associated with the appropriate `<person>` element:

```
<!-- in the XML transcription of the text -->
Printed by <persName ref="personography.xml#jorme.ysd">J.
Orme</persName>

<!-- in the XML personography file, personography.xml -->
<person xml:id="jorme.ysd" sex="1">
      <persName>
            <forename>James</forename>
            <surname>Orme</surname>
      </persName>
      <birth when="1661"/>
      <death when="1708"/>
</person>
```

Taken together, these information structures create a mode of self-knowledge in the text that enables us to pursue the kinds of inquiries and explorations suggested above, despite the many complexities and challenges that the texts themselves pose.

One important application of this XML-enabled knowledge is simple disambiguation: the problem of identifying which person is designated by a given name in the text. For instance, the phrase "King George" in a British cultural context could refer to four distinct individuals within the period covered by the WWP's textbase. While context alone may be enough for a knowledgeable reader to determine which of those four people is meant in a particular text, that context is typically

unavailable in searches or other analysis. By linking each name in a text explicitly to an external name authority or biographical record as shown above, however, we make it possible for a reader to focus attention on a particular George: to search for all texts that refer to him, to explore the other persons connected with him, or to discover biographical information about him.

This disambiguating function (which George?) of encoding personal names is matched by a complementary aggregating function, through which the markup expresses the fact that many different forms of reference all identify the same person. Here the result is to enable the reader to deal directly with the referent's personhood, rather than with the complexities of how that personhood is referenced in the text through nicknames, titles, name variants, variant spelling, or name changes resulting from marriage or ennoblement. A search for Catherine de' Medici (whose name appears in the WWP textbase variously as Catherine de Medicis, Catharine de Medicis, Catherine de'Medici) or for Boudica (variously spelled in our collection Boadicea, Boadicia, Bouadicca, Bodicea) can better express the reader's intention if treated as a request for references to a specific person rather than the text strings that *might be* references to that person.

The value of this information is even greater when context is not (or not immediately) sufficient for readers to make the disambiguation themselves. For example, Anne Bradstreet's use of "Darius" in certain poems in *The Tenth Muse* (1650) does not refer to the same person Aemilia Lanyer refers to as "Darius" in *Salve Deus Rex Judaeorum* (1611), but few readers outside of a select group of specialists would know this from information contained in either text. The detailed encoding of personal information permits the correct association of the name Darius with the appropriate referent in each case and also gives the reader direct access to the information that supports this distinction. Through the explicit link from the encoded text to the personographic record, the reader can learn that Bradstreet's Darius is Darius III of Persia, while Lanyer's is Darius I. (There are added complications here, of course, involving the question of whether the authors in question know which Darius they are referring to: do they know that there are two historical figures named Darius, and do they correctly associate the reference with one of them in particular? In some cases, the reference may be to a quasi-fictional composite, and the representation of this is dealt with below.)

These forms of information support a much stronger set of reader tools: we might say that they extend the reach of the reader's intentions further into the text, by pushing back the point at which the text yields to ambiguity or inconsistency and can no longer participate usefully in formal queries. They also permit us to foreground views of the text that take advantage of this information. Even for simple display purposes, indexes of names belonging to a given text or set of texts serves a valuable scholarly purpose, permitting readers to get a sense of the relative importance of certain names in a particular text, genre, or historical moment. Combined with more interactive approaches to displaying personal names, in which a reader can filter and sort names based on a variety of demographic criteria (such as date or place of birth, sex, marital status, religious faith, or country of residence), this type of knowledge about the people named in a text helps nuance the reader's sense of the text's position vis-à-vis other works from the same period or of the same type.

Problems and Special Cases in Working with Early Modern Names

Unlike modern personal names, which tend to follow highly regular formal patterns, names in the early modern period in Europe were often highly variable. This is particularly true before 1700, when the spelling and formatting of names often depended on the writer and the context in which he or she wrote (legal documents and personal letters offered very different contexts for naming individuals, for example). Before 1600, even family titles and surnames could shift dramatically from

one generation to the next, sometimes due to titles conferred on individuals by the monarch or acquired through marriage but at other times due simply to deliberate or casual changes in spelling. For example, Hugh Despenser, 1st Lord Despencer (1286-1326), son of Hugh le Despenser, Earl of Winchester, is variously referred to in sixteenth- and seventeenth-century texts as "de Spencer", "de Spenser," "le Despenser," and "Despenser," minor variations that often seem the result of chance more than any consistent rationale. Given that such shifts were fairly common even among educated social and political elites, it is hardly surprising that the names of obscure individuals were subject to even greater inconsistencies. This variability adds a significant challenge to the creation of name authority records and to the task of ensuring that every person, however inconsistently her name appears in written documents, can be linked to the appropriate personographic record.

Other problems with the written representation of early modern names, particularly in literary texts, arise from the evolving conventions for addressing individuals of elevated social status and rank. Peers, for instance, are rarely addressed by name in texts from this period. Instead, formal titles are typically used in the place of names: the Duchess of Newcastle, Lord Touchet, Baron Lyttleton. These titles refer (in context) to a specific person, but because they are also often inherited there may be generational ambiguities to resolve. This leaves substantial room for error when determining who, exactly, is meant when a text refers to a person only by title. In such cases, the author's personal history and known relationships, as well as the dates of composition and publication, become crucial details in determining the identity of the named individual. This can be especially problematic when the transfer of a title (often resulting from the death a person and conferral of the title on his or her heir) coincides almost exactly with the publication of a text. The historical detective work needed to resolve such cases can make gathering accurate information about people named in a text a costly endeavor in terms of time and effort.

Compounding these problems is the convention in many early modern texts to disguise or deliberately obscure the names of real people, particularly those who may have been the author's contemporaries. Thus, for instance, texts published before 1800 often contain references like "Captain E-----," "Lady P----y," or "Mrs. S." When the person named is prominent, resolving the partially obscured name can be a fairly straightforward task ("Lady P----y" in at least one WWP text is, in fact, Lady Elizabeth Percy). In the WWP's collection, late eighteenth- and early nineteenth-century poets like Sarah Dixon, Sydney Owenson, Isabella Lickbarrow, and Mary Robinson frequently adopt this practice. The "Maria" named in Lickbarrow's "On the Approach of Winter. Addressed to a Friend" appears to be a real person, the "friend" mentioned in the poem's title. But is the eponymous "Anna" who appears in a different poem in *Poetical Effusions* (1814) also a real person, or is she an invention of Lickbarrow's? In Sarah Egerton's *Poems on Several Occasions* (1703), "Alexis" refers to the actual Henry Pierce, a law clerk who was possibly Egerton's lover. But does the fact that Egerton disguises identities in this way in some poems also mean that Orabella, the subject of "To Orabella, Marry'd to an Old Man," is a real person? Without expert knowledge of the poet's social circle and imagined readership, decoding such references remains a fraught process.

The WWP's textbase, like any collection of early modern writing, contains numerous other types of problematic or difficult names as well. Among these are ambiguous names (e.g. "the Persian Darius" with no further identifying information) or names that refer to multiple people at once (for instance, "the Misses Pinckney"). Metaphorical references and indirect allusions embedded within a name ("Sappho" used to refer both to Mary Robinson and to the ancient Greek poet Sappho of Lesbos) also fall into this category, though to varying degrees. A more extreme case, perhaps, is Phillis Wheatley's invocation of "Maecenas" in the dedication to *Poems on Various Subjects, Religious and Moral* (1773). Wheatley follows classical precedent, extending back to the works of Virgil and Horace, in

which those poets invoked their patron Gaius Clinius Maecenas; later poets frequently followed this tradition by referring to their own (often unnamed) patron as "Maecenas." In Wheatley's case, the invocation of "Maecenas" signals her own mastery of the classical tradition by referencing the historic person of that name while simultaneously acknowledging, albeit in a highly coded fashion, the patronage of her own "Maecenas," Selena Hastings, the Countess of Huntingdon.

Knowing in advance that the WWP's texts would include many such examples and that our personography would need to accommodate them, we created several mechanisms for representing them in our encoding. These mechanisms, as well as some of their present limitations, are described in greater detail in the sections that follow.

Multiple reference

One common phenomenon in printed English texts is the use of a single referring string to indicate multiple people—"G. and J. Robinson, for instance, or "the Misses Pinckney." Because they represent, in effect, a compressed form of reference (they are the functional equivalent of "George Robinson and John Robinson" and "Miss Pinckney and Miss Pinckney," respectively), such multiple references demand an encoding that records information about all of the people named and that associates the text string to the personographic entry for each of the unique individuals involved. Since the textual reference is a single string, however, and often does not lend itself to the use of multiple `<persName>` elements, the WWP has developed an encoding that uses the TEI `<link>` element as an intermediary between the encoded name(s) in the text and the multiple personography entries to which a single `<persName>` element might refer. Using established TEI mechanisms for linking and pointing, the `<link>` creates an intermediate layer for specifying the type of indirection and handling gracefully several references at once.

The following example from Sydney Owenson's *Poems* (1801) demonstrates one common way the WWP uses this method to resolve multiple name references:

Source text	Encoded text
The other soul, a poor inferior,	`<l>`The other soul, a poor inferior,`</l>`
And to the body scarce superior,	`<l>`And to the body scarce superior,`</l>`
From whence it steers its flight below,	`<l>`From whence it steers its flight below,`</l>`
To Messrs. Eachus and Co.	`<l>`To `<persName ref="#eachus">`Messrs. Eachus`</persName>` and Co.`</l>`

The `@ref` attribute on `<persName>` points to the out-of-line `<link>` element (located in a separate section of the encoded file) responsible for resolving the multiple reference:

```
<link xml:id="eachus" target="personography.xml#aeacus.sua
personography.xml#rhadamant.ezr personography.xml#minos.eyp"
type="multiple"/>
```

As this example illustrates, each value of `@target` points to the personography entry for one of the three mythological figures referenced in the phrase "Messrs. Eachus and Co." The `@type` attribute indicates the motivation for this pointing—in this case, recording the fact that each of these three people is being named within a single `<persName>` element in the encoded file. This encoding thus permits each individual person to be identified with the composite reference in the text, whether for purposes of searching, glossing, or visualization.

Ambiguous reference

Using the same mechanism, the WWP also resolves cases where names referenced in the text are ambiguous. For example, in the case of an eighteenth-century text that speaks of "Walpole" without providing sufficient context to determine whether it refers to Horace Walpole or his father, Robert Walpole, the WWP would also use `<link>` to record the ambiguity. In the encoding of the text:

```
<persName ref="#walpole">Walpole</persName>
```

And elsewhere in the file:

```
<link xml:id="walpole" target="personography.xml#rwalpole.ooa
personography.xml#hwalpole.ffm" type="ambiguous"/>
```

Here, once again, this approach to the encoding permits both Walpoles to be discovered as possible referents of the text string "Walpole," with a suitable flag given to the reader indicating the reference is uncertain or ambiguous.

Metaphorical and figurative reference

While ambiguous and multiple references can be represented using the `<link>` approach described above, a different mechanism is needed to represent the special nature of metaphorical or figurative references. Because such references do not represent multiple *actual* references but rather implied or *imagined* references, using `<link>` to point to multiple targets does not accurately model the intellectual work in the text in such cases. For this reason, the WWP has created a custom attribute for `<persName>`, `@metaRef`. For practical purposes, `@metaRef` functions exactly like `@ref`; where `@ref` points to the unique `@xml:id` for the actual reference, however, `@metaRef` points to the `@xml:id` of the person being indirectly or figuratively referenced.[2] For example:

<u>Source text</u>

Come all ye tender Nymphs and sighing Swains,

Hear how our Thyrsis, Daphnis death complains

<u>Encoded text</u>

```
<l>Come all ye tender Nymphs and sighing Swains,</l>

<l>Hear how our <persName ref="personography.xml#thyrsis.auc"
   metaRef="personography.xml#jfroud.jke">Thyrsis</persName>, <persName
   ref="personography.xml#daphnis.tvc"
   metaRef="personography.xml#tcreech.zxz">Daphnis</persName> death
   complains</l>
```

Though unwieldy from the perspective of a human reader, this encoding allows us to record the implied connection between the actual people referred to (in this case John Froud and Thomas Creech, respectively) and the two figures from Virgil's *Eclogues* to whom they are figuratively compared.

---

[2] Strictly speaking, the `@xml:id` is an attribute on a `<person>` element within the external personography file. The `<person>` element functions as the basic personographic unit, acting as a container for a set of other elements that record biographical and demographic data for that person.

In the rare instance where a single name in the text alludes metaphorically to multiple other people, `@metaRef` can be combined with the use of `<link>`. That is, just as `@ref` may point from a single `<persName>` to a `<link>` that does the work of resolving multiple direct references, the same is true for `@metaRef` when it comes to indirect references. It is worth noting, however, that the WWP has encountered no examples of this phenomenon in our texts to date, leading us to imagine such occurrences, while conceivable, are likely to be rare.

While ambiguous and multiple references constitute the bulk of difficult names in the WWP textbase, several other special cases also deserve attention. A number of our texts include names that exist in a liminal space somewhere between the world of the wholly fictional and the fully historical. Unlike coded or veiled names, the presence of these hybrid names is closely aligned with questions of genre; in our experience, such names are most highly concentrated in verse forms that follow recognized conventions of the English pastoral mode, where stock pastoral names (i.e. Lysander, Coridon, Cesario, etc.) are widely used. Indeed, these names are so common in certain texts that it is hard to know whether a particular poem that mentions Coridon uses the name simply because it signifies the poet's desire to invoke the pastoral mode, based on centuries of past use, or, alternatively, because it is an indirect reference to a *specific* pastoral poem the author wants to borrow from, rewrite, or otherwise acknowledge in some meaningful way. This uncertainty about generic or stock names is further compounded when they function as coded references to real people.

Another set of special cases encountered in the WWP textbase involve various forms of authorial error in naming historical figures—what we sometimes call the "conflation problem." This typically occurs when an author conflates—either inadvertently or, possibly, as a deliberate choice—aspects of two or more people, attributing them to a single person who is then named in the text. In *The Tenth Muse* (1650), for example, Ann Bradstreet claims Artaxerxes Memnon (Artaxerxes II of Persia) is the son of Artaxerxes Longimanus (Artaxerxes I of Persia). However, historians of the ancient world generally believe Artaxerxes II was the son of Darius II of Persia, not Artaxerxes I. In another text, *An Essay to Revive the Antient Education of Gentlewomen* (1673), Bathsua Makin conflates the early Christian martyr Tiburtius with his brother Valerianus (later St. Valerian), creating a composite person she calls Tiburtius Valerianus.

Both of these examples involve incomplete or inaccurate historical knowledge but pose real difficulties when it comes to accurately representing the nature of the error in our encoding. Are obvious errors (the attribution of one person's historical actions to another person) the result of a simple mistake on the author's part? Or might such errors indicate mistakes in the texts or sources the author relied on? Or, as a third possibility, could it be that a "fact" modern scholars now consider incorrect was widely or even universally believed to be true several centuries ago? In the latter instance, it would not be accurate, strictly speaking, to identify the author's conflation of two people as an *error*, since the knowledge expressed in the original text may have actually been "true" as far as the author and her contemporaries were concerned. The problem is even more intractable in self-consciously literary texts, where the "mistake" might be deliberate, perhaps introduced for satiric or ironic purposes. Because of the inherent complexity of such situations, the WWP has chosen to approach them conservatively: only when context makes it absolutely clear that an author means to name someone other than the person she actually names do we record the intended reference—as opposed to the actual printed reference—in our name encoding. Thus, if context makes it clear that a text means "Darius I" even when the name "Xerxes I" appears printed on the page, we provide the correct name encoding for Darius I; if there is any doubt at all, we privilege the printed name and assume the author intends to speak of Xerxes I.

The presence of common figures who appear in multiple texts as—potentially—different versions or instantiations of the same person pose another significant problem. In the cases the WWP has encountered, folkloric, mythological, and fictional people are often at the center of this kind of personal multiplication or "versioning." Mary Robinson, for instance, mentions Puck and Robin Goodfellow in one of her poems; while the poem may, at some level, attempt to link them to their Shakespearean instantiations, there is little evidence in the text to suggest that intention rises to the level of figurative or metaphorical reference—meaning such references are not good candidates for encoding with `@metaRef`. At the same time, Robinson's use of these figures is not by any means original, in the sense that they are not her sole inventions; in this regard, they do share an indirect connection to other texts—*A Midsummer Night's Dream* included—that also make use of the same figures. Yet, they are not the *same* characters so much as alternative versions of the same abstract figures, reworked and reimagined by each author for her own particular purposes: Robinson's Puck is not a reference to Shakespeare but rather a rhetorical gesture of participation in a shared cultural resource. Consequently, we have chosen to record information about each distinct version of such figures, independent of their other textual instantiations: Mary Robinson's Puck and Shakespeare's Puck appear as two separate people in the WWP's personography.[3]

Planning, Organization, and Execution

Because the WWP's encoding practices have always involved tagging personal names with `<persName>`, few immediate changes to our regular encoding workflow were required as we began working on the more advanced aspects of recording personographic information. Under normal circumstances, the WWP's encoders recognize references to personal names in the course of their encoding work, tag the names appropriately, and move on. When name-like phrases appear in a text that could potentially be names of other entities (geographic locations, say, or collective names that apply to a large group of people; e.g. "Israel" to mean "the people of Israel"), encoders consult standard reference works and the WWP staff to determine the appropriate encoding.

To create the linkage between an instance of a personal name in a text and the separate personographic entry for that person, however, additional encoding work is needed. At the WWP, this involves a single encoder isolating the `<persName>` elements from a text and adding `@ref` to each (the value of `@ref` is the unique value of the `@xml:id` attribute for that person as it appears in the WWP's personography, as described above). If there is no personography record for the individual in question a new record must be created and populated with information, and a new unique name key (the value we give to `@xml:id` in the personography) generated. To manage this process more efficiently and to minimize the opportunity for errors or confusion, we have developed a partial "divide and conquer" approach to the work of encoding, disambiguating, and researching information about people named in our texts. The general workflow we have developed is described below.

---

[3] As a concession to our sense that such references to the same general cultural figure constitute an important pattern of intertextual reference—the imagined set of all written and oral reference in English to a "Puck" over the course of many hundreds of years—we may also at some future date associate these references explicitly using the same `<link>` encoding we use for multiple and ambiguous references. This approach has the advantage of offering a more accurate representation of the way such references operate across individual texts to construct a shared cultural landscape.

Our primary encoder makes a first pass through each text to which we wish to apply full name encoding, adding `@ref` to each `<persName>` for which there already existed a personography record—so long as there is no confusion about the person being named. During this initial pass, she also notes on an internal wiki page used for tracking purposes any names that are potentially ambiguous or unclear as well as any names for which no personography record already exists. Working from this tracking page, a student research assistant then performs the work of locating disambiguating information: published scholarship that made claims about the identity of obscure individuals, genealogical records, historical information—essentially, whatever might help resolve ambiguities or identify more precisely identity of unclear references. Where appropriate, she creates new personographic entries for individuals not already included in the WWP's names database. As new records are created and ambiguities resolved, she updates the tracking page with the relevant name keys for the people in question, and during subsequent passes the encoder adds these new keys to the appropriate `<persName>` elements for the file.

Once the encoder has added a `@ref` to each `<persName>` in a file, she marks that file as complete on the tracking page. At this point, we apply a simple XSL transformation that generates a report containing a list of all `<persName>` elements in that file. Any names that are missing the `@ref` attribute are flagged so that the encoder may return to the file and add the correct name keys. Each `@ref` is also checked against the personography to ensure that no errors were introduced during the encoding process (for instance, entering a name key as "jchrist.hnj" rather than the correct key, "jchrist.hjn"). Any keys that do not match a known `@xml:id` in the personography are similarly flagged for follow-up by the encoder or WWP staff.

This system works reasonably well in cases where the names in question are straightforward or exhibit predictable ambiguities that can be resolved by consulting relevant scholarship and standard reference sources. More complex ambiguities, however, complicate this workflow because they remain unresolved throughout the process. For instance, every time a file containing multiple unresolved ambiguities is checked, dozens or potentially hundreds of names without `@ref` may appear in the error report. This can make it difficult for the encoder to distinguish resolvable ambiguities (those we believe can be disambiguated with a little additional research) from unresolvable ambiguities (those we do not believe we can resolve through our own research and which therefore require consultation with expert scholars, or that will require the use of `<link>` or similar mechanisms to record the ambiguity). For this reason we briefly considered—though ultimately did not implement—using a series of "dummy" name keys, one for each type of ambiguity, as placeholders during the encoding and research phases of the project. (A name that is ambiguous in the source text might be assigned a temporary `@ref` of "ambiguous," for example.) In principle, this would permit us to filter error reports more easily, and to generate lists of particular types of ambiguous or otherwise difficult names when needed.

While this workflow is generally adequate, the nature of the project and the difficulty of finding accurate information for the more obscure people mentioned in the WWP's texts present larger questions of scale, to say nothing of the relative costs and benefits of pursuing some of the more difficult cases. Early in the project, we established a clear set of priorities based on our sense of the relative importance of each type of information to scholars and readers in the WWP's community. We granted information about authors the highest priority, on the assumption that as a thematic research collection devoted to the study of early modern women's writing, WWO would most benefit its readers by providing accurate, detailed information about the women who authored our texts. Just below textbase authors we placed the publishers, printers, booksellers, and other

individuals responsible for the production and distribution of these texts; our motive here was to support ongoing scholarly interest in such areas as the history of the book, the social production of literature, print culture, publishing and trade networks, and reception history. Below printers and publishers, we placed all other names appearing in the body of the text, whether fictional, scriptural, mythological, or historical. Finally, we gave the lowest priority to the names of subscribers, many of whom are obscure and present significant problems of identification; we assumed that most of the scholars who use WWO have relatively little interest in individual subscribers, many of whom are often mentioned only once and about which little or nothing is known.[4] The effect of this prioritization on the WWP's personography has been evident: our information for authors and publishers is fairly comprehensive but information for people named in subscriber lists is, in most cases, minimal. While this outcome does not run counter to the WWP's goals and priorities, such an outcome may be less than ideal for collections that value breadth over depth.

This outcome also gives us pause when we consider the nature of "easy" versus "hard" cases in researching and encoding personal names. From one point of view, the names that the WWP devoted the least time to—subscriber names, in particular—are the easiest names to add to a TEI personography: they are presented in a highly consistent and regular way, often in lists or tables, making them ideal candidates for rapid (or possibly automated) tagging. By the same token, names embedded in highly variable prose or verse can be difficult to identify as names; they also require encoders to remain attentive to the surrounding context, where clues about their identity or roles may appear, sometimes separated by many lines or even pages of text. From the perspective of information availability, however, we found the reverse to be more generally true: names mentioned in the content of a text are in many cases (though by no means always) the ones that yield the most information to the researcher, while the names that appear only once—often in incomplete or partial forms—in a subscriber list or advertisement often yield little if any additional details despite intensive research. In terms of the relative cost of locating useful information about the average subscriber, in other words, anything more than cursory research produces little in the way of results for a collection like Women Writers Online.

These limitations suggest that projects like ours might benefit greatly in the future from advances in areas such as named entity recognition and auto-tagging. Particularly in texts containing highly regularized structures for naming people—as in subscriber lists—a more automated approach to encoding names could have significant benefits. Coupled with routines capable of automatically generating name keys and creating database records for names that are found in such contexts (and that do not already appear to exist in the personography), named entity recognition might reduce the amount of human labor involved in locating and tagging unique personal names. Even without the use of named entity recognition, however, other forms of automation could also speed up the process of encoding the most common names in a large textbase. For names whose appearance and usage is highly regular or whose occasional variation can easily be predicted—for instance, "Jesus Christ" or "Queen Elizabeth"—automated routines that encode all occurrences across the entire collection would represent an enormous gain in encoding efficiency.

In considering the outcomes of this project to date, we must credit much of its success to the exceptional skills, dedication, and resourcefulness of the encoder and research assistant who

---

[4] The exception may be cases where prominent members of society or recognized authors subscribed to a text, which may in some cases offer clues about the text's relative in elite literary circles, its cultural influence, or its relative importance in establishing or disseminating particular ideas or literary conventions.

conducted most of the work. Our primary encoder for this project, Nora Peterson, already had substantial experience working with WWP texts at the time we began adding detailed name encoding to our texts. She was already expert in TEI/XML and familiar with our specialized encoding environment, and as a result early progress on the project took place far more rapidly than if we had needed to hire and train new encoders. Equally important to the success of the project was our research assistant, Katherine Meyers, who proved extraordinarily resourceful when it came to tracking down obscure references and locating new sources of information—whether that meant searching library stacks for nineteenth-century peerages or making transatlantic telephone calls to English archivists.

Tools and Usage

As the WWP's XML personography began to hold substantial quantities of data during the final stages of this project, we began the work of developing a set of prototype tools to facilitate interaction with this increasingly rich body of contextual information. To smooth the process of exploring how such information might support various forms of interaction, we selected a set of freely-available Web-based software and tools for our prototypes. This decision allowed us to focus our efforts on thinking about the sorts of activities personographic data could make possible, rather than the complexities of writing new code.

The first of the existing tools we began using was the Exhibit framework, originally developed by the Simile Project at the Massachusetts Institute of Technology and now freely available.[5] Simile Exhibit is essentially a data mashup platform capable of displaying a single set of data in multiple formats: maps, timelines, time series plots, table and list views, etc. One powerful feature of the Exhibit framework—and the primary reason we began using it to display some of our personographic information early in the project's development—is its support for faceted browsing and searching. Individual facets can easily be created to represent information of the sort commonly found in prosopographic sources, such as gender, cultural roles, religion, marital status, and country of residence. With minimal customization, each facet in the Simile Exhibit interface can be used to filter records based on pre-selected criteria. This makes it possible for users to approach the Exhibit interface as an alternative browsing tool for finding texts that meet the specific criteria that most interest her. For instance, a user may wish to view a map showing the birthplaces of women writers born after 1750 who were Protestant, widowed, and spoke Italian (see Figure 1).

Though adequate as an early prototype, the WWP's version of the Exhibit interface for names currently suffers from several problems that limit its usability. Foremost among these is noticeable performance degradation when accessing large sets of data (more than 2500 to 3000 discrete records). Because the WWP's personography now contains information about more than 7000 people, loading the full set of records into our Exhibit framework is not practical in terms of speed and overall stability. While a user with a fast network connection, a browser with high JavaScript execution performance, and a powerful CPU can use the Exhibit interface with the WWP's full personography, most will notice significant delays and slowdowns when interacting with the full personography. For this reason, significant work in the areas of code optimization, dynamic loading, and data caching will be needed before we can use the Exhibit framework as a public interface for

[5] See http://www.simile-widgets.org/exhibit/.

our full personography.

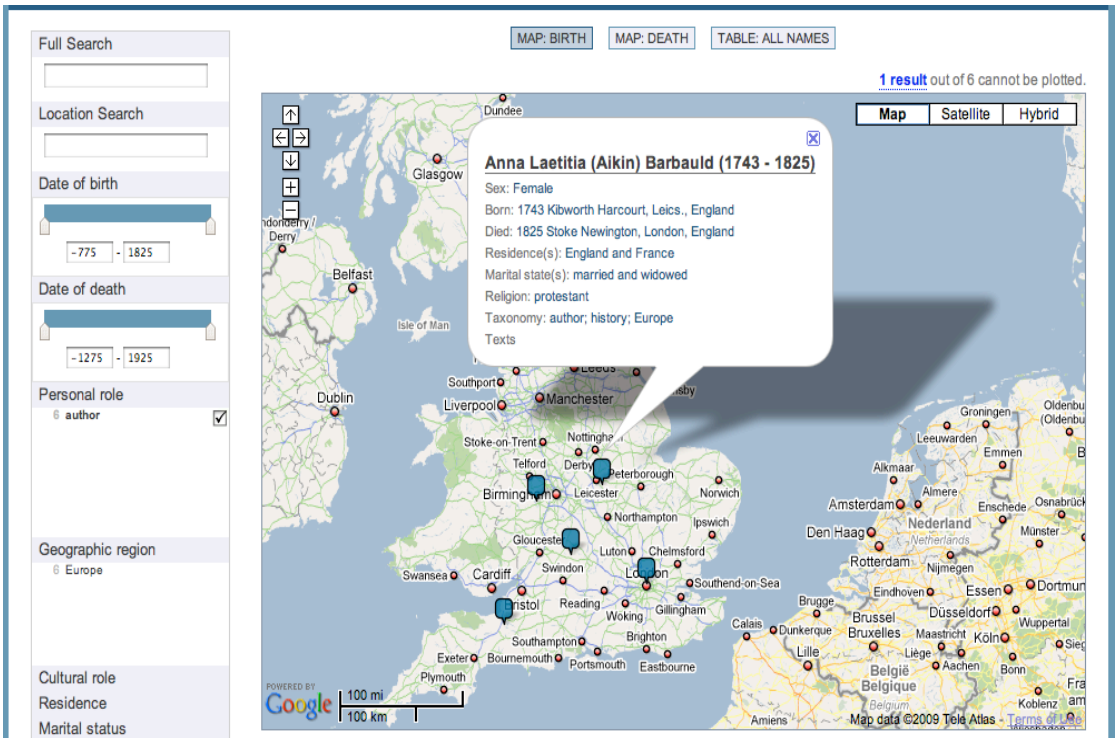

FIGURE 1. A Simile Exhibit page displaying faceted browsing capabilities

While the Exhibit framework provides something akin to a birds-eye view of the WWP's personographic landscape, we have also begun to experiment with tools that focus more specifically on visualizing relationships among people and texts. At present, we have begun the JavaScript InfoViz Toolkit (JIT), a free JavaScript visualization framework developed by Nicolas Garcia Belmonte.[6] JIT supports a core set of standard visualization types, including hypertree, rgraph, spacetree, and treemap structures; for the WWP's purposes, the rgraph and hypertree visualizations have proved especially useful for creating elegant visual representations of names and texts that can be animated to respond to user interaction. Using the JIT's hypertree visualization, for instance, we have created a simple animated graph that displays the names of people mentioned in a given text. The text is initially represented as the graph's central node, with the names of people mentioned in the text represented as outer nodes arranged in a circular pattern and connected to the text by radiating lines that resemble the spokes of a wheel (Figure 2a). When a user selects a given name, the graph morphs and redraws itself so that the selected name appears at the center and the titles of texts that mention this person appear as the outer nodes (Figure 2b). By interacting with the graph in this way, readers can develop a sense of connection between individual people and texts based on their proximity to one another in this chain of interconnection.

The JIT's rgraph visualization provides a similar approach to connecting people, though it provides a somewhat more straightforward mechanism for linking multiple nodes to one another. For this reason, we have used it to display a set of relationships between textbase authors and the

[6] Code, documentation, and demonstrations are available at http://thejit.org/.

printers/publishers who helped produce and distribute their work. The resulting view offers one way to represent graphically the publishing networks in early modern England. It shows, for instance, how certain publishers were centrally important figures in the dissemination of women's writing as well as the way in which certain women—Charlotte Smith at the end of the eighteenth century, for example—were extremely influential in the publishing world, insofar as their work was published simultaneously by many publishers working throughout Great Britain (Figure 2c).

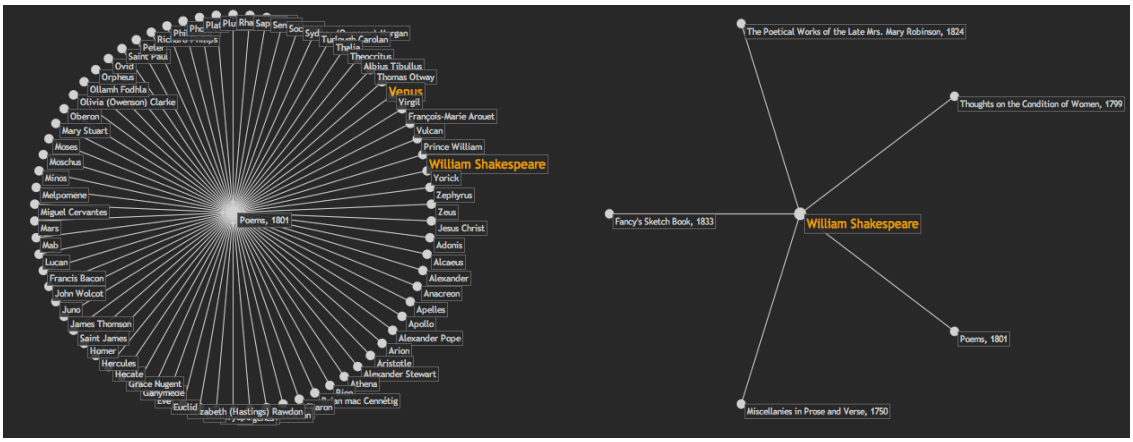

FIGURE 2A. JIT hypertree graph with a text as the center node.

FIGURE 2B. JIT hypertree graph with a person as the center node.

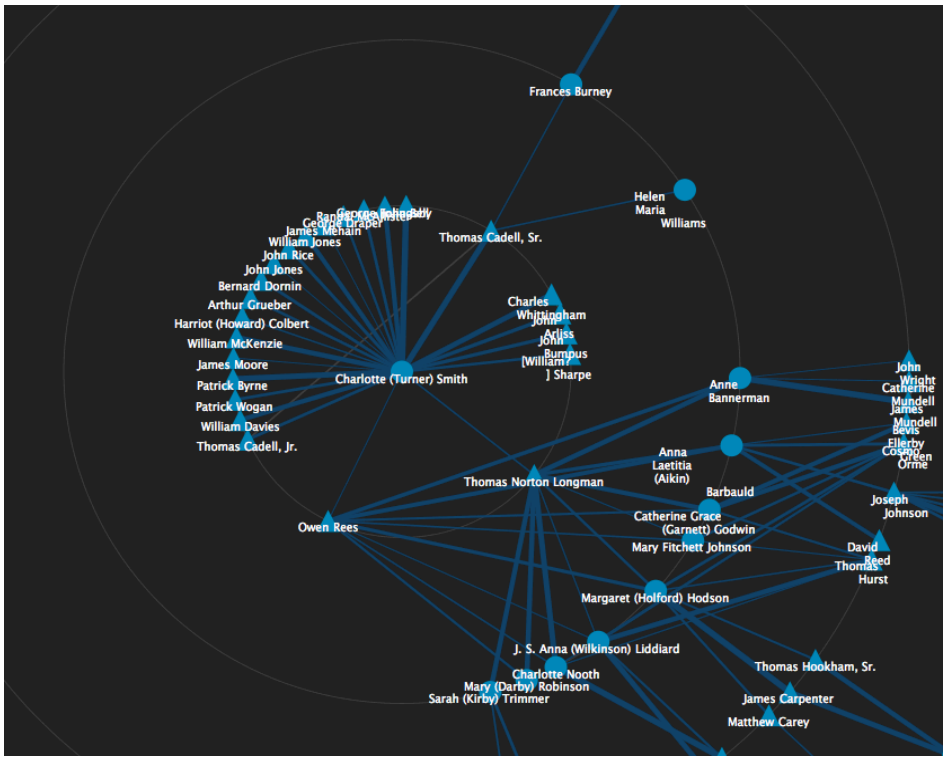

FIGURE 2C. JIT rgraph visualization showing connections between authors and publishers in a subset of Women Writers Online texts.

While both of these visualizations have been developed as stand-alone interfaces and are currently published as part of the WWO sandbox (http://golf.services.brown.edu/sandbox/: a deliberately experimental area set apart from WWO proper), we hope eventually to make them available as part of a suite of visualization tools that are integrated into the main WWO reading interface. We imagine, for instance, that one common scenario might involve a reader who encounters a name of interest in one of our texts and wishes to know more about it in relation to other texts or people; selecting that name might launch a modal window containing one or more visualizations like those described here There are, of course, numerous other possibilities along the same lines, but the goal of all of them would be to offer scholars and casual readers alike an alternative method for "reading" patterns of association and relationships among and between specific examples of early women's writing.

Along the same lines, we have also developed an initial interface for "distant reading" WWP texts in comparison to one another, according to the patterns of personal name reference they exhibit. The interface permits the user to select two texts to compare, side-by-side, in several possible views. For instance, a reader may wish to see some basic statistics about the frequency of historical versus fictional names in a particular collection of Romantic-era poems. After selecting the options that will give her this view, she might then wish to compare the results to those for a seventeenth-century collection of spiritual poems. The tool permits the user to make new selections or change viewing options on the fly, resulting in the dynamic redrawing of information for that text. As with the WWP's other prototype tools, we can imagine integrating this display into the normal reading interface, allowing the user to generate graphs for one or more texts based on a specified set of parameters (Figure 3). Incorporating such tools into more general reading aids (like glossaries or indexes of names, timelines, etc.) would provide readers with a richly varied reading interface that supports non-linear discovery and exploration in additional more linear approaches to reading.

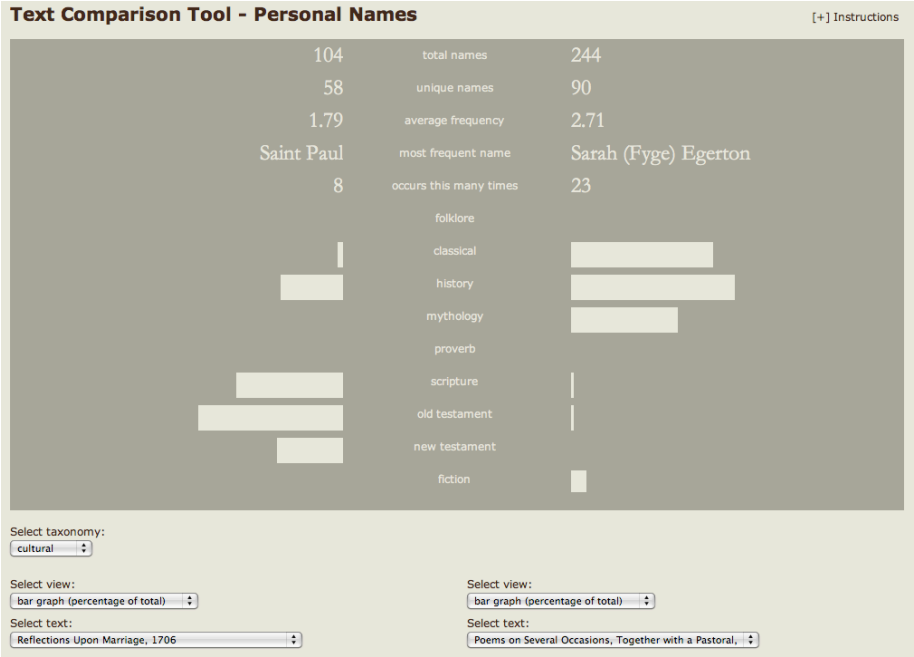

FIGURE 3. A tool for comparing two texts according to personal names classified by type.

At a general level, we imagine using some of these types of visualization tools to make broad observations about a particular text or group of texts. Even with the present prototype tools readers can get a general sense of the relative frequency of different types of names—names from classical mythology versus names from the Bible, for instance. This can lead to useful observations, particularly for students: the fact that Anne Bradstreet, who is often presented as a quintessential Puritan poet, talks about figures from classical history far more frequently than she talks about people from the Bible is strikingly evident when visualized in this way. Such seemingly simple observations can then become the starting point for a more focused investigation that might ask, for example, why a female writer in New England in 1650 would have been so interested in displaying classical—as opposed to scriptural—knowledge.

Even in the absence of specialized visualizations, however, we intend to make contextual information about people available from within our standard reading interface, and we have begun to experiment with modal windows and information popups that readers can use to find out more about the people named in a text (see Figure 4).

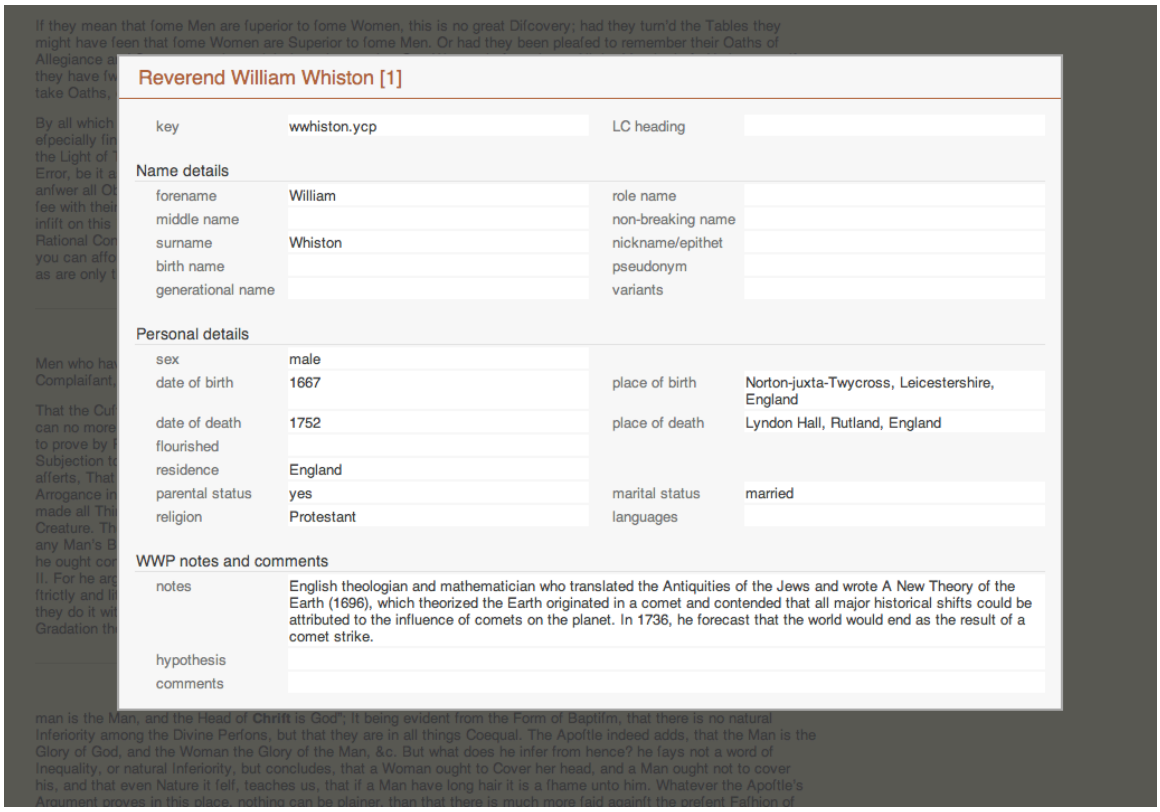

FIGURE 4. Modal view of information extracted from the WWP personography, as it could be made available within a reading interface.

Scaling Up

Because this project was focused on building a prototype, we have been addressing certain questions and design issues on a comparatively small scale—although, as indicated above, even at prototype scale the data we have developed is very substantial and tests the limits of the publicly available tools

we are using. When scaled up to include the current WWP textbase in its entirety (over 320 texts, ten times the size of our prototype subset), we can expect both the quantity and complexity of the data to scale accordingly. In addition, we intended this experiment not simply as an investigation about persons and personal names, but also by extension about all kinds of contextual information arising from named entities in our texts: places, other texts, possibly events and organizations. Hence our questions at the conclusion of this initial phase have to do with how we can extend what we've learned to the larger body of data.

Some of these questions have to do simply with interface design. The visual arrangement and default behavior of several of the tools we've experimented with are optimized for data sets containing hundreds of items (rather than dozens or thousands). For example, the map view shown in Figure 1 is already crowded in our current implementation. Even though each marker can represent an aggregation of people (using a numeric label), as the number of people represented increases the granularity of their placement will also increase: in other words, more distinct locations on the map will be populated, which in turn will mean that the markers will be closer together and (at a certain point) will no longer present an intelligible visual field to the reader. Even now, when viewing the map of Britain as a whole, there are cases where the number on the marker (showing that it represents more than one person) is obscured by another marker, so that from this distance it is difficult to get a sense of the relative density of people in different locations. Some method of showing the scale of the aggregations is needed so that the distant view remains informative. These are challenges arising not from the design of the tools but rather from the basic ratios of data to visual space.

Contextual Data and Scholarly Reading Practices

The availability of data and tools like these raises broader questions about the kinds of reading activity and intellectual engagement they support, and, considered more broadly still, how visualization and digital reading interfaces promote or inhibit particular avenues of scholarly investigation and critical reading. These questions inform the next step of our work in this area, namely the more direct incorporation of reading tools like these into WWO. At present we envision them as a suite of experiments which are deliberately kept separate from the main WWO interface—both because as tools they are still not ready for general use, and because the markup of the underlying data would need to be extended to the full textbase. However, when those obstacles are removed, the question remains of how these tools should be associated with the reader's experience of the texts. At what point(s) in the reading or research process does one want or need access to this information? And should we be conceptualizing the reader's encounter with these tools as part of a planned research exercise on their part (in which they seek out a specific tool to help address a specific need) or as a serendipitous exploratory moment, pursued on a whim, or as a constant adjunct to the normal processes of reading?

Our hypothetical answer at the moment—to be refined as we proceed—is that access to this kind of contextual information needs to be framed as a combination of these possibilities. Traditional digital publications locate access to informational facets like these as part of a distinct "search interface" which consolidates the reader's mental questions about the text into a single coordinated act of searching. A wide range of potential questions about the text (or about the collection as a whole) are thus framed as "where is phenomenon X?" or "which texts contain feature Y?" The presentation of the results as a list of hits also frames the range of sequels to the question quite tightly: the next step for the reader is assumed to be the selection of a particular hit or hits as targets to be read in detail.

The result set, in other words, is construed as a set of things that either are or are not what the reader wanted upon initiating the search; the reader's interest in them is generally assumed to be focused on their separate identity rather than on their coherence as a group.

However, this model of inquiry as "searching" has limited application, and its prevalence as an interface device arises more from the constraints of publication tools and interface design than from the ways readers actually interact with texts. When working with an unfamiliar text, conventional search paradigms force the reader to frame what they want to know in terms that they may not yet have in hand, and channel their interaction with the text through the vector of their current limited knowledge. For these reasons, it may be helpful to design the interaction so that what the text *knows about itself* (in this case, information about persons) can be made visible to the reader *as part of the reading process* rather than as a separate, digressive question that takes the reader away from the text. For example, one model we have discussed for the next iteration of the WWO interface would include a set of inspectors in the margin of the reading window, whose function is to show basic information both about the interior of the text and about its relation to the collection as a whole. Examples of the former include:

- Word frequency within the text for a selected word
- Personal identity and biographical information for a selected name
- A map showing the location of a selected place
- A graphical display (using a word cloud or similar tool) showing relative word frequencies for all words in the text
- A highly compressed graphical view of the entire text, showing the current reading location, the basic textual divisions (e.g. chapters, cantos, etc.), perhaps the distribution of the selected feature (name, word, place, etc.) across the text; this view could also be used for quick navigation by permitting the reader to jump to a selected section by clicking on it
- A tree display showing the location of the current passage within the overall XML structure of the text; this could also be used for navigation, or to see in what other structural contexts other instances of a name or word appear within the text
- A definition of a selected word (drawn from a linked dictionary)
- A keyword-in-context view of all instances of a selected word elsewhere in the text

Examples of the relation between the text and the collection include:

- Word frequency within the collection for a selected word
- A timeline showing the text's publication date in relation to the publication dates for the collection as a whole
- A map showing the text's place of publication; the same map could show in a different color the publication locations of the other texts in the collection
- A search interface that displays the results of the search as a text map (i.e. a conceptual map showing the collection as a "space" organized by time, author, genre, etc.), so that any patterns in the distribution of search results can be seen immediately, but without leaving the text one is currently reading

- A "find similar" feature that lets the reader choose facets of the text (author, date, the presence or absence of a textual structure such as lists or poetry, the presence or absence of a certain word, genre) as a basis for discovering connections between texts in the collection.

This is an extremely ambitious vision and most of these features fall outside the scope of our present skills and resources. However, these ideas demonstrate the kind of reading model we envision, one in which the interface can respond to the kinds of curiosities and casual questions that arise directly from reading an unfamiliar text: What does this word mean? Who else uses it? Is it a regional usage? Where am I in this text? How many more poems are there in this chapter? Where is Bohemia? Is this Pitt the Elder? What was his wife's name?

Questions like these may simply be like scratching a mild itch, or they may be the starting point for a more focused inquiry, and one further challenge is to gracefully accommodate the moment where that inquiry ceases to be connected with a specific text, and needs to shift (both mentally and as a matter of interface spaces) into an independent space. Quite apart from the issue of screen real estate (fitting these inspectors into a narrow margin will require ingenuity but will also necessitate some compromises in what the tools can display), there is the issue of intellectual work flow and of how these various views of the text might interact most effectively in support of the reader's research.

We also want to ask, as a prompt for further investigation in these areas, about the impact of reading interfaces on reading practices. Expert readers—in this case, for instance, scholars of the early modern period—come to even an unfamiliar text with a set of knowledge that decisively inflects what they expect to find, and that may even affect what they regard as significant or even what they are able to notice. Scholars' adeptness at reading texts through a particular interpretive frame (for instance, male homosociality, Foucaultian power structures, domesticity, etc.) is a form of intellectual efficiency: a way of not inventing a fresh set of critical terms for each and every text. The question is whether there are modes of reading available to scholars that can promote a more excavatory and less tendentious engagement with the text, by defamiliarizing the text or causing the terms of its familiarity to recede, while bringing unexpected patterns into clearer view.

This defamiliarization has already been discussed in the context of data mining approaches to text analysis, and an eloquent case is made in Stephen Ramsay's work on what he terms "algorithmic criticism": critical approaches that operate through the discovery of patterns arising from detailed computer-assisted text analysis. The reading interfaces we are exploring draw on this approach, but situate textual patterns within the larger context of cultural patterns, and broaden out the analysis of text to include textual structure and genre as well as linguistic information.

With the completion of this initial phase of our personographic encoding, the next steps are: first, to extend our personographic encoding to the entire WWP textbase (already under way); second, to scale up the tool prototypes and integrate them into the main WWO interface (now in planning); and third, to explore ways of integrating our personographic data with that of other projects, especially those (like Project Orlando) specializing in women's writing. Details on further progress will be available at the WWP site, http://www.wwp.brown.edu.