

Learning from the Past: The Women Writers Project and Thirty Years of Humanities Text Encoding

Connell, Sarah; Flanders, Julia; Keller, Nicole Infanta; Polcha, Elizabeth;
Quinn, William Reed

Northeastern University

sa.connell@northeastern.edu
<https://orcid.org/0000-0001-9202-5798>
j.flanders@northeastern.edu
<https://orcid.org/0000-0002-4199-0130>
day.n@husky.neu.edu
<https://orcid.org/0000-0003-2317-732X>
polcha.e@husky.neu.edu
<https://orcid.org/0000-0003-2395-6171>
quinn.wi@husky.neu.edu
<https://orcid.org/0000-0001-5702-9857>

Received 10/04/2017; accepted 17/06/2017
DOI: <https://doi.org/10.7203/MCLM.4.10074>

ABSTRACT

In recent years, intensified attention in the humanities has been paid to data: to data modeling, data visualization, “big data”. The Women Writers Project has dedicated significant effort over the past thirty years to creating what Christoph Schöch calls “smart clean data”: a moderate-sized collection of early modern women’s writing, carefully transcribed and corrected, with detailed digital text encoding that has evolved in response to research and changing standards for text representation. But that data –whether considered as a publication through Women Writers Online, or as a proof of the viability of text encoding approaches like those expressed in the *Text Encoding Initiative (TEI) Guidelines*– is only the most visible part of a much larger ecology. That ecology includes complex human systems, evolving sets of tools, and a massive apparatus of documentation and organizational memory that have made it possible for the project to work coherently over such a long period of time. In this article we examine the WWP’s information systems in relation to the project’s larger scholarly goals, with the aim of showing where they may generalize to the needs of other projects.

KEYWORDS

Digital humanities; XML-TEI; Women Writers Project; women’s writing; documentation; English literature; early modern texts; eighteenth century



Magnificat Cultura i Literatura Medievals 4, 2017, 1-19.

<http://ojs.uv.es/index.php/MCLM>

ISSN 2386-8295

Aprenent del passat: el Women Writers Project i trenta anys de codificació de textos en humanitats

RESUM

En els últims anys, dins les humanitats s'ha prestat gran atenció a les dades: a la modelització de dades, a la visualització de dades, a les "dades massives". El projecte Women Writers ha dedicat un esforç significatiu durant els darrers trenta anys a crear el que Christoph Schöch denomina "dades intel·ligents i netes": una col·lecció de tamany mitjà dels escrits de dones de l'edat moderna, transcrita i corregida amb cura, amb una codificació de text digital detallada que ha evolucionat d'acord amb la recerca i als canviants estàndards de representació de textos. Però aquestes dades, ja siga considerades com a publicació a través de Women Writers Online, o com a prova de la viabilitat d'enfocaments de codificació de text com els expressats a les *Text Encoding Initiative (TEI) Guidelines*, només són la part més visible d'una ecologia molt més gran. Aquesta ecologia inclou sistemes humans complexos, conjunts d'eines en evolució, i un aparat massiu de documentació i memòria organitzativa que ha permès que el projecte treballi de forma coherent durant un període tan llarg de temps. En aquest article examinem els sistemes d'informació del WWP en relació amb els objectius acadèmics a llarg termini del projecte, amb l'objectiu de mostrar on poden estendre's per cobrir les necessitats d'altres projectes.

PARAULES CLAU

Humanitats digitals; XML-TEI; Women Writers Project; escriptura de dones; documentació; literatura anglesa; textos de l'edat moderna.


Connell, Sarah; Flanders, Julia; Keller, Nicole Infanta; Polcha, Elizabeth; Quinn, William Reed.
2017. 'Learning from the Past: The Women Writers Project and Thirty Years of Humanities Text
Encoding', *Magnificat Cultura i Literatura Medievals*, 4: 1-19 

TABLE OF CONTENTS

- 1 Introduction – 3
- 2 The Information Apparatus – 6
- 3 The Tool Apparatus – 11
- 4 The Human Apparatus – 14
- 5 Next Steps – 16
- 6 Works Cited – 19



1 Introduction

In recent years, intensified attention in the humanities has been paid to data: to data modeling, data visualization, “big data”. The Women Writers Project has dedicated significant effort over the past thirty years to creating what Christoph Schöch calls “smart clean data” (Schöch 2013): a moderate-sized collection of early modern women’s writing, carefully transcribed and corrected, with detailed digital text encoding that has evolved in response to research and changing standards for text representation. But that data—whether considered as a publication through Women Writers Online, or as a proof of the viability of text encoding approaches like those expressed in the *Text Encoding Initiative (TEI) Guidelines*—is only the most visible part of a much larger ecology. That ecology includes complex human systems, evolving sets of tools, and a massive apparatus of documentation and organizational memory that have made it possible for the project to work coherently over such a long period of time. If the challenge facing digital projects is precisely how to create “smart clean data” at a useful scholarly scale and over time, studying ecologies like this one can help us understand the specifics of that challenge. What does it mean to build scholarly intelligence into such data, and what is involved (in training, communication, technical systems, documentation, and organizational memory) in doing so? In this article, we examine the WWP’s information systems in relation to the project’s larger scholarly goals, with the aim of showing where they may generalize to the needs of other projects.

The WWP was founded at Brown University in the late 1980s, and was first funded by the National Endowment for the Humanities in 1988. After remaining at Brown for twenty-five years, the project moved to Northeastern University in 2013. At its inception, the project’s urgent motivation was the rediscovery and republication of pre-Victorian women’s writing in English, at a time when most pre-1850 materials were either entirely unknown or inaccessible in rare book libraries, available only through microfilm. But early discussions of how nascent digital technologies might support the project yielded an important additional strand of inquiry: how might we understand the digital representation of such texts in terms of editorial theory, and how would these representational choices affect the meaning, research value, and informational status of the texts? The research mission of the project was thus from the outset framed around these conjoined questions of gender, textuality, and digital representation.

These questions were timely, because in 1987 the Text Encoding Initiative (TEI) began developing a set of guidelines for representing digital texts in an open, standards-based manner, which were first published in 1993 as the *TEI Guidelines for Electronic Text Encoding and Interchange* (TEI 1993).¹ The existence of such guidelines not only made it possible for projects like the WWP to proceed on a firm technical basis, but also provided a research community in which the WWP’s questions about the intersection of gender and editorial politics with digital technologies made sense. The emergence of “humanities computing” and then “digital humanities” as coherent fields of study and praxis in the 1990s and 2000s also meant that funding from public and private agencies like the NEH and the Andrew W. Mellon Foundation was available to pursue this research, and the WWP was fortunate to receive a series of major grants that supported the project’s early development: first to build the collection itself, and then to develop and publish documentation,

1. The current *TEI Guidelines* are available online at <http://www.tei-c.org/Guidelines/>.

training materials, and a workshop curriculum in text encoding. Some of this work has produced important related data sets. *Women Writers in Review*, a collection of periodical reviews of WWO texts, was produced through the NEH-funded “Cultures of Reception” initiative.² Most recently the project has begun developing a comprehensive bibliography of works quoted and cited in WWO texts, under the NEH-funded “Intertextual Networks” grant.³ Collaborative research from both of these projects is published in *Women Writers in Context*, a collection of exhibits and contextual essays building on our original Mellon-funded “Renaissance Women Online” project.⁴ All of these related data sets are open-access, supported by license revenue from *Women Writers Online*, which enables the project to continue to develop these resources after their startup phase and also to continue its education and outreach programs and its digital humanities research.⁵

One important underlying question for the project has been what we mean by “access”, “publishing”, and “dissemination”. In 1988 there were no clear avenues for publishing or analyzing the WWP’s growing collection of digital texts, which were first circulated to readers through paper printouts and included in thousands of course packets. A small number of texts were published in printed editions in an experimental series with Oxford University Press. But following the advent of the World Wide Web in 1993 it quickly became clear that digital resources would become an essential tool for humanities scholarship, and *Women Writers Online* was first published in 1999. In its early versions, the WWO interface for reading was essentially a translation of the traditional reading experience into digital form: presenting a single text at a time, albeit with navigation and searching. It expanded but did not transform what we imagine by “reading” or “access”. The early WWO search interface represented much more of a transformation, because it gave readers an opportunity to use the structural markup within the texts to create more intelligently focused searches, and it presented the search results in a way that could be used to read patterns across the entire collection (for instance through a keyword-in-context display). A somewhat later version of the interface offered complex text analysis features such as collocation and fuzzy matching through which readers could trace patterns of words throughout the collection. “Access” and “reading” in these interfaces thus represented an interplay between an individual text (construed as an object to be discovered and read) and the collection as a whole (offered for cross-cutting analysis and pattern discovery), moving in the direction of what Mitchell Whitelaw has called a “generous interface” (Whitelaw 2015) in which the underlying logic of the collection is made visible to the reader. The most recent work on the collection adds a networked dimension, enabling readers to follow connections between texts, periodical reviews, cited works, biographical information, and other forms of context. These changes have also been motivated by changes in readership: from an early cautious attachment to visual fidelity to the source material, readers have become steadily more interested in the ways digital texts can support analysis. As the word “data” becomes more familiar to humanities scholars, the idea that these texts are data becomes less alienating and more empowering. One of the WWP’s working groups is currently exploring methods of direct analysis of the underlying XML data, without a mediating interface, and we anticipate soon creating a public portal through which scholars can get direct access to WWP data.

A final important set of research questions have to do with how we understand gender, and where theorizations of gender inhabit our work. The most basic of these questions concerns what we mean

2. See <http://wwp.northeastern.edu/review>.

3. See <http://wwp.northeastern.edu/research/projects/intertextuality/index.html>.

4. See <http://wwp.northeastern.edu/context/>.

5. The source data for *Women Writers Online* is also freely available to researchers upon request; the project is currently planning an API through which the WWP’s data can be publicly exposed for reuse.

by “women’s writing”. We treat this rubric not as a literal descriptor of physical biology –which would be unverifiable, overly simplistic, and irrelevant to the goals of the collection– but rather as a term that covers a variety of different modes of authorship by people who positioned themselves as female. This includes translations of works by women, translations by women of male-authored works, co-authored works involving female authors, and cases of disputed or unknown authorship in which the work circulated at the time under an assumption of female authorship. One of our most complex cases is *The Ladies’ Diary, or, Woman’s Almanack* (1704–1801), a periodical edited by John Tipper and aimed at a female readership, containing materials contributed by readers who identified themselves as female.

The gender politics of gathering and circulating women’s writing (however defined) are fairly clear. At a deeper level, however, we need to be aware of the gender politics of the editorial enterprise itself. As scholars like Stephanie Jed, Katie King, Martha Nell Smith, Donald Reiman, and many others have shown, editorial methods have a deep-seated gender politics (Jed 1989; King 1991; Smith 2004; Reiman 1988; Sutherland and Pierazzo 2012) that informs how we think about the respective authority of different kinds of source materials, different kinds of cultural producers (authors, scribes, translators, printers, publishers, etc.), and the role of the editor him/herself. Even our understanding of where textual authority comes from –physical evidence in specific witnesses, the intentions of authors and other producers, editorial judgment– is inflected through theorizations of documentary materiality that are historically deeply gendered.

And finally, we must consider the formalization and analysis of gender as an information category within the data being produced. Formalization of categories of identity, such as gender and race, is an area that has received increasing scrutiny both in the information science community (e.g., Billey *et al.* 2014) and in the context of specific standards such as the TEI (Terras 2013). There is now a much wider recognition of the influence such categorizations exercise, not only over practical outcomes such as user interface behaviors, but also more subtly and pervasively in reinforcing cultural norms. Gender is represented in a number of different places in the WWP’s encoding, where it functions as an informational hook that supports some specific analysis. At the most basic level, the metadata for each text (which functions like a catalogue record in a library collection) contains information about the gender of each entity responsible for the work: author, translator, publisher, printer, and so forth. Within the markup of the text itself, we are also now experimenting with identifying the gender of characters in dramatic texts (with the goal of extending this work to include other genres), to support research on the gender dynamics of character interactions. To reflect the complexity of our broader understanding of gender, the descriptors used in these cases are not limited to “female”, “male”, “mixed”, and “unknown” (although at present those are the only values we have needed).

It may be helpful to conclude this introduction with a very brief orientation in the data itself, for readers who are unfamiliar with XML and the TEI. XML is a data representation system in which information structures are represented through digital codes –tags like <head> or <quote>– which demarcate, organize, and name those structures. Through XML, it is possible to define specific markup languages that describe different kinds of data, such as historical documents, or chemical formulae, or web pages, or financial transactions. The TEI is one such language, designed to describe and encode humanities research materials such as primary source documents, oral histories, linguistic data, scholarly editions, manuscripts, and many others. Although the TEI can be used to represent very simple data, it excels at providing a detailed account of editorial, interpretive, semantic, literary, and historical features of texts, which can be used to support nuanced scholarly analysis. We show more complex examples further on, but a brief sample will illustrate the basic principles of this type of information:

```

<epigraph>
  <quote source="#cavendish_olio">
    <lg type="couplet">
      <l>But if not favour'd, then my Book muft dye,</l>
      <l>And in the Grave of Dark Oblivion lye.</l>
    </lg>
  </quote>
  <bibl><author>The Duchesse of Newcastle</author></bibl>
</epigraph>

```

Fig. 1: Sample encoding

One distinctive feature of this approach (not illustrated in the above example) is that it allows for complex layering of information, in which multiple textual possibilities can be represented at once: editorial speculations, transcriber uncertainties, authorial revisions, spelling modernization. From this abundance of potential information, specific strands can then be displayed or analyzed to different ends, for instance to create a reading text for the general public, or an original-spelling edition, or a visualization showing historical changes in the language of poetry.

2 The Information Apparatus

In the digital humanities, the close interconnections between information organization, human work practices, and intellectual outcomes have been a steady preoccupation; early exemplars such as the NINCH *Guide to Good Practice* (NINCH 2002) or the case studies in the TEI's early volume on electronic textual editing (Burnard *et al.* 2006) explore these connections in detail. The WWP's ecosystem exemplifies this symbiotic relationship between information organization and human labor. Our system for managing information largely depends on our division of labor and project managerial support, and concurrently, on documentation that enables sustainable encoding and publication practices across time and across shifting WWP team members. In order to parse the various ways information is organized, documented, and managed within this symbiotic relationship, it may be helpful first to provide an overview of the project's information management systems.

The cornerstone of these systems is a database documenting the project's encoding practices, which is continuously updated based on the outcomes of encoding meeting discussions and decisions, and serves as an essential reference for encoders. Several other systems also help track different work processes and information. The WWP maintains an email discussion list which serves as a cumulative record of past discussions, where WWP team members can sort through encoding and publication conversations dating back to June of 1994. Encoding history is also captured at the micro level; for example, within each TEI file we maintain a change log of major milestones in the encoding and proofing process.⁶ Changes to specific files are also reflected in our version control system. These logs provide a record of the work completed on each file. Similarly, once a TEI file

6. Change logs are recorded in the metadata for each file using a <revisionDesc> (revision description) section. For more on the TEI's provisions for recording revisions, see: <http://www.tei-c.org/release/doc/tei-p5-doc/en/html/HD.html#HD6>

has been encoded and printed for the proofing process, we write notes and questions by hand on the proofing sheets attached to the printed copy in order to keep track of our post-encoding editorial decisions. Finally, we track the overall progress of texts through the encoding, proofing, and publication processes using simple project management tools (currently, a tool called Trello⁷). All of these resources for information management enable WWP team members to refine both the textbase and our encoding practices in response to new texts and changing tools. We are constantly testing our encoding practices and documentation of those practices against the texts in our collections and adjusting our encoding, or refining our documentation, as needed. The endurance of the project thus depends on a continued attention to organized information and documentation.

To show how these informational systems work in practice, it is useful to sketch out how an encoding question might be engaged across these various resources. For example, the encoding process might reveal an unusual feature within the text, something that doesn't fit with current practices. This was the case recently when an encoder encountered a poem in Charlotte Smith's *Elegiac Sonnets* (1797),⁸ in which a number of poetic lines were omitted and replaced with asterisks. When encountering an encoding question like this, the first step would be to check the WWP's internal documentation. For this specific problem, the encoder would likely review the narrative section of our documentation on "Typography and Special Characters" to see if this phenomenon had already been addressed.⁹ A next step could be to explore the encoding discussion list, searching for terms like "omission", "redaction", and "asterisks". The encoder could also use XPath, a query language for searching within XML documents, to look through the WWP's textbase of TEI files for similar examples of typographic redaction, such as other instances of multiple asterisks and dashes within features like poetic lines. As an encoder moves through these different resources, she can collect examples, noting if there are any potential inconsistencies between the documentation and the encoded files, or between different approaches to encoding the same phenomenon.

After conducting this research in the WWP's information systems, the encoder would then take her question and her findings to one of the WWP's weekly encoding meetings, or to our project manager's office hours. The question is then discussed by multiple WWP team members, and resolved collaboratively. The results of this discussion could take multiple forms, such as adding a new example in the internal documentation to help address similar questions in the future, making changes to the WWO schema, or deciding whether we can use our current encoding practices to address the question. If it's not possible to represent a phenomenon in current encoding practices and we are not sure that phenomenon is widespread enough to merit changes to the schema, we may instead add a note to the text's interface display. For instance, Lady Eleanor Davies' *The Benediction* (1651) includes two typographic characters representing the eye and the horn of the lamb, the sun and the moon, and the letters "O" and "C" for "Oliver Cromwell" —we chose to include a note explaining these multiple layers of meaning rather than relying on the encoding alone.¹⁰

In the case of the redacted poetry lines in Smith's *Sonnets*, we decided to change our schema to add an <elision> element for typographic representations of redacted text. We came to this decision based on encoder research, after realizing that similar typographic redactions appear across our textbase in both poetry and prose. In the excerpted mark-up below (Figure 2), we have used the

7. <http://trello.com>.

8. This text has been published in WWO at: <http://www.wwp.northeastern.edu/texts/smith.sonnets.html>. Access to Women Writers Online requires a subscription; for more information on licensing and setting up free trial access, please see <http://www.wwp.northeastern.edu/wwo/license/> or email wwp@northeastern.edu.

9. See http://www.wwp.northeastern.edu/research/publications/documentation/internal/#!/entry/typography_narrative

10. This text has been published in WWO at: <http://www.wwp.northeastern.edu/texts/davies.benediction.html>.

<elision> element to represent a gap in the poem where three lines of asterisks indicate that the remainder of the line group has been redacted.

```
<l>Beneath accumulated horror, finks</l>
<l>The defoliate mourner!</l>
<elision>
  <lb/>* * * * *
  <lb/>* * * * *
  <lb/>* * * * *
</elision>
```

Fig. 2: Encoding redacted text using the <elision> element in Charlotte Smith’s *Elegiac Sonnets* (1797)

The development of this specific element required several schema changes and edits to our internal documentation. Our current documentation entry for <elision> explains that “the <elision> element is used to mark instances where significant portions of text have been omitted or redacted”, clarifying that we do not use this element for redactions of names, e.g. “Mr.—”.¹¹ This definition may be altered or expanded as we encounter more examples of textual redaction, and we are still working to identify all of the appropriate uses of <elision> in previously encoded files in our textbase. Recently, an encoder came across an example of redacted text in Elizabeth Craven’s *A Journey through the Crimea to Constantinople* (1789) that is more self-censorship than editorial excision. Unlike Smith’s use of redaction, which appears once in the text to indicate an abridgement, Craven uses the em-dash frequently across her epistolary travel narrative to represent censored fragments of text within the collected letters. The outcome of discovering these distinct examples, and developing encoding practices to best represent them, is that both our markup and our documentation are refined over time.

As suggested above, an important result of these practices is that the document analysis process does not end when a text has been encoded—or even when that text is proofed and published. Rather, document analysis continues over generations of encoding refinements and technological developments. Each text in Women Writers Online thus represents our best current understanding of how we can model that text through our markup. As we encode additional texts, that understanding continues to change, leading to iterative refinements of each text in the collection.

Sometimes, as in the case of <elision>, we make immediate changes to our encoding practices when we encounter textual features that we are unable to represent with the existing schema, particularly when those features seem to be semantically significant or otherwise in line with our editorial priorities.¹² In other cases, we might attempt to represent textual features within the bounds of the current schema until we can determine whether adjustments are necessary, as for instance in the WWP’s approach to acrostics. The issue of acrostics was raised during the project’s

11. See http://www.wwp.northeastern.edu/research/publications/documentation/internal/#!/entry/elision_element.

12. A contrastive example may be useful in clarifying the matter of significance: the WWP recently encoded a text that displayed several instances of shifted type, which occurs when the pressure of the press causes type to move around between impressions. We determined that—like uneven baselines, type size, and the weight and length of ruled lines—shifted type fell within the category of textual features that we silently regularize. Accordingly, we adjusted the project’s documentation to record our handling of shifted type, rather than changing our encoding practices to represent this phenomenon. For more on silent regularization, see the project’s editorial declaration at: http://www.wwp.northeastern.edu/about/methods/editorial_principles.html.

work on Frances O'Neill's *Poetical Essays* (1802),¹³ which contains, among several other acrostic poems, "An Acrostic for a Great Lady" addressed to Lady Anne Barnard. The first stanza of this poem reads:

Lift up, my soul, O Muse, inspire my lays,
And join to sing your favorite lady's praise,
Descending here with all your charms of sense,
Your brightest beams and noblest influence.

The other stanzas of the poem spell out the remainder of Lady Anne's name through the initial letters of its rhymed couplets.

In terms of its encoding, this text is one of the earliest in WWO; a change log entry indicates that transcription work began in May of 1989, a full decade before the publication of Women Writers Online in 1999. The change log records several major updates to the text, as the WWP's encoding practices (and, in fact, the TEI) developed; during one of these revisions, the question of how to encode the names of persons expressed through acrostics was raised. A message posted to the discussion list in November of 1998 records that several options were discussed; the initial decision was to encode each separate letter with a <persName> element and use the @next and @prev attributes to link these, as the most "direct and accurate" approach. The implementation of this decision several days later is documented in the file's change log, which reads: "Tagged persnames appearing vertically in acrostics". That is, the encoder tagged each letter of Lady Anne's name with a <persName> element and then linked these with attributes that represent the sequence of letters as a complete word. Essentially, each letter points to those that come before or after it, indicating that what might appear to be fifteen individual <persName>s is instead the fifteen letters of a single person's name (see Figure 3).

```
<lg type="quatrain">
  <l><persName xml:id="pn1" next="#pn2">L</persName>ift up, my foul, O Mufe, infpire my lays,</l>
  <l><persName xml:id="pn2" next="#pn3" prev="#pn1">A</persName>nd join to fmg your favourite lady's praife,</l>
  <l><persName xml:id="pn3" next="#pn4" prev="#pn2">D</persName>efcending here with all your charms of fenfe,</l>
  <l><persName xml:id="pn4" next="#pn5" prev="#pn3">Y</persName>our brighteft beams, and nobleft influence.</l>
</lg>
```

Fig. 3: Initial approach to encoding an acrostic name

This solution may be direct, but it is far from efficient and, as the initial discussion list post acknowledges, the approach is also "messy" in that it requires reassembly of the name from multiple elements. In July of 1999, the WWP revisited this issue and reached a different solution, creating a new element: <acrostic>. This new approach made it possible to treat the words expressed in acrostics as part of the hypertextual information contained within a text,¹⁴ using a @target attribute to point from the generated word to its instantiation in the text itself (in this case, the acrostic poem whose first lines form Lady Anne's name; see Figure 4).

13. This text has been published in WWO at: <http://www.wwp.northeastern.edu/texts/oneill.poetical.html>.

14. Rather than being marked on each separate letter of a name, <acrostic> elements are contained by the <hyperDiv>, a special division for the hypertextual features of texts, such as notes.

```

<hyperDiv>
  <acrostics>
    <acrostic target="#poem01">
      <persName>Lady Anne Barnard</persName>
    </acrostic>
    <acrostic target="#poem02">
      <persName>Sir Francis Burdett</persName>
    </acrostic>
  </acrostics>
</hyperDiv>

```

Figure 4: Updated encoding for acrostic names (the encoding in this case has been edited to include only two example <acrostic> elements from this text)

The WWP's internal documentation provides key details on the purpose and contents of <acrostic>:

The <acrostic> element is used to encode the word or phrase spelled out by an acrostic poem, permitting acrostics to be searched, counted, and otherwise processed explicitly...The complete word formed by the acrostic is encoded as the content of <acrostic>, with phrase-level tagging applied as appropriate when the word is a personal name, place name, foreign-language word, etc. Each <acrostic> requires a @target that points to the @xml:id of the smallest element in the main text containing the entire acrostic. Typically this will be an <lg>, but on rare occasions it may be some other element.¹⁵

The documentation goes on to cover additional specifications for working with acrostics, such as how to handle the formatting of the original text when transcribing the word formed by the acrostic poem.

In the case of acrostics, the project changed the WWO schema to add two new elements (<acrostic> and <acrostics>) after first testing an approach that drew on established markup (linking individual <persName>s). As this example shows, the WWP's information structures make it possible to see not just how the project encodes various textual features (as recorded in the internal documentation and the schema¹⁶) but also how the project has encoded such features in the past, along with the rationales behind different approaches (as recorded in the listserv) and even how those decisions have manifested in individual documents (as recorded in the change logs). The addition of new elements such as <acrostic> and <elision> is also an example of TEI customization, through which projects like the WWP can delineate their encoding practices, as defined in and constrained by their schemas, in order to enforce consistency and make it possible to encode textual features, such as acrostics, with elements not included in the base TEI guidelines.¹⁷

15. See http://wpp.northeastern.edu/research/publications/documentation/internal/#!/entry/acrostic_element.

16. The mechanisms that TEI projects like the WWP use to define our encoding practices also allow for documentation of those practices. In essence, "the TEI guidelines describe an XML language in which the schema is first modeled using a system called ODD (One Document Does it all). The ODD format is a document that contains XML schema fragments and their documentation; it also contains mechanisms for expressing specific choices and constraints, such as the application of local controlled vocabularies or the omission of specific elements" (Flanders and Jannidis 2014: 231).

17. TEI customization can take many forms; many of these are restrictions, such as creating controlled lists of values for attributes to ensure consistency in encoding. The TEI provides extensive mechanisms for customization –for more on these, see <http://www.tei-c.org/Guidelines/Customization/index.xml>.

The WWP's models for representing texts are tested, refined, and applied to new textual phenomena over time, a long-term approach to encoding made possible by detailed and multivalent information structures. We work with the expectation that both our encoding practices and our publication capabilities will continue to develop, and so our current work is deeply diachronic, drawing on the information about modeling texts we have gathered over the past three decades while working in anticipation of capabilities for publishing those texts that we have not yet developed.

3 The Tool Apparatus

The WWP has always encoded with the understanding that the emergence of new tools will influence our practices over time. The tool apparatus at the WWP contributes to our long-term and ongoing approach to document analysis and markup. We use various tools not only to make document-level decisions about encoding, but also to consider corpus-level changes. These tools allow us to shuttle between close readings of individual documents and a large-scale, XML-based analysis of the entire corpus. We have begun referring to this latter process as “distant encoding”.

Distant encoding refers to the two lenses we now use while encoding. With one view, we can scan the entire corpus using various XML-based languages. This reveals the encoding patterns of some textual phenomenon. For instance, advertisements in WWO's collection typically are for books and usually contain <bibl> elements (for bibliographic references), with nested <title> and <author> elements; using XPath to look for the descendants of <bibl> within <advertisement> shows that other bibliographic elements such as <publisher> and <edition> may appear. If an encoder is marking up an advertisement that also contains the price of the book, he or she may be unsure about whether that should be treated as part of the <bibl>. Searching for the <measure> element (used to encode prices) to determine whether these are within or immediately following <bibl> elements in advertisements shows that, in fact, both cases are common. Looking more closely at specific examples from the collection will reveal that we follow each text's own lead on whether prices are part of bibliographic references; where these are on the same line as the rest of the reference and where the typography indicates that they are part of the same unit as other bibliographic details, we encode <measure> within <bibl>. On the other hand, if the prices are on a separate line and typographically distinct from the advertised book's bibliographic information, then <measure> is encoded outside of <bibl>. Taking a large-scale view of texts in WWO helps us balance the technical regularization of XML with each text's linguistic and semantic nuances. Sometimes, though, the linguistic or semantic make-up of a text resists the schema in a way that requires us to reconsider the schema. Our close reading and document analysis of a text can ask us to reflect on our current and past encoding practices. Moments where these come into conversation can lead to large-scale changes to both the schema and previously encoded and published texts. Distant encoding, then, is a negotiation between each text and the corpus. The two mutually inform each other and work dialogically.

There are many tools that the WWP uses to support distant encoding; we will focus here on those that influence our day-to-day encoding practices most significantly: XPath, XSLT, XQuery, and oXygen. Oxygen is an XML-aware text editor with XPath capabilities. It provides us with a virtual environment in which we transcribe and encode texts into the tree model structures of XML. Oxygen also enables validation, a process that checks whether the encoding is consistent with the WWP's schema, and it suggests appropriate tags, attributes, and values and provides

error messages if documents are ill-formed or invalid.¹⁸ We now take these tools for granted, but in its early years (before SGML/XML software was readily available) the WWP used cumbersome workarounds, such as a tag-counting process to determine each text was properly encoded. If the number of start tags and end tags did not match, encoders would know that one was missing (though not necessarily which one!). Oxygen helps our encoding team to address these kinds of issues as part of the encoding process itself.

Oxygen also supports a variety of related XML technologies, including XPath, XSLT, and XQuery functionality. XQuery and XSLT are two tools we use to run systematic, encoding-based reports at the WWP; XQuery is an XML-based query language, and XSLT is a language for manipulating and transforming XML data. WWP staff use these to make programmatic updates to the project's various TEI collections, and encoders can also use these tools to query those same collections to see how elements are used in practice, not just how our documentation says they *should* be used. XPath is another language for searching and navigating XML data. Together, these are powerful tools for engaging not just with markup, but with the ever-evolving process of marking up. Because they treat the XML document as a tree structure, rather than simply as a sequence of characters, XPath and XQuery enable searches based upon that structure: for instance, as above, to look at <measure> both as a descendant and a sibling of <bibl>. These tools can work both at the level of the individual text and at the level of the collection, providing us with a panoramic and extra-textual sense of encoded patterns in the corpus and constituting the core functions of our "distant encoding".

This new collection-level perspective has in turn changed the ways encoders think about texts, now that they can query the entire textbase. Rather than treating textual phenomena as unique to a single text, we often discuss our encoding in a more global sense. A question intended for one text might relate to a cluster of other texts that feature the same phenomenon. For example, Chelsea Clark, an encoder at the WWP, has done extensive work with XPath and regular expressions to discover instances of textual notes that are used to indicate the contents of a section, a phenomenon that possesses characteristics of both <note> and <argument> elements. Our documentation defines arguments as rhetorical devices that are "descriptive of the contents of the text that follow them", while notes include editorial and authorial footnotes, endnotes, marginal notes, and inline notes. The question of how to treat arguments when they are formatted as notes, and of whether to create a taxonomy of notes (distinguishing, for example, between notes that provide citations, describe contents, or offer more information on subjects discussed in the main text), has been an ongoing discussion since 1998. At that time, it was decided to continue encoding arguments as notes because there was not yet an efficient method of searching through the database to create such a taxonomy, nor was there a clear sense that the benefits of maintaining such a taxonomy would outweigh the labor involved.

However, this question of taxonomizing notes re-emerged while Clark was encoding Lucy Hutchinson's 1679 *Order and Disorder*, in which marginal notes appear regularly, both pointing to relevant biblical passages and briefly describing the contents of the poem. Using our collection-level XPath functions to query the entire textbase, it was now possible to reconsider the labor and payoff of developing special handling for marginal notes whose function overlaps with arguments. The research process that Clark followed shows how XML-aware tools can be used to ask and answer encoding questions. She began by sifting through all the notes in the database and selecting only the ones that might include arguments. Her XPath removes all the notes that are unlikely to contain arguments, by eliminating those that contain bibliographic citations, quotes, titles, dates, or other features indicating that the note served a bibliographic function; it also eliminates <note

18. For more on validity and well-formedness in XML, see Birnbaum (2015).

type="WWP">, which designates a purely internal note. After removing all the WWP notes that are unlikely to contain arguments, there were still 4,052 notes to examine. To narrow these results, Clark focused first on Lanier's 1611 *Salve Deus Rex Judaeorum*¹⁹ as a case study and found that most arguments tend to start with the same set of prepositions and articles ("the", "a", "of", and "to"). By combining searches for these terms with the XPath above, Clark located a number of example notes that describe textual contents and also identified particular texts in which this feature appears regularly. At the moment, this project is still in a phase of data collection and discussing whether to make schema changes that would taxonomize notes by their contents and functions. Clark's work thus highlights how even published texts remain responsive to ongoing document analysis. Her work incorporates new modes of markup that begins with document analysis and then folds in corpus analysis.

Another long-term encoding question has recently shifted into the phase of making corpus-wide changes and reconfiguring the schema. The WWP had initially followed standard TEI practices for the time in using the @type attribute to categorize stage directions, which are encoded with the <stage> element. The suggested values for @type on <stage> are "enter", "exit", "setting" and several others, including "mixed" for cases when more than one type of action is at stake.²⁰ In 1999, Syd Bauman, Senior XML Programmer/Analyst at the WWP, wrote in our discussion list that "mixed" as an attribute value "leaves a bit to be desired", because it "turns out that many, many stage directions contain more than one of the other types, and thus would get assigned 'mixed'". This makes searching "useless" as there is no finer granularity to query what might be included in "mixed" stage directions. Bauman also points out that "'mixed' presents so little information" that "encoders may feel compelled to try to sort out when to use 'mixed' versus when to figure out which type is most prominent or important for a given stage direction". Both decisions (choosing "mixed" or selecting the most prevalent value) lead to some information loss (Bauman 1999).

Although this problem was recognized in 1999, there was no available way to work with the 40,320 potential combinations of different values that might be used for @type on <stage>. At the time, the WWP decided to "continue to use 'mixed' liberally" until a better mechanism for encoding multiple values for this attribute presented itself. Now that we can easily traverse the XML trees within the entire textbase, we have revisited the issue of mixed values for @type on <stage> and have decided to use multiple values rather than "mixed".²¹ As a first stage in updating the encoding, Bauman changed all "mixed" values on @type to "UNKNOWN" as an efficient way to locate them; an encoder then reviewed each of these and added appropriate values. The end result is that we can now more precisely describe stage directions which contain multiple types of information, such as an entrance that is accompanied by a description of how characters appear and a relation of the actions they perform onstage:

19. This text has been published in WWO at: <http://www.wwp.northeastern.edu/texts/lanier.salvedeus.html>.

20. See <http://www.tei-c.org/release/doc/tei-p5-doc/en/html/ref-stage.html>.

21. The WWP initially added a new attribute, @kind, on <stage> because the TEI did not allow @type to have multiple values at that time. Soon after we completed our first review of the encoding, an update to the TEI permitted @type to have multiple values on <stage> and we adjusted our own practices to use @type again.

```
<stage type="entrance business modifier"
  rend="align(center)slant(italic)">Enter <persName
  rend="slant(upright)">Clarina</persName> and
  <persName rend="slant(upright)">Ifmena</persName>, dreft like one another in every thing,
  <lb/>Laughing and beholding one another.</stage>
```

Figure 5: Encoding of a stage direction with multiple values for @type, from Aphra Behn's *The Amorous Prince, or, the Curious Husband* (1671)²²

Here, the characters' "entrance" is noted as well as some stage "business" ("Laughing and beholding one another") and a "modifier" describing the characters' appearance ("drefit like one another in every thing").

The ability to represent an array of dramatic interactions has also made it more practicable to expand on our list of accepted values for @type on <stage>. For example, we now have "prop" for stage directions that simply name the objects that appear onstage; "present", which gives the names of characters who are present in a scene without indicating their entrance; and "remains" for any character that stays on stage for dramatic effect or a soliloquy. These additions reflect our continued refinement of the taxonomies we use to represent texts through encoding, both drawing on tools like XPath and XQuery to adjust past encoding and also supporting more detailed inquiries in the future. We read texts differently with a new understanding of these recent schema changes.

As these examples show, the tool apparatus at the WWP multiplies the way we read texts and corpora together and functions as part of a wider, ongoing research structure. It allows us to make our data cleaner and smarter through the gradual dialogue between text analysis and corpus analysis. While we use the tool apparatus to make our data smarter and cleaner, we also work to facilitate future WWP encoders as they continue to clean and improve the data as well.

4 The Human Apparatus

Human(ist)s are at the heart of all branches of the WWP's work, whether considering the evolution of the WWP's data in response to research and changing standards for text representation, reflecting on the connection between information organization and human labor, or contemplating the ways in which technological advances have shaped the project's practices. We have developed research initiatives, generated data, learned technologies, encoded texts, curated exhibits, and worked together to ensure that the present textbase reflects our current understanding of those subjects most central to our mission: gender, textuality, and digital representation. The ways in which the project has developed, outlined in the sections above, are mirrored by the evolution of training, encoding, and research practices of WWP employees, volunteers, and working groups.

The WWP training documents, in fact, offer a genealogy of the project's intellectual labor. Successful training of new encoders requires a degree of familiarity with the history of the project's technologies and encoding refinements, some of them 30 years in the making. The WWP has an archive of training materials dating back to 1996—these evolved from very early hotsheets on specific tasks into a more comprehensive and coherent curriculum that starts with basic orientation and works towards more specialized guidelines. Current encoders have the benefit of drawing on a greater amount of past materials, as well as better tools for navigating the whole collection. That said, the project's move from Brown to Northeastern revealed how crucial the human element is

22. This text has been published in WWO at: <http://www.wwp.northeastern.edu/texts/behn.amorous.html>.

to this history. The move imposed a hiatus during which we were unable to hire new encoders, and when encoder training was restarted at Northeastern, much had changed about our encoding environment: tools, work processes, and approaches to training and oversight were all adapted to the new working environment. As a result, our training materials had to be updated to reflect these changes. During this transition period, the new cohort of encoders had to draw heavily on the project's documentation as a guide to practice. They also worked in a quasi-archaeological spirit to rediscover the project's recorded history: notes from meetings and discussions of earlier encoding decisions and rationales, recorded in the project's electronic discussion list. This process immersed the encoders in the documentation and the project's work history, re-establishing a disrupted intellectual continuity; most importantly, it also allowed the new cohort to provide feedback and shape future training materials.

The training process begins by introducing new encoders to XML and the TEI, after which they begin by learning document-level encoding practices with exercises and homework assignments in oXygen.²³ This reflects long-standing practice; in the archive of training guides, a 1998 document reveals that the initial training of new encoders has consistently started with a document analysis of the first text assigned (or selected) for encoding. The encoder is advised to skim the text to become familiar with the overall contents, to sketch the structure of the text in outline form, to list the divisions that might fall within the main sections, and to note any puzzling or anomalous textual features that may need special treatment. Document analyses (then and now) are often discussed at encoding meetings or with the project manager. This process allows for more efficient encoding; it helps to uncover idiosyncrasies and general patterns early on, so that the encoding of the text can be planned out in advance to avoid wasted effort.

Analysis of the training documents also reflects changes in technology used by the WWP: the 1998 instructions refer to SGML files, as opposed to XML files. The archive of training documents further reveals how tools have altered the WWP's proofreading processes. Changes in these processes are especially illustrative of the ways technology has informed the project's human labor. While the overall process—each encoded text receives two rounds of proofreading, corrections entry, and then corrections checking—has been the same since 1996, improved mechanisms for programmatically detecting errors have enabled the WWP to pilot processes in which the text's encoder also performs a series of systematic checks prior to printing and proofing. This process allows the encoder to assume preliminary proofreading responsibilities, and will, hopefully, make a second round of proofing unnecessary in many cases.²⁴

Finally, once an encoder is proficient with markup and typically after he or she has completed the initial capture of texts in a range of genres, the encoder begins learning strategies for corpus-level distant encoding, requiring technologies such as XPath, XQuery, and XSLT. Support for these more advanced topics includes in-person training sessions, formal training materials, and materials from the WWP's advanced seminars and workshops.

The work of the encoding staff is situated within the larger ecology of the Women Writers Project, including the permanent staff, the steering committee, and the project's research partners, external collaborators, and alumni. It is also embedded in the broader set of research questions that animate the WWP's more specialized initiatives such as the study of reception, readership,

23. For a sample of a recent training curriculum, please see <http://dsg.northeastern.edu/wiki/TrainingMaterialsFall2016>.

24. As with much of the work that the WWP does, this new proofing process also draws on past practices; in this case an earlier mechanism for programmatic error detection that became obsolete when the project switched editors from Emacs to oXygen.

and intertextuality. One important effect of this embeddedness is that it establishes intellectual continuity between the practical, editorial, and theoretical questions that arise in the encoding process, and the scholarly and interpretive work that is enabled by the resulting published collection. It is by now common to acknowledge the intellectual significance of data design and representation within digital humanities projects,²⁵ but the continuities are unusually striking in this case, where encoding and transcription decisions so directly and visibly affect the downstream analysis, and where those doing the encoding are also collaborating on the research that encoding supports.

Another important effect of this embeddedness is that this work reinforces the digital humanities research being undertaken by graduate students in other areas of their degree programs. In some cases, the connections are quite direct. For example, Elizabeth Polcha's work on Eliza Hamilton's 1796 epistolary novel, *Translation of the Letters of a Hindoo Rajah*, had important connections to her own research on eighteenth-century colonial literature. Nicole Keller's encoding of early issues of *The Ladies' Diary* expanded the scope of her dissertation, which focuses on evidence in astronomy texts of the long eighteenth century. William Quinn has built a Python-based XML parser for the WWP that can visualize the flows of literary influence within the WWO corpus; this work has folded back into his own research, examining intertextuality in an XML-encoded corpus of Modernist journals. More generally, the WWP's encoding work and the expertise it engenders contribute to the research environment for graduate students in Northeastern's Digital Humanities Certificate program,²⁶ many of whom undertake projects that engage with markup in more experimental ways. The WWP's ongoing inquiries into encoding methods and their supporting apparatus of documentation and data curation, which has evolved over such an extended period, provide a frame of reference within which students can orient themselves as they start to plan projects of their own.

5 Next Steps

As the project prepares for its fourth decade, its intermediate and long-term goals are focused on finding more powerful and scholarly ways of exploiting and exposing our data. Much of our ongoing work will be focused on encoding new texts for publication in WWO while continuing to refine our encoding practices, information structures, and publication systems, as discussed above. We will also continue to add new materials to Women Writers in Review and Women Writers in Context, so that all the texts we publish will be part of a network of linked content.

As the WWP's newest collection—released in November of 2016—Women Writers in Review represents an important part of the textual network we are building. We are planning to focus substantial attention on both corpus and interface development for WWiR in the next few years. In expanding the collection, we will prioritize improving representation of North American texts; we will also add additional formats and genres, such as commonplace books. We have some immediate goals for the WWiR interface, including more advanced search and navigation options, both static and dynamic visualizations, and user-generated thematic tagging. We are also gathering feedback on interface needs from our user community and from a group of teaching partners who are developing

25. See for instance Bauer (2011).

26. For more information on the certificate program, please see <https://www.northeastern.edu/cssh/english/graduate/graduate-certificate-in-digital-humanities/>.

assignments and activities with both the WWiR and WWO collections.²⁷

We are working now on updating the WWO interface to offer greater flexibility in its display—for example, enabling users to select whether they see errors or their corrections, abbreviations or their expansions, original or modernized orthography, and so on. We are also planning to make it possible for readers to act on encoding that marks titles, names of persons and places, quotation and dialogue, and various document structures in verse, drama, and prose. For example, we imagine users constructing visualizations of where poetry and prose appear in a text or set of texts, or building queries that take advantage of the information available in the markup to search for individual terms in certain contexts (find “grace” but only when it appears inside of a quotation), or even querying the markup itself (find all of the locations in which <castList>s appear).

Another WWP project that involves both expanded encoding and interface development is our collaboration with *The Almanacks of Mary Moody Emerson: A Scholarly Digital Edition* to pilot encoding and display for manuscripts.²⁸ Working with the *Almanacks* team, we have published sixteen “folders” of Emerson’s writing thus far. In the future, we plan to publish the remainder of Emerson’s *Almanacks*, expand the WWO corpus to include additional manuscript texts, and update the WWO display options to better accommodate manuscript materials.

In the fall of 2016, the WWP began work on Intertextual Networks, a three-year research project focusing on intertextuality in early women’s writing and funded by a grant from the National Endowment for the Humanities. As part of this project we are expanding our encoding to include explicit identification of the sources of quotations, allusions, and citations in the WWO collection. We are also developing a comprehensive bibliography of those sources, which we will make openly available at the WWP Lab.²⁹ We have assembled a team of research collaborators who will each pursue a research project engaging with materials from WWO, to be published in *Women Writers in Context*.³⁰ We will also be developing interface tools for exploring intertextual connections and are partnering with other projects focused on early women’s writing, such as the RECIRC project (The Reception and Circulation of Early Modern Women’s Writing, 1550–1700).³¹

From these initiatives, and the infrastructure they build on, several points emerge that may inform the development of other projects. First, as the WWP’s history demonstrates, a long-term research project will realistically need to make changes to its data, approaches, and tool set over time. What makes these changes tolerable—or even possible—is the deeper apparatus of documentation and clear intellectual rationale that can guide those changes in a coherent direction. This apparatus is what enables a prototype to mature into a genuinely long-term project, by treating each experiment or update as a process that can yield future insight in unforeseeable ways. Although historically the digital humanities field has been treated as intractably fast-moving and present-oriented, the decades-long history of projects like the WWP, Perseus, Orlando, the Walt Whitman Archive, and others shows that digital humanities can also operate at much longer scale.³² We have described here systems, practices, and tools that support this longer-term work, albeit perhaps at the cost of faster development in other areas. A second, related point of emphasis here is that

27. For more on the teaching partner program, please see <http://wwp.northeastern.edu/wwo/teaching/pedagogical-dev.html>.

28. See <http://marymoodyemerson.net>.

29. See <http://wwp.northeastern.edu/wwo/lab/index.html>.

30. For more on the work of our research collaborators, please see this list of abstracts: <http://wwp.northeastern.edu/research/projects/intertextuality/collaborators.html> and the posts categorized under Intertextual Networks at our blog: <http://wwp.northeastern.edu/blog/category/intertextual-networks/>.

31. See <http://recirc.nuigalway.ie>.

32. See <http://www.perseus.tufts.edu>, <http://orlando.cambridge.org>, and <http://whitmanarchive.org/>.

work of this kind requires us to create data that will be intelligible to future users –who may or may not be ourselves– and usable with future tools. As this history has shown, projects like the WWP were at work before there were suitable tools for many key operations and have survived profound changes in the tool set at all levels, including a shift from mainframe computers to local networks to cloud-based systems, as well as migrations across at least three generations of publication tools. Despite these evolutions, the WWP has been able to use and build its core collection in a form that remains essentially unchanged, thanks to the fact that SGML/XML and TEI are open standards for which there are open tools and a broad international research community. In this context, the documentation produced by long-term projects not only serves the local goals of organizational memory and consistent work practices, but also makes a contribution to broader community goals and histories.

Finally, what this history illustrates is the challenge of modelling a large and growing set of texts, particularly while our understanding of what it means to model texts in digital form is rapidly changing. In the early 1990s, digital editors debated the merits of image-based and markup-based editions; in the early 2000s the emergence of Web 2.0 focused attention on contributory editions and community-driven annotation; in the current decade, linked open data offers an entirely new paradigm for representing cultural and textual networks. That historical sequence could be understood as a series of technological developments, but it is simultaneously a series of inquiries into how we construct, circulate, and engage with texts. Digital humanities, distinctively, understands those histories as being tightly coupled and its characteristic work practices – illustrated in this case study– span and comprehend both.

6 Works Cited

- Bauer, Jean. 2011. 'Who You Calling Untheoretical?', *Journal of Digital Humanities*, 1.1 <<http://journalofdigitalhumanities.org/1-1/who-you-calling-untheoretical-by-jean-bauer/>> [accessed 20-07-2017]
- Bauman, Syd. 1999. 'Encoding Meeting Minutes, 1999-04-14', *WWP Encoding Discussion List* [internal electronic discussion list], 14/04/1999 [accessed 20-07-2017]
- Billey, Amber; Drabinski, Emily; Roberto, K.R. 2014 'What's Gender Got to Do With It? A Critique of RDA Rule 9.7', *Cataloguing and Classification Quarterly*, 52.4: 412-421 <<https://doi.org/10.1080/01639374.2014.882465>>
- Birnbaum, David. 2015. 'What is XML and Why Should Humanists Care? An Even Gentler Introduction to XML' <<http://dh.obdurodon.org/what-is-xml.xhtml>> [accessed 20-07-2017]
- Burnard, Lou; O'Brien O'Keefe, Katherine; Unsworth, John (ed.). 2006. *Electronic Textual Editing* (Modern Language Association and Text Encoding Initiative Consortium) <http://www.tei-c.org/About/Archive_new/ETE/Preview> [accessed 20-07-2017]
- Flanders, Julia; Jannidis, Fotis. 2016. 'Data Modeling', in *A New Companion to Digital Humanities*, 2nd edn, ed. by Susan Schreibman, Ray Siemens and John Unsworth (Oxford: Wiley; Blackwell), pp. 229-237
- Jed, Stephanie. 1989. *Chaste Thinking: The Rape of Lucretia and the Birth of Humanism* (Indianapolis: Indiana University Press)
- King, Katie. 1991. 'Bibliography and a Feminist Apparatus of Literary Production', *Text*, 5: 91-103
- NINCH (National Initiative for a Networked Cultural Heritage). 2002. *The NINCH Guide to Good Practice in the Digital Representation and Management of Cultural Heritage Materials* <<http://www.ninch.org/guide.pdf>>
- Reiman, Donald. 1988. 'Gender and Documentary Editing: A Diachronic Perspective', *Text*, 4: 351-59
- Schöch, Christof. 2013. 'Big? Smart? Clean? Messy? Data in the Humanities'. *Journal of Digital Humanities*, 2.3 <<http://journalofdigitalhumanities.org/2-3/big-smart-clean-messy-data-in-the-humanities/>> [accessed 20-07-2017]
- Smith, Martha Nell. 2004. 'Electronic Scholarly Editing', in *A Companion to Digital Humanities*, ed. by Susan Schreibman, Ray Siemens and John Unsworth (Oxford: Blackwell) <<https://doi.org/10.1002/9780470999875.ch22>>
- Sutherland, Kathryn; Pierazzo, Elena. 2012. 'The Author's Hand: From Page to Screen', *Collaborative Research in the Digital Humanities*, ed. by Willard McCarty and Marilyn Deegan (London: Ashgate), pp. 191-212
- TEI Consortium. 1993-2017. *Guidelines for Electronic Text Encoding and Interchange* <<http://www.tei-c.org/P5/>> [accessed 20-07-2017]
- Terras, Melissa. 2013. 'On Changing the Rules of Digital Humanities From the Inside', *Melissa Terras's Blog*, 27-05-2013 <<http://melissaterras.blogspot.com/2013/05/on-changing-rules-of-digital-humanities.html>> [accessed 20-07-2017]
- Whitelaw, Mitchell. 2015. 'Generous Interfaces for Digital Cultural Collections', *Digital Humanities Quarterly*, 9.1 <<http://www.digitalhumanities.org/dhq/vol/9/1/000205/000205.html>> [accessed 20-07-2017]