



the yellow nineties online

edited by Dennis Denisoff and Lorraine Janzen Kooistra

“I STILL PREFER BOOKS”: NARRATING THE GENTLE INTRODUCTION TO XML MARKUP

By Constance Crompton (Project Manager of *The 1890s Online*)

A version of this paper was presented at the Digital Humanities Symposium: Visualizing the Archive held at Ryerson University April 23, 2010 to mark the official inauguration of The 1890s Online. The name of the project was changed from The 1890s Online to The Yellow Nineties Online in December 2010.

In April 2010, as we were preparing to launch *The 1890s Online*, we had three visitors come to the 1890s Digital Studio. They were on a campus-wide tour and so could only stay for a few moments. The principal investigators, Lorraine Janzen Kooistra and Dennis Denisoff, described the scope of the project to them and presented our newly minted image html. Two of our three visitors were very enthusiastic. The third, though friendly, hung back, away from the screens we were using to exhibit the site. As she left I overheard her say that she liked the look of *The 1890s Online* but, as she said of herself, “I still prefer books.”

The phrase “I still prefer books” provides a lens through which to consider *The 1890s* team’s experience encoding in TEI in the context of digital humanities pedagogy, collaboration, and publication. The purpose of this paper is to introduce some of the conceptual challenges that the team has faced when marking up texts in TEI. The unspoken concerns implied by the comment “I still prefer books”—that computers are hard to use; that it is unclear why we need digital humanities projects; and that digital publishing may eclipse traditional modes of storing and retrieving text—are ones that *The 1890s* team has itself had to address in its site design and encoding practices.

TEI Markup – Its Practice and Purpose

In 2005 Dennis Denisoff and Lorraine Janzen Kooistra presented their plan for *The 1890s Online* project at the inaugural workshop of the Networked Infrastructure for Nineteenth-Century Electronic Scholarship (NINES) at the University of Virginia. They returned with two things to report: first, the NINES team was enthusiastic about the project; and second, we needed to encode our texts using TEI, the extensible markup language of Text Encoding Initiative. TEI is a descriptive markup language that allows one to overlay semantic meaning onto a text in the form of computer- and human-readable tags. TEI is comprised of scholars from multiple North American and European academic institutions who have created a markup language in keeping with the needs of humanities scholars.

When my passion for fin-de-siècle culture brought me to *The 1890s Online* in 2006, I arrived wholly unaware of TEI. As a novice encoder, I was thrilled at the prospect of helping the principal investigators publish a scholarly edition of *The Yellow Book* online. Like any user who had more experience with texts in a codex form than in an online form, during my first months I had to accept that the project's digital tools and processes would eventually make conceptual sense to me. Like many coders, I learned how to encode in TEI before I learned the rationale and purpose of markup. Zachary Devereaux, the head research assistant on the project in 2006, sat me down in front of his text editing software, TextEdit, and explained which TEI tags the project was using. He admonished me, above all else, to open and close my tags properly. The encoding process took enormous concentration; since we did not have any schema files, TextEdit could not tell us whether or not our code was valid. Another research assistant, Ruth Knechtel, gave me a copy of the TEI's "Gentle Introduction to XML" to acquaint me with the theoretical rationale for our markup language. I could decipher only about a third of the Introduction's explanations. I still did not know why we were

inserting tags into the text, but I was certain of one thing: for every opening tag, there had to be a closing one.

Later that year, the project invested in an XML editing software application, Oxygen, which prompted us to use the right tags and to format them correctly. Our initial attempt to proofread our XML was rather dispiriting. Our XML was suffused with syntactical errors. As you can see in Figure 1, Oxygen marks all incorrect tagging with red lines, highlighting mistakes we had overlooked when encoding in TextEdit. And of course, we could not get our documents to validate. To find the answer to our lingering questions and to learn more about how to encode in TEI, Ruth and I attended the University of Victoria's Digital Humanities Summer Institute in June 2007 for a workshop run by Julia Flanders and Syd Bauman of Brown University's *Women Writers Project*.

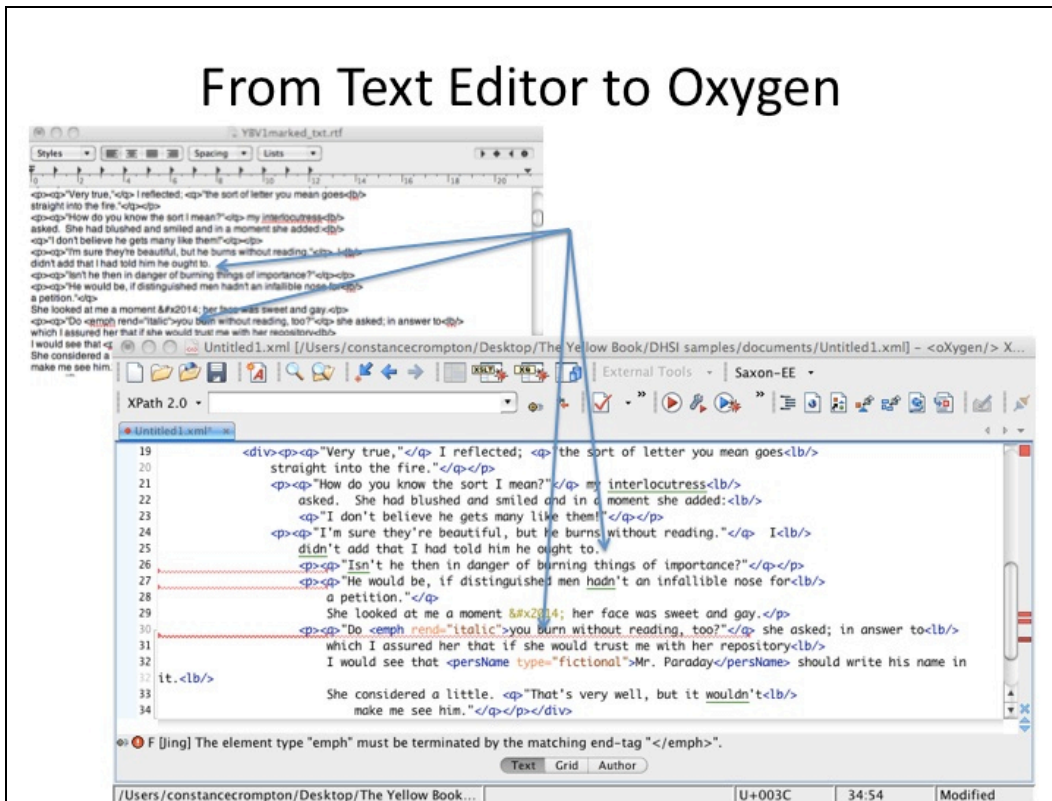


Figure 1

At the Summer Institute, I learned that XML stands for eXtensible Markup Language, a descriptive language used to mark up the meaning of a text. XML does not transform the text in any way. Unlike basic html, XML does not provide browsers with instructions for how to display text. XML's central purpose is to encode meaning to the text. The semantics of conventional printed text is generally obvious to humans but imperceptible to computers. As shown in Figure 2, XML lets encoders be explicit about meaning. XML is not a proprietary language, nor does it require special editing software (although editing in Oxygen certainly makes encoding easier). Since XML is extensible—that is, it does not have a pre-defined set of tags—each project team is free to design tags that best describe its particular text. The unbounded freedom to invent tags, however, makes it hard to share documents with digital scholars beyond the context of any individual project. The TEI has responded to this dilemma by devising a set of tags to increase the interoperability of XML-encoded documents. By using a standard set of tags, and a standard way of modifying them to meet each project's needs, scholars can more readily share marked-up text with other TEI users.

Semantic Markup

Moore, George. *Héloïse and Abélard*. London: Cumann Sean-eolais na h-Eireann, 1921.

```

<biblStruct>
  <monogr>
    <author><lastName>Moore</lastName><firstName> George</firstName></author>
    <title level="m"><persName type="fictional">Héloïse</persName> and <persName
type="fictional">Abélard</persName>.</title>
      <imprint>
        <pubPlace> London:</pubPlace>
        <publisher> <foreign xml:lang="ga">Cumann Sean-eolais na h-Eireann</foreign>,</
publisher>
        <date> 1921.</date>
      </imprint>
    </monogr>
  </biblStruct>

```

Figure 2

Readers are often unaware of how effortlessly they interpret, or "decode," the meaning signaled by fonts or formatting. For example, scholars and students are likely to know from experience that the first line in Figure 2 is a bibliographic entry. Drawn from the project's biography of George Moore, the entry is for his monograph *Héloïse and Abélard*. One cannot assume, however, that all users and computers will be able to gather the meaning of this string of characters. The markup does not change the text, but it does clarify its meaning. In Figure 2, for example, the <biblStruct> tags indicate the entry's status as a bibliographic record. The <monogr> tag signals that *Héloïse and Abélard* is a monograph, not a pamphlet or magazine article. Finally, the <foreign> tag, which carries the attribute "xml:lang" and the value "ga," signals that the publisher's name is in Gaelic.

The 1890s Online's encoded text can be manipulated to generate new research questions. As Franco Moretti has noted, encoded material helps us answer some questions, but almost inevitably provokes others (4). For example, if the citation in Figure 2 came from an exhaustive bibliographic list of everything Moore ever wrote, from love notes to articles, the encoded text could be used to organize the list in a meaningful way. A program could scour the list for monographs, correlate them to publishing dates, and present them on a timeline, showing the user when Moore's works were issued and reissued. The timeline would not offer a causal explanation of Moore's waning and waxing popularity, but it would provide the groundwork for an investigation into the demand for Moore's books.

The number of descriptive tags that any project uses is a matter of editorial discretion. When first marking up Volume One of *The Yellow Book*, *The 1890s Online* team was inclined to mark up in detail. Eager to do justice to *The Yellow Book's* semantic substance, we tagged place names, geographic regions, dates, each character's fictional status, quotation marks, and foreign words. We even imagined a second layer of coding for each document. This extensive markup of course increased the time it took to encode each text, but we envisioned our efforts meeting the needs of as many of the scholarly users we could imagine for

The Yellow Book's text and images. Our original encoding for Volume One would allow scholars to trace the London boroughs mentioned in *The Yellow Book*, or to visualize the distribution of fictional and nonfictional characters in the magazine, or to rank the contributors based on their foreign-language use.

In 2008 we were confronted by a conundrum: for users who wanted to use *The 1890s Online* in the ways listed above, our markup was useful. However, the time it took to encode each document decreased the utility of the site for users who wanted to access the full print run of *The Yellow Book* and the associated scholarly content. In early 2009 *The 1890s Online* adopted a simpler tag set, one that does not attempt to imagine all the inquiries users might make of the text.

Digital Humanities – Practice with Purpose

I have introduced TEI and some of the semantic problems that *The 1890s* team has had to face in marking up its editions. What I wish to address now is what it means to us to be encoding *The Yellow Book* and other 1890s material in the context of the digital humanities. There is no official code of ethics for digital humanities, but there is a spirit or an ethos that is embodied in digital humanists' commitment to teaching, teambuilding, and representing culture. Scholars in this field of study are particularly willing to give explanations, share information, and help novices in the pursuit of humanities questions. Most of the team has had direct experience with the pedagogical and collaborative impetus that has given rise to the field's implicit code of ethics. For example, in 2005, *The 1890s Online* principal investigators were met with considerable enthusiasm and support when they first proposed the project to NINES. In 2007, when I rather naively asked what a path was, my DHSI instructors explained the concept to me in a way that left me feeling enthused at having grasped it, rather than crestfallen at not having understood the concept earlier. Laura Mandell, Associate Director of NINES, has been incredibly open with her xslt teaching resources and has even sent Ruth Knechtel screencasts when Ruth could not get our xslt files to work. This ethos of openness and community support has material consequences that we have to

take into consideration at all levels of the project: from what we will encode and how we mark it up, to how we design the front end of the site for optimal navigation and how we work together as a team.

It is not enough to design a site that is useful to users with a wide range of computing experiences and aptitudes. We also aim to extend the same courtesy and aid to our teammates as others have extended to us. As Lorraine Janzen Kooistra and I have specified elsewhere <http://www.1890s.ca/HTML.aspx?s=Elec_Scholarship.html>, digital humanities projects like *The 1890s Online* require that we work in “teams (sometimes very large teams) of people with varied and distinct skill sets, knowledge bases, time commitments to, motivations for, and even understandings of, the project.” Digital projects require collaboration, as Julia Flanders has noted, even if the institutional structures are not in place to facilitate teamwork. Beyond expediting access to textual material, any project requires the particular expertise of each collaborator to represent a text in ways that are intellectually provocative (12, 17).

Commitment to collaboration also requires working with our users. Underlying the statement “I still prefer books” is the intimation that computing is complicated, intimidating, and often incomprehensible. It is true that computing is difficult, particularly when users and producers have no community to help them navigate through the computational universe. Jenna McWilliams, a doctoral student in Learning Sciences at Indiana University, has used the term “tinkering” to describe the many attempts that it takes to solve computing problems (32). Before booting up or logging in, nascent digital humanities scholars need the freedom to tinker with the support of a responsive community. I was particularly aware of the importance of community building when, in the summer of 2010, I designed a blog-like site that now serves as the central training and communication hub for the principal investigators and the research assistants on *The 1890s Online* team. Everyone has equal access and administrative permissions on the communication and training website. Although it officially serves as a collaborative space online to share our

knowledge with one another, it also provides a space to tinker and experiment with fundamental rules of encoding while building new pages or responding to one another's posts.

Finally, implied by the observation "I still prefer books" is a fear that digital humanities is a threat to the printed word. There has been a rush from print to digital publishing and, with it, an upending of the revenue models that support traditional publishing practices. In the last two years, we have witnessed a remarkable restructuring in publishing; for example, The University of Michigan started offering digital editions of most of its texts with print versions available only on demand. E-readers have become commonplace. Print editions of daily newspapers, such as *The Montreal Gazette* and *The Ottawa Citizen*, seem in danger of disappearing completely. The publishing landscape is certainly changing, but the threat to printed material is not coming from the digital humanities. Digital humanities resources and tools are more than just replicas of books online. In addition to providing access to previously scarce material, scholarly websites offer academic researchers and other users the opportunity to develop an enriched understanding of the text by providing expedited access to contextual material. *The 1890s Online* team's challenge for the future is to maintain a site that is useful to digital humanities experts as well as novices (who may prefer books), to provide collaborative learning and teaching opportunities to humanities scholars, and to offer access to analytical tools and material that give us new insights into 1890s periodicals and their publishing and cultural contexts.

© 2010, Constance Crompton, York University

Constance Crompton is a research associate and Project Manager at *The 1890s Online*. She also co-directs *Lesbian and Gay Liberation in Canada 1964-1975: An Online Research Database and Community Resource*, a project scheduled for online publication in 2012. She is a doctoral candidate in York and Ryerson University's joint program in Communication and Culture.

Works Cited

- Flanders, Julia. "The Productive Unease of 21st-century Digital Scholarship." *Digital Humanities Quarterly* Summer (2009): 1-27. Print.
- McWilliams, Jenna. "Closing Remarks for the AERA 2010 Annual Meeting." *Understanding Complex Ecologies in a Changing World*. Denver, CO, 2010. Print.
- Moretti, Franco. *Graphs, Maps, Trees: Abstract Models for a Literary Theory*. New York: Verso, 2005. Print.
- Text Encoding Initiative. "A Gentle Introduction to XML - TEI P5: Guidelines for Electronic Text Encoding and Interchange." Web. 20 May 2010.