

Willett, Perry. "The Victorian Women Writers Project: The Library as a Creator and Publisher of Electronic Texts." *The Public-Access Computer Systems Review* 7, no. 6 (1996). (Refereed Article)

1.0 Introduction

Librarians, who are busy building new digital library services while maintaining traditional print-based services, are not generally looking for major new job responsibilities. Nevertheless, using my experience as the general editor of the Victorian Women Writers Project (VWW Project), [\[1\]](#) I would like to outline a number of reasons why librarians need to be knowledgeable about and involved with the creation and publication of electronic texts. [\[2\]](#)

Faculty and students are beginning to view the World-Wide Web as a valuable resource for research and instruction. As they become more familiar with the Web and more convinced of its value, they will want to create their own electronic resources. They will require advice and assistance for many technical, procedural, and theoretical problems. It is likely that they will be unaware of efforts to create open and archival standards for character-based electronic texts.

Librarians have been involved in the creation of the Text Encoding Initiative's (TEI) *Guidelines for Electronic Text Encoding and Interchange*, [\[3\]](#) and they have a long tradition of creating and supporting other open standards for electronic information interchange (e.g., MARC).

2.0 The Victorian Women Writers Project

The Victorian Women Writers Project began in the spring of 1995 when an undergraduate student approached the staff of the Library Electronic Text Resource Service (LETRS) at Indiana University Libraries, asking about the availability of certain electronic texts. The student had taken a course in Victorian poetry, and had used *The English Poetry Full-Text Database* produced by Chadwyck-Healey [\[4\]](#) to look for other works by some of the poets that he had read in class.

The English Poetry Full-Text Database is based upon *The New Cambridge Bibliography of English Literature (NCBEL)*, [\[5\]](#) and it includes writers considered canonical when *NCBEL*'s predecessor, the *Cambridge Bibliography of English Literature*, was first published in the 1940s. Many writers now studied in classes and included in anthologies are not in the *Cambridge Bibliography*, and, consequently,

they are not in *The English Poetry Full-Text Database*. (This is not a criticism of this database--it contains the works of over 1,300 poets.)

The student wondered whether we planned to add other poems to this database. We explained that it was a commercial publication, and, as such, we had no plans to add to it. He then asked what he could do to make these works available electronically. Sensing that this was an opportunity for LETRS to become involved in producing electronic texts, I worked with him and Professor Donald Gray, the senior Victorian literature specialist in the English Department, to draw up a list of authors who were important or interesting enough to be considered, but whose works were not in the Chadwyck-Healey database. A quick scan of this preliminary list showed that most of the writers were women, so we decided to focus the project on their work.

I worked with the student to train him in textual encoding, editing and proofreading skills, and poetic forms of the late 19th century. He was very bright and had good proofreading skills, which proved to be very important, but he was largely innocent of both computers and poetry. He and I finished nine works by October 1995 for the opening day collection of the VWW Project.

3.0 Investigation of Electronic Text Formats and Standards

The decision about what electronic format to use for the VWW Project's texts was of critical importance, because it would determine their longevity and usability. The choice had to be made between creating either graphical images of texts or character-based transcriptions of them.

3.1 Images

Since creating graphical images of pages only requires scanning them, it is initially very easy. However, this method raises serious long-term storage and retrieval issues. There are currently no acceptable archival image format standards. The GIF format has been claimed by CompuServe as its proprietary format. Although JPEG is an open standard, it involves data loss, which complicates its archival use. TIFF is archival standard, but it cannot be used on the Web without converting it to GIF or JPEG, or without requiring a special TIFF viewer. Collections of graphical images provide access that is only slightly better than the original text, because the words in the electronic version cannot be searched. As anyone who has used the Web knows, image files can be very large, and they take a long time to transmit.

3.2 Character-Based Representations

Character-based files provide much better access, and they can be transmitted much more quickly than image files; however, they also present serious problems.

3.2.1 Proprietary Formats

It is essential that the format of electronic documents be carefully considered, for there are many choices, ranging from proprietary formats, such as WordPerfect, to nonproprietary formats, such as HTML. Proprietary formats are inappropriate for the long-term storage and use of electronic texts because of changing hardware, operating systems, and application software. Documents created using MultiMate a decade ago can no longer be used, because the company has long since ceased to exist. Most contemporary word processors cannot convert Multimate files to their own proprietary format. Ten years is a long time for software companies, but a brief time in library terms.

In general, documents encoded in a program's proprietary format can only be used with that program for the particular uses which that software allows. For example, documents created with Adobe Acrobat can only be used with that software. A user cannot create a concordance of the text or use other textual analysis software on it.

Even when conversion is possible between formats, such as between WordPerfect and Microsoft Word, it is rarely simple or flawless. Librarians need to ensure that the representation format will allow the text to be used under a variety of operating systems and on a variety of hardware platforms. Even texts distributed on CD-ROM, bundled with their own search software, may only work in a single operating system/hardware environment.

3.2.2 Standards

Librarians cannot hope to base research collections of electronic texts on proprietary formats. If we hope to be able to use texts created today in the future, their format cannot be based on proprietary software. Librarians need to insist upon the use of open standards for formatting electronic documents.

3.2.2.1 ASCII

ASCII text is a widely embraced standard. Project Gutenberg and other similar projects insist that texts must be in ASCII form, without any tagging (i.e., unencoded ASCII). However, the basic ASCII character set is limited to a narrow range of characters and numbers, and it provides for no standard way of representing any other characters. The extended ASCII character set provides for some frequently used Western accented characters, but this is not common to all platforms. Also, unencoded ASCII provides no standard way of documenting bibliographic sources, creating intra- or inter-textual links, or indicating the various structural, syntactic, or semantic elements of a text. For example, if one wished to find all occurrences of a particular word in a collection of unencoded ASCII electronic verse, the computer would be unlikely to be able to distinguish between bibliographic, other extra-textual information, and the text itself. It could not even tell in a standard way where a poem begins and ends. Consequently, simple unencoded ASCII is not a good representation format for most electronic texts.

3.2.2.2 SGML

The Standardized Generalized Markup Language (SGML) is the best and the most widely accepted

representation format for electronic texts. [6] SGML guidelines for markup languages require the use of angle brackets to enclose the name of elements, which are commonly called "tags." Entity references are used to replace files, character strings or words, or more commonly, accented or foreign characters (e.g., an accented "e" is represented by é). Attributes are used to further refine and define elements (e.g., within the <A> tag in HTML, HREF and NAME are commonly used attributes). All of these elements are defined within a Document Type Definition (DTD). Along with many other markup languages, both HTML and the TEI *Guidelines* are forms of SGML, and they have the features described above. (HTML has a Document Type Definition called HTML.DTD, but many users are not aware of its existence because it is hidden by Web browsers.)

3.3 Representation of a Poem Using Three Standards

Yet, even within the domain of open standards, there are choices. HTML is a widely accepted open standard following SGML rules, but it cannot be considered an archival format for literary texts.

3.3.1 ASCII

Figure 1 shows an unencoded ASCII version of Amy Levy's "Straw in the Street," a poem from the VWW Project's collection.

Figure 1. Text of "Straw in the Street"

Straw in the Street.

STRAW in the street where I pass to-day
Dulls the sound of the wheels and feet.
'Tis for a failing life they lay
Straw in the street.

Here, where the pulses of London beat,
Someone strives with the Presence grey;
Ah, is it victory or defeat?

The hurrying people go their way,
Pause and jostle and pass and greet;
For life, for death, are they treading, say
Straw in the street?

The poem consists of a title (which should appear in bold type) and three stanzas. The first word of the

first line should also be in bold type. The last line of stanzas one and three are indented.

3.3.2 HTML

Figure 2 shows an HTML version of the poem.

Figure 2. HTML Version of "Straw in the Street"

```
&#160;&#160;&#160;&#160;<H2>Straw in the Street.</H2>
<P>
<B>STRAW</B> in the street where I pass to-day<BR>
Dulls the sound of the wheels and feet.<BR>
'Tis for a failing life they lay<BR>
&#160;&#160;&#160;&#160;Straw in the street.<BR>
<P>
Here, where the pulses of London beat,<BR>
Someone strives with the Presence grey;<BR>
Ah, is it victory or defeat?<BR>
<P>
The hurrying people go their way,<BR>
Pause and jostle and pass and greet;<BR>
For life, for death, are they treading, say<BR>
&#160;&#160;&#160;&#160;Straw in the street?<BR>
```

Since HTML focuses almost exclusively on a document's appearance, much of the poem's structure must be encoded using tags not designed for verse. There are a number of objections to this encoding:

- It is unclear from the encoding where the poem begins and ends
- The title is not identified as such, but only formatted to appear in larger type using the <H2> tags.
- The stanzas are not clearly encoded, but only separated using the <P> paragraph tag.
- The verses themselves are only separated using the
 line break tag.
- The inclusion of a hard space code () is the only way possible to create a blank space in most browsers, but it requires the introduction of characters foreign to the text.
- Given the limitations of HTML tagging, it would not be possible to display all lines of verse (or

poems) that contain a certain word or phrase, because the search software wouldn't know where a line or poem begins and ends.

HTML is appropriate for many uses, but not as an archival format for important literary and historical texts.

3.3.3 TEI *Guidelines*

Figure 3 shows a TEI *Guidelines* version of the poem.

Figure 3. TEI *Guidelines* Version of "Straw in the Street"

```
<DIV0 TYPE="poem">
<HEAD>Straw in the Street.</HEAD>
<LG TYPE="stanza">
<L><HI>STRAW</HI> in the street where I pass
to&hyphen;day</L>
<L>Dulls the sound of the wheels and feet.</L>
<L>&rsquo;Tis for a failing life they lay</L>
<L REND="indent1">Straw in the street.</L></LG>
<LG TYPE="stanza">
<L>Here, where the pulses of London beat,</L>
<L>Someone strives with the Presence grey;</L>
<L>Ah, is it victory or defeat?</L></LG>
<LG TYPE="stanza">
<L>The hurrying people go their way,</L>
<L>Pause and jostle and pass and greet;</L>
<L>For life, for death, are they treading, say</L>
<L REND="indent1">Straw in the street?</L></LG>
</DIV0>
```

Each major structural element of the poem is enclosed in special tags that clearly identify it. The entire poem is enclosed by a <DIV0> generic division tag. From the use of this tag, it is clear where the poem begins and ends. Each stanza is enclosed with a <LG> line group tag. It is further defined with the TYPE="stanza" attribute. Lines have been encoded with separate <L> tags. Indentations are clearly indicated using the REND="indent1" attribute. Most SGML-aware editors can format text using these attributes (e.g., to provide indentations before lines or line breaks between stanzas) without introducing entities or characters that do not actually appear in the text.

SGML can be considered an archival format, because the text can be used with a variety of software, on a variety of operating systems, for a variety of purposes. It can be used with search and display software over a network, such as the Open Text search engine. It can be displayed over the Web using a helper application from SoftQuad called Panorama Free.

Since the TEI *Guidelines* standard was especially designed for encoding literary and linguistic texts, the VWW Project decided to use it as the archival format for its texts.

Electronic texts formatted according to the TEI *Guidelines* can be easily converted to other formats, such as HTML. I use a simple script, written in the Perl programming language, to convert the TEI versions of texts in the VWW Project to HTML. Among other string substitutions, this script file changes </L> end-of-line tags in the TEI format to
 line break tags in the HTML format. John Price-Wilkin has described other on-the-fly SGML to HTML conversion strategies. [\[7\]](#)

4.0 Continuing Development of the VWW Project

As the VWW Project has evolved, its goals have changed.

Initially, the VWW Project was envisioned as complementing *The English Poetry Full-Text Database*, and, therefore, it concentrated on digitizing poetry. However, it soon became apparent that women in this period wrote in a wide variety of genres, and, in recognition of the importance of these other genres, the VWW Project was expanded to include novels, children's books, political pamphlets, religious tracts, and other forms of expression.

Many of the works included in the VWW Project are rare. For example, only half of the items in a preliminary bibliography of about 50 works were available at Indiana University, which has the vast Lilly Library of rare books and manuscripts. Many works digitized by the VWW Project have been obtained via interlibrary loan.

Even though many of these writers were very popular and critically well received in their own times, most of the writers included in the VWW Project are somewhat lesser known today than such contemporaries as Elizabeth Barrett Browning, Elizabeth Gaskell, George Eliot, and Christina Rossetti.

Given the rarity of the texts and how little known many of them are, many readers will be encountering these works for the first time, and they will want to access the entire text. Consequently, the Project's initial goal of creating a searchable database was changed to creating an archival collection of texts that are widely available on the Web.

The VWW Project is now a virtual collection, which is more inclusive than any single physical collection, available to anyone with access to the Internet. The challenge ahead is to expand the collection. I have been fortunate in that the Indiana University Libraries and my department chair have

been very supportive in allowing me to continue this work. I have also spent a great many evenings, weekends, and holidays working on various texts. Several people have asked me where I will turn my attention once the VWW Project is done, but I currently do not see any end in sight.

The breadth of writing by women in this period seems staggering, and much of it was either printed in limited press runs, not collected by libraries, or has disintegrated along with so many other works from the late nineteenth century (many of which were printed on acid-based paper). Carol Poster writes of the urgency of recovering these works, noting that failure to do so "will be the permanent silencing of the majority of popular female Victorian novelists by permitting physical disintegration of their works." [8]

As literary critics rediscover women's writing from this period, the immediate hurdle they must overcome is finding works by these writers. A recent discussion on the VICTORIA list [9] demonstrated researchers' frustrations with the limited in-print availability of works by women writers of this period. One professor said that she routinely sends inquiries to publishers as to whether they would be interested in reprinting some of these works, and the answer is invariably "no."

Electronic publishing has changed radically in the past year due to the growth of the Web. Opportunities for individuals to make editions and collections available to a wide audience have rapidly expanded. I have received several notes from instructors who have used works from the VWW Project in their classes, and one graduate student even reported the importance of certain editions for her dissertation. Considering that the VWW Project first went online in October 1995, this demonstrates its potential and that of similar electronic collections.

5.0 Conclusion

The basic relationship between publishers and libraries should change as a result of electronic publishing. Librarians, in conjunction with faculty, researchers, and students, should have a much larger role in determining not just what is collected, but what is published.

Librarians need to be well informed about electronic text issues as they affect both commercial and locally developed products. Librarians must be prepared to advise faculty and others about how to create effective electronic texts that have lasting value, and they should be willing and able to create electronic texts themselves if necessary.

The amount of effort required to prepare an electronic edition of a text is not trivial, and it should not have to be redone after a few years. Other campus departments and organizations, such as computing centers, audiovisual departments, and multimedia labs, may not share the library's concern for open standards and archival formats. Open standard formats provide the greatest possibility for the long-term archival storage of electronic texts in a rapidly changing computing environment.

By utilizing SGML and the TEI *Guidelines*, the Victorian Women Writers Project is creating a

collection of electronic texts that will remain useable in spite of technological changes in document delivery tools, such as the Web.

Notes

1. See <URL:<http://www.indiana.edu/~letrs/vwwp/>>.
2. This article was developed from a talk delivered at Yale University on March 14, 1996. I would like to thank Ann Okerson of the Yale University Libraries for inviting me. I would also like to thank Julie Bobay and Carolyn Sherayko of the Indiana University Libraries for their comments and advice in preparing this article. All remaining errors, of course, are mine only.
3. C. M. Sperberg-McQueen and Lou Burnard, eds., *Guidelines for Electronic Text Encoding and Interchange* (Chicago; Oxford: Text Encoding Initiative, 1994). Includes "A Gentle Introduction to SGML." This two-volume set is also available at <URL:<http://www.hti.umich.edu/docs/TEI/>>.
4. *The English Poetry Full-Text Database*, release 4 (Cambridge, UK: Chadwyck-Healey, 1995).
5. George Watson, ed., *The New Cambridge Bibliography of English Literature* (Cambridge: Cambridge University Press, 1969-77).
6. *Information Processing--Text and Office Systems--Standard Generalized Markup Language (SGML)*, ISO 8879 (Geneva: The International Organization for Standardization, 1986). (Also reprinted in Goldfarb--see bibliography.)
7. John Price-Wilkin, "A Gateway Between the World-Wide Web and PAT: Exploiting SGML Through the Web," *The Public-Access Computer Systems Review* 5, no. 7 (1994): 5-27.
8. Carol Poster, "Oxidization is a Feminist Issue: Acidity, Canonicity, and Popular Victorian Female Authors," *College English* 58, no.3 (March 1996): 2.
9. See <URL:<http://www.indiana.edu/~libref/victoria/96/05b/0054.html>>. For further discussion, search the archives of the VICTORIA list at <URL:<http://www.indiana.edu/~libref/victoria/vic96.html>> for messages with "noncanonical" or "fin de siecle" in the subject line.

Bibliography

Alachuler, Liora. *ABCD--SGML: A User's Guide to Structured Information*. Boston: International Thomson Computer Press, 1996.

Goldfarb, Charles F. *The SGML Handbook*. New York: Oxford University Press, 1990.

Robinson, Peter. *The Transcription of Primary Textual Sources Using SGML*. Oxford: Office for Humanities Communication, 1994.

Turner, Ronald C., Timothy A. Douglass, and Audrey J. Turner. *Readme.1st: SGML For Writers and Editors*. Upper Saddle River, NJ: Prentice Hall, 1996.

Van Herwijnen, Eric. *Practical SGML*, 2nd ed. Boston: Kluwer Academic Publishers, 1994.

For links to many different SGML sites, see the LETRS Web page at <URL:<http://www.indiana.edu/~letrs/related-links/index.html#sgml>>.

About the Author

Perry Willett, Head, Library Electronic Text Resource Service (and General Editor, the Victorian Women Writers Project), Indiana University Libraries, Bloomington, IN 47405. Internet: pwillett@indiana.edu.

About the Journal

The World-Wide Web home page for *The Public-Access Computer Systems Review* provides detailed information about the journal and access to all article files: <URL:<http://info.lib.uh.edu/pacsrev.html>>.

Copyright

This article is Copyright © 1996 by Perry Willett. All Rights Reserved.

The Public-Access Computer Systems Review is Copyright © 1996 by the University Libraries, University of Houston. All Rights Reserved.

Copying is permitted for noncommercial, educational use by academic computer centers, individual scholars, and libraries. This message must appear on all copied material. All commercial use requires permission.