Crossroads
The ACM Student Magazine

Women Writers Project

by **Benjamin Fan**

Introduction

To most people, it may seem that the history of markup languages began in 1991 with the advent of HyperText Markup Language (HTML). However, since the mid 1980's the Women Writers Project (**http://www.wwp.brown.edu/**) at Brown University has extensively used markup languages and has seen and participated in the evolution of markup language from SGML to HTML to XML.

What Is The Women Writers Project?

The Project's mission is simple: transcribe works by women writers into an electronic form and make them readily accessible to scholars. The texts are mainly works written in English and from the pre-Victorian period (before 1850 A.D.). To date, the Project has encoded over 200 texts ranging from a 1664 A.D. cookbook with recipes such as ``To make a Pigge pie'' to the speeches of Queen Elizabeth I of England.

On a simple level, the Women Writers Project may seem similar to Project Gutenberg (**http://promo.net/pg/**), a project which seeks to make the text of books available electronically, stored as pure ASCII text files. But unlike Project Gutenberg, the Women Writers Project is concerned with more than just words.

On the purpose of text encoding

> ``The point all along has been to create texts in an electronic form which can be used in a variety of ways without changing the underlying text encoding. You should be able to display online, search online (and by online we mean both in a web browser and in other forums), print out,....

> generate a Braille book, generate an audio output, etc. etc. etc., all from a single unchanging source.'' [3]

-- Carole Mah, Programmer/Analyst, Women Writers Project

What Is Text Encoding?

The words which comprise the main text might be considered the most important part of a book. However, the use of an electronic medium presents an opportunity to record much more information about a book than just its text. Meta-information describing the book and its contents can also be encoded along with text, making it more useful to scholars and researchers.

On the more physical ``book'' level, encoded information can include details such as the occurence of page breaks in the original text, changes in font, and handwritten notes in the margins. On the ``content'' level, information can be stored about the text's structure and grammatical layout. Meaning can be encoded as well, for example noting that some instances of the word ``rose'' might indicate a woman's name while in other instances of the same word the flower is meant.

Although this may seem like more information than is useful or necessary, having as much information as possible about each work allows for the distribution of multiple versions with differing levels of detail, according to the needs of each user. Accordingly, the Women Writers Project's text encoding strategy [6] is to record as much of the information about a text as possible. After looking into the options that were available at the time the Project began, it chose to encode its texts in a format called Standard Generalized Markup Language (SGML).

On Content Vs. Appearance

> ``We feel it's important to emphasize that the textual information (the words, their structure) is the real research instrument: it's what gives you powerful access to the meaning of the texts and lets you analyze their historical and cultural significance. With the current emphasis on digital images, PDF files and similar presentation-oriented forms of delivery, people too often think that by providing an image or a presentation-based version of the text you've provided everything a reader needs. We feel strongly that this is not true--the image of the page is useless for

> everything except looking at.''[2]

-- Julia Flanders, Project Manager, Women Writers Project

Standard Generalized Markup Language (SGML)

Many people are familiar with HyperText Markup Language (HTML), as the language of web pages. However, back in 1986 when the Women Writers Project began, scholarly text encoding was quite new and so was SGML, the then-recently developed standard for descriptive text encoding.

Designed to be a meta-markup language, SGML does not have a set list of tags used to markup text. Instead, it is the language used to create and define these tags. HTML is probably the best known application of SGML. As an example, with SGML you could define a tag, ``<FRUIT>'', to indicate that certain text refers to or is fruit: <FRUIT>orange</FRUIT>. You could specify that an orange is a citrus fruit, <FRUIT type = "citrus">orange</FRUIT>. You could also specify whether the text marked by the tag <FRUIT> requires an ending tag </FRUIT>, or whether the tag is a milestone element which can stand alone.

The ability to create your own markup tags is useful. However, what SGML does not do is define what a tag like <FRUIT> might mean, what it is used for, and exactly how and when it should be used. As its name suggests, SGML can be used for general markup, but it is up to each user to decide on the specific details of how to develop and use SGML tags.

This is the position that the Women Writers Project found itself in in the early 1990's. They had chosen SGML in which to encode their texts-- partially because it was the only such standard at the time and partially because they felt that its flexible nature would allow it to be adapted to the Project's needs. But, because there were no set guidelines for using SGML to mark up scholarly texts (at the time there was also no software which could display SGML documents!), before they could begin to encode any texts, the Women Writers Project had to decide exactly what tags they would use and which text elements they would encode.

To make matters worse, every other scholarly text encoding project at the time was also forced to create its own guidelines and tags for encoding text. Soon there were a diverse number of somewhat similar, but incompatible, guidelines for encoding text.

Each project knew that its encoding scheme would be incompatible with the others, making it difficult for scholars to research texts from different projects. They also realized that in the future it might be necessary to change their individual encoding schemes or to adopt new guidelines, and they would then have the task of converting or re-encoding all of the texts which had been encoded previously. It soon became clear that a standard set of guidelines was needed.

Examples of SGML's flexibility [4]

How do you encode text and a page break?

If you are more interested in the ``book'' than the ``text'' you could look at lines instead of paragraphs:

```
<page n="3">
...[etc, lines 1-23]
<line n="24">foo bar</line>
<line n="25">blah blah</line>
<line n="26">chicken</line>
<line n="27">rabbit</line>
</page>
<page n="4">
<line n="01">dog cat</line>
<line n="02">la la lah</line>
<line n="03">antlers</line>
...[etc, rest of lines for this page]
</page>
```

If you are interested in the ``text'' you could consider a page break a ``milestone element'' which can occur anywhere:

```
... [etc]
<p>
...[etc]
<lb>foo bar <!-- lb is a milestone element -->
<lb>blah blah
</p>
<p>chicken
<lb>rabbit
```

```
<pb n="4"> <!-- pb is a milestone element -->
<lb>dog cat
</p>
<p>la la lah
<lb>antlers
[etc...]</p>
<p>...</p>[etc]
```

What do you do if each page is a block element and there is a page break in the middle of a paragraph (a common occurrence)?

The wrong way-- The following example is invalid SGML because of nesting errors:

```
<page>
<p>...</p>
<p>... <!-- wrong to end enclosing element
            before nested element finishes -->
</page>
<page>
...</p>
<p></p>
...
</page>
```

A widely used solution-- break the paragraph into two sections, end one section on the first page, and begin the second section on the next page, linking the two artificially broken halves back together again using attributes:

```
<page>
<p id="p001">...</p>
<p id="p002a" next="p002b">...</p> <!-- artificial end -->
</page>
<page>
<p id="p002b" prev="p002a">...</p>
<p></p>
...
</page>
```

The Text Encoding Initiative (TEI)

In the late 1980's, many of the key people and organizations in the field of text encoding formed the international Text Encoding Initiative Project (TEI) (**http://www.uic.edu/orgs/tei/**) with the goal of developing guidelines for the encoding of scholarly texts. With input from its member organizations, and volunteers from various universities and text encoding projects (including substantial contributions by Women Writers Project staff and student workers), the TEI worked on its guidelines for a number of years, publishing the first official version of its Guidelines For Electronic Text Encoding and Interchange (TEI P3) (**http://www.uic.edu/orgs/tei/p3/**) in 1994.

The TEI P3 Guidelines, and specifically its formal Document Type Definition (DTD), specify how text structures and punctuation are to be marked up, and provide base tags for common types of texts such as prose, drama, and verse. The TEI DTD is simply an application of SGML, and the TEI Guidelines also specify how individual projects can further extend the TEI DTD to adapt it to their particular needs. The Women Writers Project's own DTD follows the TEI Guidelines and extends the TEI DTD.

With the publication of the TEI Guidelines, the Women Writers Project and other text encoding projects could finally begin to encode their texts in earnest.

Example Of Text Markup

An example of text markup using SGML and the TEI DTD. From *TEI Lite: An Introduction to Text Encoding for Interchange* [**1**]:

```
<P>At the outset of its work, the overall goals of the TEI were
defined by the closing statement of a planning conference held at
Vassar College, N.Y., in November, 1987; these <SOCALLED>Poughkeepsie
Principles</SOCALLED> were further elaborated in a series of design
documents.  The Guidelines, say these design documents,
<!-- <NOTE PLACE="foot"> <TITLE LEVEL="U">TEI ED P1:   Design
Principles for Text Encoding Guidelines</TITLE> (Chicago, Oxford:
Text Encoding initiative, 1989); like other TEI technical documents,
available from the TEI.</NOTE> --> should:
<LIST TYPE="bullets">
<ITEM>suffice to represent the textual features
     needed for research;</ITEM>
<ITEM>be simple, clear, and concrete;</ITEM>
<ITEM>be easy for researchers to use without special-purpose
```

```
software;</ITEM>
<ITEM>allow the rigorous definition and efficient processing of
texts;</ITEM>
<ITEM>provide for user-defined extensions;</ITEM>
<ITEM>conform to existing and emergent standards.</ITEM>
</LIST></P>
```

The New Markup Languages: HTML and XML

HTML is the most popular text encoding format used today. One question which comes up often is ``why doesn't the Women Writers Project just use HTML?'' Although it is widely used and would allow texts to be accessible from any web browser, HTML is merely one of many applications of SGML. Because of its simplicity, HTML has too many limitations to be used for scholarly texts.

One of HTML's limitations is the inability to encode structural information such as page breaks and metrical lines in poetry. Because scholars use information about page breaks and page signatures (how pages are physically bound together during printing) to describe and to reference specific locations in text, this is a serious limitation. Another limitation of HTML is the inability to encode semantic information such as our example of the flower ``rose'' vs. woman's name ``rose''. This makes it difficult for researchers to do intelligent searches and analyses on texts.

With these encoding disadvantages come additional disadvantages in visual presentation. Although text encoding projects consider this to be one of their least important considerations, people do have an interest in actually seeing the texts displayed. HTML's rudimentary text layout capabilities make it even less of an option for text encoding purposes.

Dissatisfaction with HTML has recently led to the development of a new standard, Extensible Markup Language (XML). Some people might think of XML as a more powerful version of HTML, but this would be wrong. XML is not just another markup language, but is in fact a metalanguage which allows you to define your own markup languages. On a simple level, XML could be considered a less-complex, web-enabled version of SGML which has retained much of SGML's functionality. According to Julia Flanders, Project Manager for the Women Writers Project, ``There are a lot of things SGML can do that XML cannot, but at least half of them are things we (and many others) [do not use].'' [2]

Just as the Women Writers Project uses SGML to encode their scholarly texts so that information about the texts can be easily used, XML allows other people to encode their own text documents and to make the data contained within more useful. XML may also replace HTML as the primary language for displaying text because it can do everything that HTML can as well as many things it cannot. As Carole Mah, Programmer/Analyst with the Women Writers Project, puts it, ``...it would be nice not to have to display using one DTD (e.g. HTML) and search using another (e.g. TEI).'' [**3**]

Even so, the Women Writers Project sees no change in its use of SGML to encode its texts. SGML is powerful and has the capability to be extended for any possible future application, and in any case it can always be easily converted to XML or HTML for web publishing purposes. Says Flanders, ``Our current plan is to continue to use SGML internally, but attempt to ship XML out the door.'' [**2**]

An Example Of A Women Writers Project Text

A paragraph and associated footnotes from *Thoughts on the Condition of Women*, 1799, by Mary (Darby) Robinson [**5**] (unformatted text only):

```
About the year of Christ, 500, Ama-
lasuenta, the daughter of Theodoric king
of the Goths, and wife of Eutharic who
was made consul by the emperor Justin,
was celebrated both for her learning and
her wisdom. Princes are said to come
and advise with her, and admire her un-
derstanding*. She took upon her the
administration of affairs, in the name of
her son, Athalaric, who was left king, at
eight years of age; and whom she in-
structed in all the polite learning before
unknown to the Goths+.

[...]

*If the great men of the present day paid more at-
tention to the genius and good sense of some British
women, they would be considerable gainers by the
```

```
experiment.

+Query. Might not the society of some living
English women, if properly appreciated, tend to the
reformation of certain gothic eccentricities; as well as,
by comparison, produce more masculine energies?
Men would be shamed out of their effeminate foibles,
when they beheld the masculine virtues dignifying the
mind of woman.
```

The same text in SGML:

```
<p rend="pre(&ldquo;)">
About <date value="0500">the year of
     <name>Chri&s;t</name>, 500</date>,
     <persname key="Amalasuen.dol">Ama&shy;
<lb>la&s;uenta</persname>, the daughter of
     <persname key="Theodoric.rbq">Theodoric</persname> king
<lb>of the <name>Goths</name>, and wife of
     <persname key="Eutharic.qab">Eutharic</persname> who
<lb>was made con&s;ul by the emperor
     <persname key="JUstin.mwl">Ju&s;tin</persname>,
<lb>was celebrated both for her learning and
<lb>her wi&s;dom.
     <emph rend="case(smallcaps)">Princes</emph> are
     &s;aid to come
<lb>and advi&s;e with her, and admire her un&shy;
<lb>der&s;tanding<anchor rend="pre(*)" id="rta012"
     corresp="rtn012">. She took upon her the
<lb>admini&s;tration of affairs, in the name of
<lb>her &s;on, <persname
     key="Athalaric.ggt">Athalaric</persname>, who was left
     king, at
<lb>eight years of age; and whom &s;he in&shy;
<lb>&s;tructed in all the polite learning
     <emph rend="slant(italic)">before
<lb>unknown to the <name>Goths</name></emph>
     <anchor id="rta013" corresp="rtn013"
rend="pre(&dagger;)">.</p>
```

```
[...]

<note id="rtn012" target="rta012" rend="place(foot)pre(*)">
<p>If the great men of the pre&s;ent day paid more at&shy;
<lb>tention to the genius and good &s;en&s;e of &s;ome Briti&s;h
<lb>women, they would be con&s;iderable gainers by the
<lb>experiment.
</p></note>

<note id="rtn013" target="rta013" rend="place(foot)pre(&dagger;)">
<p>Query. Might not the &s;ociety of <emph
     rend="slant(italic)">&s;ome</emph> living
<lb><emph rend="slant(italic)">Engli&s;h women
     </emph>, if properly appreciated, tend to the
<lb>reformation of certain <emph rend="slant(italic)">gothic
     </emph> eccentricities; as well as,
<lb>by compari&s;on, produce more ma&s;culine energies?
<fw type="catch" rend="align(right)">Men</fw>
<pb>
<lb>Men would be &s;hamed out of their <soCalled
     rend="slant(italic)">effeminate</soCalled> foibles,
<lb>when they beheld the ma&s;culine virtues dignifying the
<lb>mind of woman.
</p></note>
```

Future Work And Developments

In addition to the Women Writers Project, other women writer text encoding and related projects have been developed. These include the Orlando Project of the University of Alberta (**http://www.ualberta.ca/ORLANDO/**) and the Victorian Women Writers Project at Indiana University (**http://www.indiana.edu/~letrs/vwwp/**).

There is currently a Text Software Initiative among scholars to self-develop software specifically for working with and displaying scholarly texts. This arises from the increased number of users of encoded scholarly texts and also from the current lack of commercial products suitable for this purpose.

The Women Writers Project has also been working on Renaissance Women Online

(http://www.wwp.brown.edu/rwo/RWOoverview.html), a three-year initiative to study how the use of electronic texts in the humanities can change how teaching and research is conducted. Through this initiative, marked up texts by women writers from the Renaissance period will be available online for teaching and research purposes.

Acknowledgements

I would like to thank Carole Mah, Julia Flanders, Syd Bauman, and the rest of the Women Writers Project for their invaluable assistance with this article.

References

1

Burnard, L. and Sperberg-McQueen, C.M. *TEI Lite: An Introduction to Text Encoding for Interchange*, **http://www.uic.edu/orgs/tei/intros/teiu5.html**, June 1995.

2

Flanders, J. and Bauman, S. Private Email, April 9, 1999.

3

Mah, C. Private Email, April 8, 1999.

4

Mah, C. Private Email, April 9, 1999.

5

Robinson, M. *Thoughts on the Condition of Women*, SGML text, Brown University Women Writers Project, 1999.

6

Women Writers Project. Methodology for Transcription and Editing (draft), **http://www.wwp.brown.edu/encoding/EdPrinciples.html**, 1999.

Biography

Benjamin Fan is an undergraduate computer science student at the University at Buffalo.