2010-11-16

Digital Publishing and Preservation Using XML

Welzenbach, Rebecca; Schaffner, Paul; Hawkins, Kevin

Exercise: TEI encoding 3

This exercise is based on one used as part of the DHO Summer School 2010, held at the Royal Irish Academy and Trinity College Dublin.

So far we have created digital editions of printed material. Now we will create an edition of a handwritten manuscript of some notes by Walt Whitman.

The exercise assumes that you have a working installation of <oXygen/> XML Editor version 12 as well as the following files:

- `whitman_27.jpg`
- `whitman_28.jpg`
- `whitman_42.jpg`
- `whitman.win.xml`
- `whitman.mac.xml`
- `whitman-transcription.win.txt`
- `whitman-transcription.mac.txt`
- `tei_ms.rng`
- `whitman.css`

You might want to work with a printout of `whitman_27.jpg`, `whitman_27.jpg`, and `whitman_42.jpg` rather than viewing them on your computer.

Part A: Getting started

1. Open the <oXygen/> (with a blue icon, not the "author" mode with a red icon).
2. Go to **File → Open…** and open `whitman.win.xml` or `whitman.mac.xml` (depending on whether you're working in Windows or in Mac OS X).
3. Go to **File → Save As…** and name the file `mywhitman.xml` **in the same directory as the other files for this exercise**.

We will start with this template instead of using those built into <oXygen/> because it already has some metadata filled in and specifies the use of a schema (`tei_ms.rng`), a customization of the full TEI tag set including only tags needed for manuscript encoding.

1. Open `whitman-transcription.win.txt` or `whitman-transcription.mac.txt`.
2. Copy and paste the contents of the transcription between `<ab>` and `</ab>` in the body of the XML document.

The `ab` element is for an "anonymous block", which contains an "arbitrary, component-level unit of text." As the source document contains notes rather than complete verse, it does not make sense at this point to tag the content using either `p` (for a paragraph) or `lg` (for a group of lines, such as a stanza). So we'll put *all* of the text content of the source document in one single `ab` element.

Part B: Marking page breaks and line breaks

In our digital edition of Whitman's notes, it's important to encode page and line breaks in order to reproduce the source document as closely as possible.

Page numbers are indicated in the transcription as a number contained in square brackets. TEI indicates page numbers using the `pb` element. Note that the `pb` element is an *empty element*: instead of an opening and closing tag, it has no closing tag at all. The single tag is written with a slash after the element name, before the closing angle bracket, like this: `<pb/>`. In order to indicate the page number of the page beginning following the `pb` element, use the `n` attribute. Attributes on empty elements go before the closing slash—for example, `<pb n="3"/>`.

1. Replace each page number in brackets with a `pb` element with an `n` attribute.

Line breaks are encoded using the `lb` element, which, like `pb`, is empty and can have the `n` attribute indicating the number of the line following the tag. Line numbers are not indicated in the transcription, so we will need to add them.

2. Insert `lb` tags with an `n` attribute containing the line number at the beginning of each line. (You will have to number the lines yourself. Restart the count at "1" for each new page.)

You will see that the lines in the transcription do not follow the lines in the manuscript. You can reformat the not-yet-encoded transcription in the XML file in order make it easier to compare against the source document.

Part C: Marking scribal deletions and additions

On page 27 there is one instance of a scribal deletion. Whitman originally wrote "a locust" but then deleted "a" and used instead "the". He did this immediately: we can tell this because "the" follows immediately on the line, and "blossoms" is plural, not singular as we would expect had he finished writing "a locust blossom".

TEI provides elements for both additions and deletions. Is this a deletion plus an addition? Or is it just a deletion? It could be interpreted either way, but since "the" is clearly not an insertion made as an edit, we will treat this as a deletion standing on its own.

1. To mark this as a deletion, insert the word "a" in the appropriate place at the beginning of line 12 and tag it with the `del` element (for "deletion").

In our digital edition, we want to indicate how this deletion was made. (Was the word crossed out with a single stroke, was it scribbled out to make it barely visible, or was this part of the page torn off?)

2. To indicate the appearance of the deletion, use the `rend` attribute (for "rendition") with the value "overstrike".

On page 28, at the beginning of line 2, there's an instance of a scribal deletion plus addition. The word "spirit" has been added above the line and is clearly a replacement for the deleted text, which is

illegible. For such a substitution we have the `subst` element. In a transcription of illegible text, use the `gap` element.

3. Encode the scribal deletion plus addition on page 28, line 2, as follows:

```
<subst>
    <del rend="overstrike">
        <gap reason="illegible"/>
    </del>
    <add place="above">spirit</add>
</subst>
```

Note that the `del` element contains no text content but only a single child element, `gap`. You might also have noted when creating the `place` attribute that <oXygen/> presents a list of allowed values for this attribute. This attribute has a controlled vocabulary of "places", as opposed to some other elements (like `type`) which allow any value.

4. If you'd like, use **Format and Validate** to make your XML easier to read. Note that <oXygen/> distorts the locations of empty elements like `pb` and `lb`. If you don't like the way it looks, undo your action and format manually.

Part D: Viewing the transcription in a browser

Use <oXygen/>'s built-in XHTML transformation to view the digital edition in a browser.

1. Choose **Document → Transformation → Configure Transformation Scenario**.
2. Choose TEI P5 XHTML.

While there is some built-in handling for the display of deletions and gaps, it is not especially easy to read. As an alternative to <oXygen/>'s XSLT transformation to XHTML, let's use a CSS stylesheet to display the XML directly in a browser without transforming to XHTML. Our encoded text includes a reference to the stylesheet to use when viewing the file in a web browser.

3. Open `mywhitman.xml` in a web browser.

You'll see the file rendered quite differently, with metadata displayed in the browser and the insertions and deletions rendered in a way that mimics the appearance in the source document.

Part E: Marking more scribal additions and deletions

We'll now encode content on page 42, but we should indicate in our digital edition that we have omitted the pages between 28 and 42. If you were building a corpus of material by randomly sampling from various sources, you would omit material and indicate that the reason is "sampling", so we will use this term as well.

1. Insert `<gap reason="sampling"/>` before `<pb n="42"/>`.

Page 42 is rather more complex than pages 27 and 28, so read the entire page before you start tagging. The transcription we have provided contains only the final text, not any deletions.

You will notice that there are places where text has been deleted but nothing added in its place, as well as places where text has been added without an accompanying deletion. In these cases, you should use `add` and `del` on their own, not in combination inside of a `subst` element. The TEI allows nesting of `subst` elements within `add` and `del` for cases where text was added or deleted and then the revision was further revised.

Let's start with encoding line 5 and then go to line 3, which is a bit more difficult.

2. On line 5, add tags to mark that "how" has been deleted and replaced with "where" above the line and that "sound" has been added above the line between "the" and "man's".

Your encoding should look like this:

```
<lb n="5"/>the west,
<subst>
    <del rend="overstrike">how</del>
    <add place="above">where</add>
</subst> the <add place="above">sound</add> man's
```

3. On line 3, add tags to show that "you show me how" has been deleted in two separate acts, with one part replaced with "I imagine" and the other with "where". (As editor of this edition of the text, you might instead decide that "you show me how" has been replaced with "I imagine where" in a single act.)
4. Markup up "imagine" as deleted and replaced with "see". This requires putting a `subst` (with `add` and `del`) inside the add containing "I imagine".

Your encoding should look like this:

```
<lb n="3"/>When <subst>
    <del rend="overstrike">you show me</del>
    <add>I <subst>
            <del>imagine</del>
            <add>see</add>
        </subst>
    </add>
</subst>
<subst>
    <del rend="overstrike">how</del>
    <add place="above">where</add>
</subst>
```

5. Finish encoding page 42. If you can't read a word, insert in its place a `gap` element. Note that last two lines have been entirely deleted, so they are not in the transcription.

If you have any questions, please ask!