VWWP Encoding Setup and Validation

UPDATE TO REFLECT NEW XUBMIT INTERFACE!

Checklist

- Did you setup your working folder on your computer?
- Were you able to successfully login to the Xubmit repository to access your TEI file?
- If you are working from your personal computer, did you install and test the plugins to add subject/genre terms, preview the encoding and validate using schematron?

If you encounter problems with any of these or anything else in these instructions, please email Angela Courtney and Michelle Dalmau as soon as possible.

Below are detailed steps for setting up your encoding environment, tracking errors or problems encountered along the way with both the fulltext generation via OCR (Optical Character Recognition) and markup/encoding, and validation.

Accessing and Setting Up the Files for Encoding

1. Create a working folder on your Desktop called "vwwp"
2. Download the PDF of the text to assist in encoding
    a. Go to the VWWP Encoding Log and find the text assigned to you
    b. Right click on the link in the "PDF" column (there will be a link with the text VAB####) and click on "Save link as"
    c. Save the PDF in your "vwwp" folder
3. Access the XML "TEI shell file" in the "Victorian Women Writers Project" repository in Xubmit: http://algernon.dlib.indiana.edu:8080/xubmit/
    a. Select "Victorian Women Writers Project" and click the "go" button
    b. Login using your IU network ID and password
    c. Select file assigned to you by clicking on and highlighting the row (the row will become bright blue)
    d. Click "Download" button; Find file in your default "Downloads" directory
    e. Move the TEI Shell file you downloaded from the **Victorian Women Writers Project** Xubmit repository to your "vwwp" folder
4. Open up the Oxygen XML Editor, select File =>Open and find your XML file to begin encoding

At the end of your encoding session, you should upload your XML/TEI file so that we can maintain a backed-up version. To do this:

1. Make sure your file is valid.
2. Click on the "Submit File" link
3. Choose the file by browsing your file system
    a. **NOTE:** please do NOT rename your XML file, this will cause problems in the database
4. Create a brief log message summarizing the work you have completed thus far (e.g., "In progress; encoded through page 45")
    a. When you are **ready to submit the file for the final time**, your log message should read "Ready for Review" and the file **must be valid**
5. Set the file status to: **In progress** and when the encoding is complete and ready for editorial review, set the status to **Pending review**
6. Click on the "Submit" button

return to top

How to go about encoding the document

Anonymous Block or Comment

Remove the `<ab>` and `</ab>` tags found within the `<body>` of the shell file. These were used to auto-generate a valid TEI file ("TEI shell") but **should not appear in the final version**.

To help with periodical validation as you are encoding significant chunks of text, you can:

- Move the start <ab> tag continually down so you can work with a manageable chunk of text and still validate.
- Comment out the chunks of text you are not encoding:

```
<text>
<body>
<p>This is text I am working with.  I have TEI tags around it.</p>

<p>But below there's hundreds of pages of text I can't encode before
validating so I will need to position that open {{<ab}}> at the start of the
pages I won't be encoding.  See below.</p>

<ab> Lots of pages of text within these tags .... </ab>
</body>
</text>

<!--  OR   --->

<text>
<body>
<p>This is text I am working with.  I have TEI tags around it.</p>

<p>But below there's hundreds of pages of text I can't encode before
validating so I will need to comment out that text.</p>

<!-- Commented text is created by using the less than symbol, exclamation point and two hypens at the start of
the text AND two hypens followed by the greater than sign at the end of the text. Lots of pages of text I am
commenting out to facilitate validation. -->
</body>
</text>
```

In both cases, the start and end `<body>` and `<text>` tags need to remain in the document (e.g., they should not be enclosed by the comments punctuation else your document won't validate).

return to top

Preview Your Encoding

At any time during the encoding or submission process you can preview your markup on the Web. You can do this one of two ways, the first being the most efficient:

1. Preview in Oxygen XML Editor
   a. Right-click within the text of your XML/TEI file
   b. Select "Plugins" => "XTF Preview"
   c. Select the repository "Victorian Women Writers Project" from the drop down menu
   d. Click "OK"; a browser window will launch revealing the full text Web display
2. Preview in Xubmit
   a. Launch Xubmit: http://algernon.dlib.indiana.edu:8080/xubmit/
   b. Select "Victorian Women Writers Project" and click the "go" button
   c. Login using your IU network ID and password
   d. Select file assigned to you by clicking on and highlighting the row (the row will become bright blue)
   e. Click "Preview" button; a browser window will launch revealing the full text Web display

Warning

The full text preview does not yet account for every styling and rendering nuance. It is important that you DO NOT modify the encoding to force a more desirable Web display. Instead, inform Angela Courtney and Michelle Dalmau and we can review the style sheet outputting the HTML (Web view).

The Preview functionality is meant to give you an idea of how your text will be rendered on the Web. What you see in this Preview will not necessarily match the end result.

return to top

Validation

You will want to validate often while encoding. This will help you uncover encoding errors early on in the process as opposed to letting errors compound, making troubleshooting more challenging. You will engage in two kinds of validation: schema and schematron, both described in more detail below.

Schema validation should be frequent and on-going. Schematron validation can occur at strategic points of encoding, say in quarter or half chunks.

As part of uploading your file after an encoding session, validation must be performed in the Xubmit text repository (see above), which will check both schema **and** schematron validation. It is preferred that uploaded files are valid to both forms of validation (described in detail below), but if the file is a work in progress and time does not permit, it is okay to upload an invalid schematron file. **However, All files must be valid to the SCHEMA when uploaded.** When a **completed file** is uploaded, **it must be valid to both the schema and schematron.**

Install Schematron Validation "Plugin"

- Consult Oxygen DLP Plugins for installation instructions (if necessary)

Schema Validation

This project uses a schema customized from TEI P5, which makes sure the XML document is well-structured and valid. Your Oxygen editor automatically knows this information. The Oxygen editor will show you a green box towards the top-right of the editor if the file is valid or a red box if the file is invalid.

To validate an XML/TEI file already open in the Oxygen editor:

- Select the "validate document" icon (red checkmark)
- Or in the menus, select "Document" => "Validate" => "Validate Document"

If the file is valid, you will see:

- Green box towards the top-right of the editor
- Green box on the bottom of the editor that reads: "Document is valid"

If the file is invalid, you will see:

- Red box towards the top-right of the editor
- Red box on the bottom of the editor that reads: "Validation-failed. Errors: #"
- Error messages in a bottom pane
  - Click on each error message to position the cursor **near** the error (often the error is somewhere before the cursor, often a line or more of code above)
  - If the error message is cut off, right-click on the error message and select "Show message"

Once you fix the errors, re-validate the document. All documents must be valid to the schema at the end of an encoding session.

return to top

Schematron Validation

⚠ **Schematron being Updated**

We are working on new schematron validation so ignore this level of validation for now. – Michelle Dalmau, 3/7/2013

Schematron is an extra layer of validation that is able to check for proper formation and usages of identification schemes, type attributes (e.g., chapter, section, etc.), etc. To perform schematron validation, you will, while in the Oxygen editor:

1. Right-click within the text of the XML/TEI file that requires validation
2. Select "Plugins" => "XTF Validator"
3. Select the repository "Victorian Women Writers Project Ongoing" from the drop down menu
4. Click "OK"
5. Select a directory into which you wish to save the report
6. Name the file (e.g., vwwp_report.html, save as ".html"); make sure the name of the report ends in .html so that you will be able to view the report in a web browser
7. Click "Save"
   a. On a PC, the report will automatically open when it is saved; on a Mac you must manually open the report
8. Review the HTML report

The Schematron report uses two types of flags to indicate errors or potential errors:

- **Errors**
  - Display like this:

> Change $Encoder's Name in the name element to your first and last name.

  - Errors **must** be fixed before you upload the final version of your book. No error messages should appear in the Schematron report for a finished book.
- **Warnings**
  - Display like this:

> The front matter will generally contain a titlePage element.

  - Warnings indicate that your book does not follow a pattern that appears in *most* of the books in this collection. Read the message and check your encoding. Consult with Angela Courtney or Michelle Dalmau to determine if your book is an exception or if you need to make a change. If you decide that your book in an exception, then the Schematron report for your final book will still contain these warning messages and that is okay.

To diagnose and fix Schematron validation issues:

1. Copy and paste the "Xpath" query from each "flag" or "error" that the report generates (e.g., /TEI.2[1]/teiHeader[1]/fileDesc[1]) into the Oxygen "XPath 2.0" field located on the top-left of the Oxygen editor
2. Hit the "enter/return" key
3. Double-click on the "description" that appears on the feedback pane (below the content pane)
   a. This will take you either right to or near the vincinity in which the error/warning exists.
   b. Reference the schematron report to help you locate the exact area that needs to be fixed.

After fixing an error, please re-validate the document and verify that your fix was successful. Remember that all errors must be fixed before the final version of the book is uploaded.

return to top

OCR Errors in Full Text Generation

Full text generation is made possible by Optical Character Recognization (OCR), a process by which scanned images are converted into editable text. As you encode, you may encounter textual artifacts created by the OCR software (strange characters, omitted letters or diacritics, etc.). **Please correct the text as you encode.** Always verify the text against the page images in the PDF file.

- If the amount of correction and/or re-keying exceeds five total pages, please document the problem at: VWWP OCR Issues. Angela Courtney and Michelle Dalmau will monitor this page and get back to you as soon as possible.
- The OCR of title pages, tables of contents, and other front and back matter is often poor. These pages must be fixed and should not count toward the five total pages of correction in the above guideline.

return to top

Encoding Problems

If you encounter a problem or a question during encoding, please document the problem/question at: VWWP Encoding Problems. Angela Courtney and Michelle Dalmau will monitor this page for feedback.

return to top