Intro to Textual Encoding and XML/TEI

Institute for Editing Historical Documents - 2019

Presentation by: Serenity Sutherland

Connections to yesterday’s material

- Camp Edit hopes to help center you into an editorial community (Nick)
  - TEI is an already established community within documentary editing.
- Think about audience needs (Jennifer)
  - When making decisions about markup, we think about our audience: not only the information about the text they would like to see, but how they might like to see it.
- Digital Organization (Cathy)
  - How might TEI help (or hinder) your efforts to organize your transcription, annotation, verification process, etc…
- Engagement, Interactivity, Digital Affordances
  - TEI affords more than a static document (that scanned PDF “electronic edition”). It allows the user to interact, engage, and truly make a document dynamic.

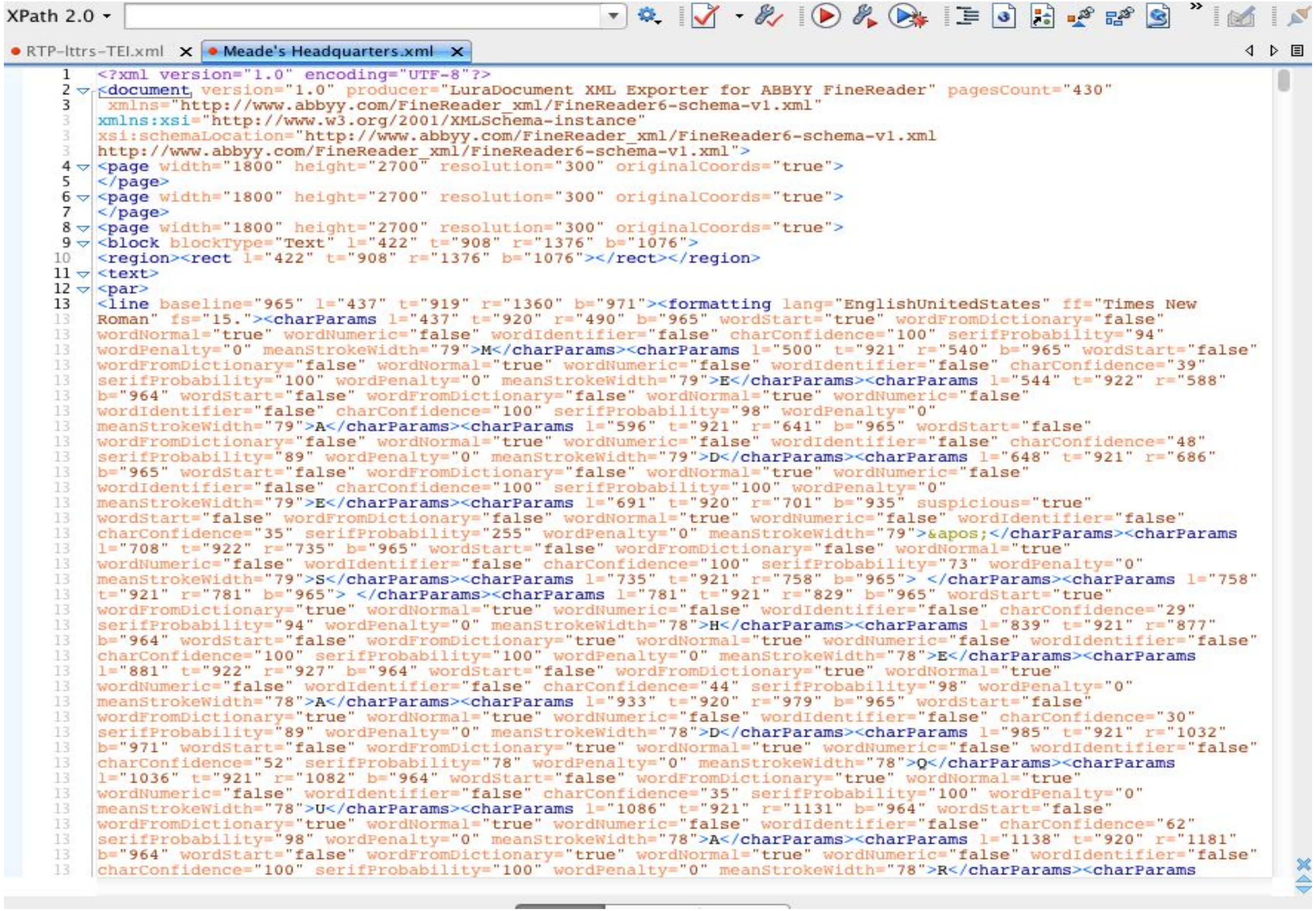
Before jumping in, let’s conceptualize TEI

TEI is XML -- XML is not TEI.

TEI is meant to be human-readable.

Descriptive markup languages (usually) formally separate information (the text in a document) from the meta-information (information *about* the text in a document).

- We use “tags” to highlight information *about* the text.
- Tags are part of markup language which is part of encoding. As soon as you make your first <p> tag, you’re encoding in a markup language! Easy peasy!



```
<?xml version="1.0" encoding="UTF-8"?>
<document version="1.0" producer="LuraDocument XML Exporter for ABBYY FineReader" pagesCount="430"
 xmlns="http://www.abbyy.com/FineReader_xml/FineReader6-schema-v1.xml"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xsi:schemaLocation="http://www.abbyy.com/FineReader_xml/FineReader6-schema-v1.xml
http://www.abbyy.com/FineReader_xml/FineReader6-schema-v1.xml">
<page width="1800" height="2700" resolution="300" originalCoords="true">
</page>
<page width="1800" height="2700" resolution="300" originalCoords="true">
</page>
<page width="1800" height="2700" resolution="300" originalCoords="true">
<block blockType="Text" l="422" t="908" r="1376" b="1076">
<region><rect l="422" t="908" r="1376" b="1076"></rect></region>
<text>
<par>
<line baseline="965" l="437" t="919" r="1360" b="971"><formatting lang="EnglishUnitedStates" ff="Times New
Roman" fs="15."><charParams l="437" t="920" r="490" b="965" wordStart="true" wordFromDictionary="false"
wordNormal="true" wordNumeric="false" wordIdentifier="false" charConfidence="100" serifProbability="94"
wordPenalty="0" meanStrokeWidth="79">M</charParams><charParams l="500" t="921" r="540" b="965" wordStart="false"
wordFromDictionary="false" wordNormal="true" wordNumeric="false" wordIdentifier="false" charConfidence="39"
serifProbability="100" wordPenalty="0" meanStrokeWidth="79">E</charParams><charParams l="544" t="922" r="588"
b="964" wordStart="false" wordFromDictionary="false" wordNormal="true" wordNumeric="false"
wordIdentifier="false" charConfidence="100" serifProbability="98" wordPenalty="0"
meanStrokeWidth="79">A</charParams><charParams l="596" t="921" r="641" b="965" wordStart="false"
wordFromDictionary="false" wordNormal="true" wordNumeric="false" wordIdentifier="false" charConfidence="48"
serifProbability="89" wordPenalty="0" meanStrokeWidth="79">D</charParams><charParams l="648" t="921" r="686"
b="965" wordStart="false" wordFromDictionary="false" wordNormal="true" wordNumeric="false"
wordIdentifier="false" charConfidence="100" serifProbability="100" wordPenalty="0"
meanStrokeWidth="79">E</charParams><charParams l="691" t="920" r="701" b="935" suspicious="true"
wordStart="false" wordFromDictionary="false" wordNormal="true" wordNumeric="false" wordIdentifier="false"
charConfidence="35" serifProbability="255" wordPenalty="0" meanStrokeWidth="79">&apos;</charParams><charParams
l="708" t="922" r="735" b="965" wordStart="false" wordFromDictionary="false" wordNormal="true"
wordNumeric="false" wordIdentifier="false" charConfidence="100" serifProbability="73" wordPenalty="0"
meanStrokeWidth="79">S</charParams><charParams l="735" t="921" r="758" b="965"> </charParams><charParams l="758"
t="921" r="781" b="965"> </charParams><charParams l="781" t="921" r="829" b="965" wordStart="true"
wordFromDictionary="true" wordNormal="true" wordNumeric="false" wordIdentifier="false" charConfidence="29"
serifProbability="94" wordPenalty="0" meanStrokeWidth="78">H</charParams><charParams l="839" t="921" r="877"
b="964" wordStart="false" wordFromDictionary="true" wordNormal="true" wordNumeric="false" wordIdentifier="false"
charConfidence="100" serifProbability="100" wordPenalty="0" meanStrokeWidth="78">E</charParams><charParams
l="881" t="922" r="927" b="964" wordStart="false" wordFromDictionary="true" wordNormal="true"
wordNumeric="false" wordIdentifier="false" charConfidence="44" serifProbability="98" wordPenalty="0"
meanStrokeWidth="78">A</charParams><charParams l="933" t="920" r="979" b="965" wordStart="false"
wordFromDictionary="true" wordNormal="true" wordNumeric="false" wordIdentifier="false" charConfidence="30"
serifProbability="89" wordPenalty="0" meanStrokeWidth="78">D</charParams><charParams l="985" t="921" r="1032"
b="971" wordStart="false" wordFromDictionary="true" wordNormal="true" wordNumeric="false" wordIdentifier="false"
charConfidence="52" serifProbability="78" wordPenalty="0" meanStrokeWidth="78">Q</charParams><charParams
l="1036" t="921" r="1082" b="964" wordStart="false" wordFromDictionary="true" wordNormal="true"
wordNumeric="false" wordIdentifier="false" charConfidence="35" serifProbability="100" wordPenalty="0"
meanStrokeWidth="78">U</charParams><charParams l="1086" t="921" r="1131" b="964" wordStart="false"
wordFromDictionary="true" wordNormal="true" wordNumeric="false" wordIdentifier="false" charConfidence="62"
serifProbability="98" wordPenalty="0" meanStrokeWidth="78">A</charParams><charParams l="1138" t="920" r="1181"
b="964" wordStart="false" wordFromDictionary="true" wordNormal="true" wordNumeric="false" wordIdentifier="false"
charConfidence="100" serifProbability="100" wordPenalty="0" meanStrokeWidth="78">R</charParams><charParams
```

The best TEI/XML analogy...

XML is a syntax.

Subject Verb Object.

TEI is a vocabulary that uses that syntax.

I see the dog.

What is XML?

e**X**tensible: XML can be expanded and adapted to include any number of descriptive frameworks and terminologies, i.e. it’s an open-ended mode of structured description

Markup: XML is a metadata format that allows for information about a document or its contents to be added within a single file structure. This means that you can add information about contents without having to alter those contents themselves.

Language: XML is a syntax with rules, just like other languages have rules.

XML Syntax and Format

Open and Close Tags: XML tags start with an open tag in brackets <tag> and end in a close tag </tag>. If a tag is “empty” and contains no child elements, it may be rendered as an “empty tag” in format <tag/>.

Attributes: Attributes are information contained within the tags themselves. They are formatted like this : <tag attribute=”value”/>.

Root: you must have a “root” element. This means that every XML file will have one “root” into which everything else will be contained. From another angle, this means that every XML document has a clear beginning and end point, defined by the root element. All other elements and content are “children” of the root.

Hierarchical: No element can be a child or a parent of itself, they must always nest like Russian dolls or tree branches.

Yes: <book><chapter1><page>...</page></chapter1></book>

No: <book><chapter1><page>...<chapter2>...<page/></chapter1></chapter2></book>

Plain Text: an XML file contains no special characters or styling, these must all be added in the form of metadata using elements and attributes.

XML is meant to be human-readable...

```
<note>
  <to>Tove</to>
  <from>Jani</from>
  <heading>Reminder</heading>
  <body>Don't forget me this
weekend!</body>
</note>
```

Above example from https://www.w3schools.com

XML vocabularies include

DocBook
EPUB
.docx (WordML)
EAD
TEI

Each guideline specifies what terms are recognized by the tagset and what combinations and arrangements are allowed.

Let’s break down the TEI

element

```
<title>Letter from Frances Miller Seward to William
Henry Seward, 1 January 1837</title>
```

From the Seward Project

```
<title>Notebooks and Unpublished Prose
Manuscripts</title>
```

From the Walt Whitman Archive

To see a larger list of common elements by group type:
https://wwp.northeastern.edu/outreach/seminars/_current/handouts/elementList.xhtml

Within elements you can have attributes

Vague example:

```
<element attribute=”value”>example</element>
```

More specific examples:

```
<hi rend="italic">Walt Whitman Archive</hi>

<biblScope type="page">169-170</biblScope>
```

Sometimes, elements may require more than one attribute.

```
<persName ref="psn:CHEm_233" cert="medium">Mrs Cheadell</persName>
```

Metadata within your TEI

All kinds of juicy details about the text in your document can show up in the TEI. This information doesn’t display in the public facing portion of your document, but helps to track the documents as they move through your workflow.

```
<funder>The National Endowment for the Humanities</funder>

<orgName xml:id="duk">
  Trent Collection of Whitmaniana, David M. Rubenstein Rare Book & Manuscript Library, Duke University
</orgName>

<msIdentifier>
  <institution>University of Rochester</institution>
      <repository>Rare Books and Special Collections</repository>
</msIdentifier>
```

To think about...TEI Customization

The process of altering the TEI schema and documentation to match your needs.

Changes may include:

- choosing which parts to use or omit
- changing the names of elements or attributes
- restricting the values of attributes
- adding new elements
- adding new attributes

Slide made with inspiration from: https://wwp.northeastern.edu/outreach/resources.html

Actually doing the TEI...

Decide on a text editor...

- Highly suggest oXygen

  Download

  https://www.oxygenxml.com/xml_editor/download_oxygenxml_editor.html

Why oXygen?

- Has parsing capabilities and will help you write error-free XML documents. It validates your XML against a schema (more on that later), and visually communicates this via a red (error) or green (error free) box.

Processing your XML/TEI

You might wonder how oXygen knows so much about your document. The application is reading the information at the top of your XML file, in the declarations and (sometimes) other references within your file.

You can use declarations to associate your file with various schemas, namespaces, and other useful resources or scripts, such as XSLT files to transform your XML document into various and/or multiple outputs.

The TEI and you

Overall, think of your journey with TEI as a set of phases. Today, phase 1, practice tagging and going green (box, that is).

Tomorrow, phase 2, try using TEI’s Roma customization.

The day after? Read through the entire TEI Guidelines in one sitting… just kidding. Maybe never do all of it in one sitting. Not light reading!

TEI can help digital editions to...

- Choose a core set of TEI tags
- Document that set and the decisions you made about how to implement markup
- Customize with Roma/schema for constrained validation
- Establish TEI boilerplate

Resources

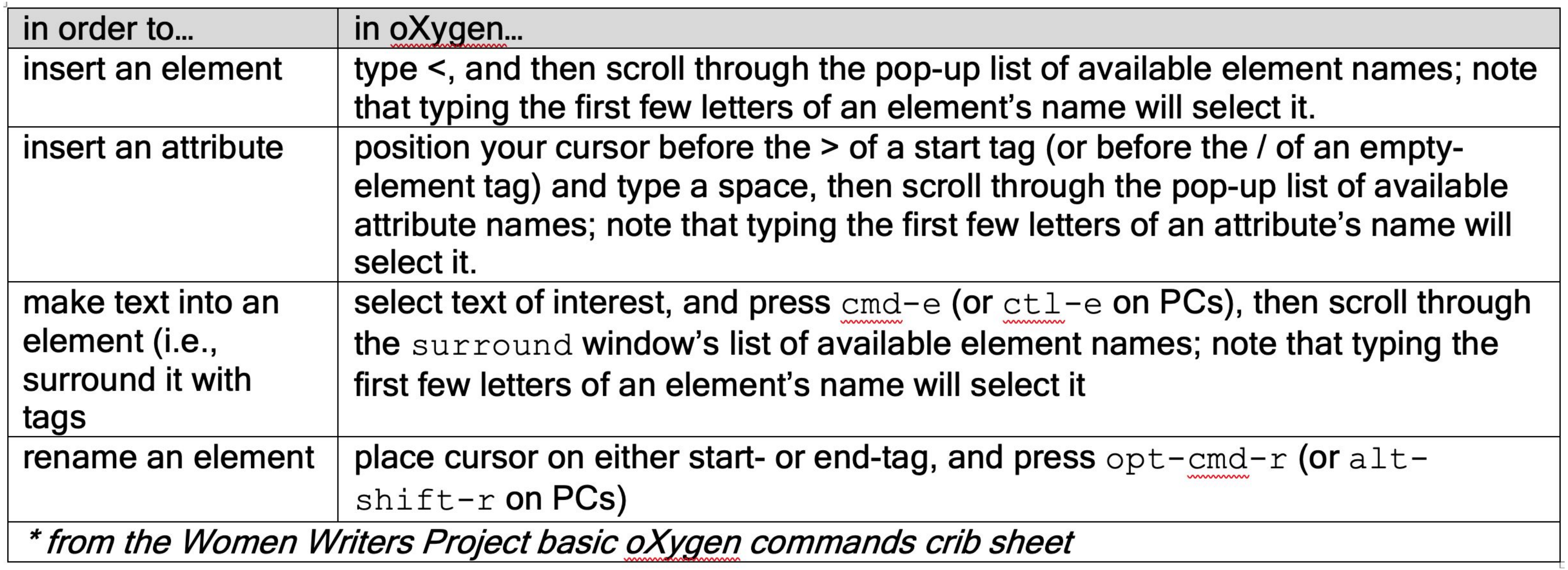
My favorite, and a handy resource for quick tutorials is “TEI by Example:” http://teibyexample.org/TBE.htm

David Birnbaum, “What is XML and Why Should Humanists Care? An Even Gentler Introduction to XML,” obduron.org, 28 August 2015), http://dh.obdurodon.org/what-is-xml.xhtml.

“A Gentle Introduction to XML” TEI P5 Guidelines Version 3.5.0, https://tei-c.org/release/doc/tei-p5-doc/en/html/SG.html.

Northeastern’s Women Writers Project is a good example of both “how to” and “why?”: https://wwp.northeastern.edu/

Quick Reference



in order to…	in oXygen…
insert an element	type <, and then scroll through the pop-up list of available element names; note that typing the first few letters of an element's name will select it.
insert an attribute	position your cursor before the > of a start tag (or before the / of an empty-element tag) and type a space, then scroll through the pop-up list of available attribute names; note that typing the first few letters of an attribute's name will select it.
make text into an element (i.e., surround it with tags	select text of interest, and press `cmd-e` (or `ctl-e` on PCs), then scroll through the `surround` window's list of available element names; note that typing the first few letters of an element's name will select it
rename an element	place cursor on either start- or end-tag, and press `opt-cmd-r` (or `alt-shift-r` on PCs)
** from the Women Writers Project basic oXygen commands crib sheet*	

Some TEI Tags to Practice With

Using TEI-All Schema, practice encoding your document's basic layout structure, identifying people and places, and adding some annotations.

Layout/Structural Tags:
<div> and <div type="chapter"> etc.
<p> and/or <l>
<pb/>

Identifying People, Places
<persName key="">
<placeName key="">

Annotations
<note>

Advanced Question:
How would you tag the location of topics/themes in your document(s), including places where they are not mentioned explicitly? For example, your collection may contain discussion of slavery, without ever using the word slavery.

Not ready to start encoding in TEI? Start thinking about your document's structure with this document analysis worksheet: https://wwp.northeastern.edu/outreach/seminars/_current/handouts/document_analysis.tei

Supported by:

Association for Documentary Editing

&

National Historical Publications & Records Commission

THE ASSOCIATION FOR DOCUMENTARY EDITING · VERITAS EX DOCUMENTIS · ADE

NATIONAL ARCHIVES

NATIONAL HISTORICAL PUBLICATIONS & RECORDS COMMISSION