Introduction to Text Encoding for Research and Scholarship

Sarah Stanley
Assistant Digital Scholarship Coordinator
scstanley@fsu.edu

FLORIDA STATE UNIVERSITY
LIBRARIES

Before we get started

- Download <oXygen/> at https://www.oxygenxml.com/xml_editor/download_oxygenxml_editor.html
- Get a trial license: https://www.oxygenxml.com/xml_editor/register.html#get_trial
- Pick a text from Project Gutenberg (or if you want to transcribe something, pick a text from Google Books, HathiTrust, or some other resource of digitized texts)

Boring Preface About XML

- TEI uses XML (for now) as a standard for describing texts in a structured manner
- XML is
  - A set of specifications for encoding and marking up documents
  - “Ordered Hierarchy of Content Objects” (OHCO) model
  - Thinks of documents as containing nested, rather than linear structures. Text is not a string, but rather a set of nested objects (words in paragraphs, paragraphs in divisions, etc.)
  - Recommended Reading:
    - Text: A Massively Addressable Object, Michael Witmore

Is XML the same as TEI?

- XML is a set of standards for markup
  - It dictates the structure of documents, but not the vocabulary
- TEI uses XML
  - TEI, as a scholarly standard could adopt a method of description that *isn't* XML
  - TEI is the descriptive vocabulary practices
- There are XML languages that aren't TEI!
  - MathML, MML (music), HTML, MEI
  - Individual languages define vocabulary

Rules of XML (aka Well-formedness)

XML only has 3 rules (these apply to ***all*** XML languages!)*

- Everything needs to be properly delimited
- No overlap!
- One root element

*not the same as the rules of *specific* markup languages!

Boring Preface About XML

Let's talk about XML.

- nested structure
- comprised of "elements"
- elements may be further described by "attributes"
- attributes have "values"

All characters properly delimited

- <elementName>content</elementName>
- <emptyElement/>
- <elementName attName="attValue">content</elementName>
- & < > (special characters must be properly delimited!)

Test your knowledge!

1. <name type="person>Abe Lincoln</name>
2. <name type="place">Tallahassee<name>
3. <name="organization">Florida State University</name>
4. <name type="person">Robert M. Strozier</name>
5. <name type="person" type="place">Washington</name>

No overlap!

This is one of the harder things to wrap your head around conceptually, but it is *super important* if you are going to to XPath searches.

All elements need to be nested *inside* each other like matryoshka dolls.

Example of Overlap

Our encoder said,
<said sameAs="#s2">The beginning of a bit of speech starts in one paragraph.</said>

 <said xml:id="s2">And yet, it continues on to the next!</said> She wondered how she could possibly represent it.

Describing Text Structures

- Mark divisions of the text with
  - Specify what type of divisions they are
  - type="chapter", type="preface"
- Mark paragraphs or other medium-level chunks
  - <p> for paragraph
  - <ab> for “anonymous block” (commonly used for bible translations where textual chunking ≠ paragraph
- Information about poetic stanzas and lines
  - <lg> = line group (stanza)
  - <l> - poetic line

Info about physical features

- Breaks in the text
  - <pb/>, <lb/>, <cb/>
- Damage to physical text
  - <damage>
  - <gap/>, <unclear> with @reason
- Paratextual details
  - Figures, ornaments, “formework”
- <surface> for describing physical pages

Phrase-level information

- Names: <persName>, <placeName>, <orgName>
- Quotations and speech
- Rhetorical devices
  - <emph> for emphasis
  - <soCalled> for ironic distance
  - <distinct> and <foreign> for foreign languages or use of dialect/accents
- ????

Regularizing Messy Text Data

<date when=”2016-02-10”> 10 February 2016 C.E. </date>

<date when=”2016-02-10”> Tenth day of Feb., 2016 A.D </date>

<distance unit="mile" measure="1"> 1 mile </distance>

<distance unit="mile" measure="1"> 8 furlongs </distance>

<placeName ref="#newyork"> New York City </placeName>

<rs ref="#newyork"> “The Big Apple” </rs>

<persName ref="#georgeeliot"> Mary Ann Evans </persName>

<persName type="pen" ref="#georgeeliot"> George Eliot </persName>

Now that you know the basics, let's get started!

Some things to try

1. Wrap all textual divisions (i.e. chapters, letters, poems, prefaces) in
   a. Add the type attribute (with a descriptive value) to
2. Wrap all poetic stanzas in <lg> and all poetic lines in <l>
3. Wrap all paragraphs in <p>
4. Wrap headings/titles in <head>
5. Wrap printed page numbers in <fw> (formework)

FLORIDA STATE UNIVERSITY

LIBRARIES

LIB.FSU.EDU