WHO WANTS TO BE A STAR ARCHIVIST?

You want to make available a 50 page manuscript journal by Sam Houston online. How should you present it?

a) Page images (JPEGs)
b) Image with OCR behind it (like Google Books)
c) PDFs
d) As a web page (HTML)
e) Transcription encoded in TEI (Text Encoding Initiative)
f) It depends.
Five questions to ask in deciding how to digitize texts

- What are the goals of the project?
- What best serves our end-users?
- What will last, so that our initial investment has long-term value?
- What does our community say is the best way to approach such a project? (Best practices)
- What is feasible?
Outline of Talk

• What is TEI?
• How does XML enables flexibility in representing & manipulating data?
• What is the relationship between TEI & EAD?
• What are some cool ways that TEI can be used?
• What does it take to create a TEI document?
What is TEI?

- TEI= Text Encoding Initiative
- Guidelines for representing texts in electronic form. Focused on humanities & social sciences.
- Separates content from presentation
- Includes guidelines for marking up:
  - Novels
  - Plays
  - Poems
  - Letters & manuscripts
  - Dictionaries
  - Linguistic corpora
What is TEI’s history?

- Emerged from a 1987 meeting of representatives from text archives, research projects & scholarly societies
- Guidelines have gone through 5 versions (proposals, in their language) between 1990 and 2007
- Beginning with TEI P.4, TEI is expressed in XML
Why do we need TEI?

- Enable analysis of texts
- Make explicit features of a text so that they can be processed by computer applications
- Support range of output formats (HTML, PDF, Braille reader, etc)
- Long-term preservation
What does TEI look like?

<div n="1" type="letter">
  <head>Letter from Col. W.R. Boggs to Thomas O. Moore, July 29, 1862</head>
  <opener>
    <pb n="1" facs="aa00151_0001.gif"/>
    <dateline>
      <placeName>
        <settlement>Milledgeville</settlement>,
        <region><choice>
          <abbr>Ga</abbr>
          <expan>Georgia</expan>
        </choice></region>
      </placeName>
      <date when="1862-07-29">July 29 1862</date>
    </dateline>
    <lb/>
  </opener>
TEI & XML

- XML, or Extensible Markup Language, is a meta-language, offering a syntax used to define and create markup languages. XML provides rules for describing the structure of a document.
- TEI is one of 100s of XML “applications”
- XML is…
  - A common, standards based approach for structuring and storing information
  - A family of related technologies for processing and publishing information
- In essence: “XML is a cross-platform, software and hardware independent tool for transmitting information.” (W3Schools, Introduction to XML)
XML: Why all the fuss?

- "XML offers the power, precision, control, and flexibility that should appeal to librarians at the gut level. It provides a unifying foundation for future development due to its extensibility (suitable for evolving standards), sophisticated hyperlinking, modularity (permitting reuse of information in multiple contexts with different requirements), and relative simplicity." (Dick R. Miller, "XML: Libraries' Strategic Opportunity")
Structure vs. Presentation:
XML vs. HTML

Xavier Xylophone
*Exuberant XML*
Xpert Boox
<?xml version="1.0" encoding="UTF-8" standalone="yes"?>
<!DOCTYPE html PUBLIC "-//W3C//DTD XHTML 1.0 Transitional//EN" "http://www.w3.org/TR/xhtml1/DTD/xhtml1-transitional.dtd">
<html xmlns="http://www.w3.org/1999/xhtml">
<head>
</head>
<body>

<h1>HTML Version</h1>


</body>
</html>
<?xml version="1.0" encoding="UTF-8"?>
<book>
  <author type="primary">Xavier Xylophone</author>
  <title>Exuberant XML</title>
  <publisher>Xpert Boox</publisher>
</book>

[Note: This is a made-up schema, not TEI]
XML vs. HTML

The problem with "What You See Is What You Get" is that what you see is all you've got. (Brian Kernighan)

<table>
<thead>
<tr>
<th>HTML</th>
<th>XML</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fixed set of elements</td>
<td>Enable users to define the elements they need</td>
</tr>
<tr>
<td>Focused on presentation</td>
<td>Focused on structure and syntax and describes what data is; facilitates information retrieval, analysis, and exchange</td>
</tr>
<tr>
<td>Meant only to be used for presentation of documents in Web browsers</td>
<td>Can be used as a storage format for word processors, data interchange format, &amp; preservation format.</td>
</tr>
</tbody>
</table>
Advantages of XML

- Reusability
- Sustainability
- Information Exchange/Interoperability
- Extensibility
- Semantic & Structural Richness
Advantages: Reusability

- "Build once, use many"
- Separates presentation from content
- Multiple outputs possible, more versatile digital libraries, e.g.:
  - Web
  - e-book
  - Pdf
  - Braille reader
  - Database of personal names
  - Index
Advantages: Sustainability

- Non-proprietary, open standard
- Well-supported
- Human and machine readable
- Unicode character sets (language-independent)
- Recommended by Digital Preservation studies:

  “XML is of the greatest importance for **digital preservation**, not just because of this widespread uptake, but also because it protects the Achilles’ heel of digital documents: the dependence on obsolete operating systems and application software. It does this by being **platform- and software-independent**. The **separation of content, structure and appearance** plays an important role here.”

Advantages: Information Exchange/Interoperability

- Hardware and software independent
- Hence, fosters exchange of data between different systems and applications …
- … which, in turn, tends to foster user access to information
- XML can be transformed programmatically according to the user's needs. For instance, businesses can receive data from another company's system and translate it for their own.
Advantages: Extensibility

- XML provides common rules for encoding and markup syntax
- User communities are free to extend XML by developing XML implementations of their metadata & document-tagging standards
- Through namespaces, an XML document can encode data in different schemas within the same document.
Advantages: Semantic & Structural Richness

- XML enables one to make explicit the structural features of a document
  - Chapters, paragraphs, archival containers
- XML enables one to make explicit the semantic features of a document
  - Personal names, place names, dates
- The XML markup can then be used by applications to search, retrieve, and display features of interest in useful ways
Both are markup standards based on SGML (which evolved into XML)
Daniel Pitti examined TEI while developing the EAD DTD (Document Type Definition) but decided it wouldn’t work for finding aids

Still, EAD was “made as consistent with TEI as possible: the basic TEI header structure was incorporated into EAD, and element names and attributes conflict as little as possible.” (http://www.loc.gov/ead/ag/agagcntx.html)
Integrated Guide to Whitman's Poetry Manuscripts:
TEI & EAD @ the Walt Whitman Archive

- Goal: “develop a virtual collection of Whitman poetry manuscripts”
- Lists “all identified poetry manuscripts located in archival repositories throughout the United States and United Kingdom.” Includes 29 repositories.
- Listing generated from EAD finding aids
- Provides access to images and/or transcription of original item in TEI, if available
- Links to EAD finding aid
- SAA awarded the Whitman Archive the C.F.W. Coker Award for their work on this project
What can you do with TEI?

TEI offers a great deal of power and flexibility in representing textual information. For instance, with TEI you can:

- Enable specialized searches
- Present different versions of document
- Compare documents
- Provide context for terms
- Make possible sophisticated textual analysis
Enable specialized searches

- Example: UVA Electronic Text Ctr.
- Limit search to particular fields (elements) such as title, date, subject, genre, publisher, etc.
Dear cousin

I have no doubt
But what you think by my long silence
that I have forgotten you tho Ill have
you to know thatsuchis not the case I
have bin waiting to find out where we had
to be stationed we are at winchester, now
I dont know how long weel have
to stay [unclear: here] , I am in hopes that
we will stay here for some time,
we have elegant water and a
plenty of it, and a plenty of good
pervision so far, and a fine chance of
beutiful young Ladies, and the kind
est that I ever saw in my life,
and the most beautiful Country
that I ever saw thay have fine
Crops over here, and not [unclear: mutch]
likely hood of a fight the yankeys

Original Version

Dear cousin,

I have no doubt
but what you think by my long silence
that I have forgotten you, though I'll have
you to know that such is not the case. I
have been waiting to find out where we had
to be stationed. We are at Winchester now.
I don't know how long we will have
to stay here, I am in hopes that
we will stay here for some time.
We have elegant water and a
plenty of it, and a plenty of good
pervision so far, and a fine chance of
beautiful young Ladies, and the kindest
that I ever saw in my life,
and the most beautiful Country
that I ever saw. They have fine
Crops over here, and not [unclear: much]
likelihood of a fight. The yankees

Modernized Version

**Compare Different Versions of a Literary Text: Dickinson’s “Faith is a Fine Invention”**

<table>
<thead>
<tr>
<th>Version</th>
<th>A660</th>
<th>X</th>
<th>Version</th>
<th>P1891</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>&quot;Faith&quot; is a fine invention</td>
<td></td>
<td>1</td>
<td>FAITH is a fine invention</td>
</tr>
<tr>
<td>2</td>
<td>When Gentlemen can see</td>
<td></td>
<td>2</td>
<td>For gentlemen who see</td>
</tr>
<tr>
<td>3</td>
<td>But Microscopes are prudent</td>
<td></td>
<td>3</td>
<td>But microscopes are prudent</td>
</tr>
<tr>
<td>4</td>
<td>In an Emergency</td>
<td></td>
<td>4</td>
<td>In an emergency</td>
</tr>
</tbody>
</table>

**The Versioning Machine**: “a framework and an interface for displaying multiple versions of text encoded according to the TEI”

- Dynamically compare transcriptions & page images

**Verse embedded in letter to Samuel Bowles.**

**Published as poem XXX in the second volume of Todd and Higginson's Poems of Emily Dickinson.**
Provide Context for Keywords in the Text

The dreams of childhood realized, the hopes of early manhood fully accomplished, I returned home with stories of travel for ears which, alas the day! were closed to my voice by the solemn seal of death.

Whether, that I have seen the sunrise flush the brow of Remeses at Abou Simbal, and touch with passionate, yet gentle and trembling caresses—as a lover would touch the lips of his maiden love, dead in her grave, the cold lips of Memnon at old Thebes; through the stately halls of Karnak, and look up the stream of time from the summit of the mastaba, I have knelt at the Sepulchre, and felt the coolness of my forehead in Gethsemane—whether all
Analyze Texts with TAPOR

TAPoR Tool Broker

TAPoRware Find Collocates (HTML)

Parameters
- HTML tags: body
- Word/Pattern: bachelor
- Context: Words
- Context length: 10
- Sort criteria: Co-occurring words by frequency
- Display as: HTML

Submit □ Show results in new window?

Summary: There are 168 unique words other than those in the stop list, there are 315 words other than those in the stop words. There are 607 words in total including the stop words.

<table>
<thead>
<tr>
<th>Words</th>
<th>Counts</th>
</tr>
</thead>
<tbody>
<tr>
<td>Page</td>
<td>14</td>
</tr>
<tr>
<td>Heart</td>
<td>8</td>
</tr>
<tr>
<td>Image</td>
<td>6</td>
</tr>
<tr>
<td>Reveries</td>
<td>5</td>
</tr>
<tr>
<td>Book</td>
<td>4</td>
</tr>
<tr>
<td>Man</td>
<td>2</td>
</tr>
<tr>
<td>Surely</td>
<td>2</td>
</tr>
<tr>
<td>Great</td>
<td>2</td>
</tr>
<tr>
<td>---and</td>
<td>2</td>
</tr>
<tr>
<td>Pseud</td>
<td>2</td>
</tr>
<tr>
<td>Friends</td>
<td>2</td>
</tr>
<tr>
<td>Dignity</td>
<td>2</td>
</tr>
<tr>
<td>Light</td>
<td>2</td>
</tr>
<tr>
<td>Years</td>
<td>2</td>
</tr>
<tr>
<td>Dusty</td>
<td>2</td>
</tr>
</tbody>
</table>

TAPoR Tool Broker

TAPoRware Visual Collocator

Parameters
- Element: body
- Pattern: bachelor
- Exclude Glasgow Stop Words
- Context: Words
- Context length: 5
- Sorting: Co-occurring words by frequency

Submit □ Show results in new window?
What do you need to do to make TEI texts available online?

- Convert scanned image to textual format
  - Through OCR (would need to clean up resulting OCR; OCR doesn’t really work with handwritten docs)
  - Through (double)-keying

- Add TEI markup
  - You can use an XML editor such as Oxygen, or programmatically add some markup (e.g. by saving Open Office file as TEI)
  - Different approaches to markup:
    - Light markup: chapters, paragraphs, etc
    - Heavier markup: names, critical apparatus, linguistic features, etc.
How to serve TEI online

- Use XSLT stylesheet(s) to convert XML file to variety of formats, including HTML
- Make available online by:
  - Storing HTML on the server
  - Transforming on the fly using XML publishing technologies such as Cocoon, AxKit, etc. or XML database such as eXist, Xindice, or
  - Trusting the browser to make the transformation by applying CSS (most newer browsers support XML)
WHO WANTS TO BE A STAR ARCHIVIST?

You want to make available online a 50 page manuscript journal by Sam Houston. How should you present it?

f) It depends.

- TEI offers a number of advantages, such as the ability to re-purpose, search, analyze, preserve, and manipulate TEI files
- However, TEI also requires a significant investment of time and resources to implement.
Resources

- Lisa’s bookmarks on TEI: http://www.diigo.com/user/lspiro/tei
- The TEI Consortium: http://www.tei-c.org/index.xml
- Brown Women Writer’s Project (TEI training materials, workshops, and more): http://www.wwp.brown.edu/encoding/
Bonus slides
TEI Design Goals

- provide a standard format for data interchange
- provide guidance for the encoding of texts in this format
- support the encoding of all kinds of features of all kinds of texts studied by researchers
- be application independent
Compare different editions of a text

- Critical edition presents variants in 5 different versions of *Reveries of a Bachelor* (1849-1884)
- Can move dynamically between different versions of text by clicking on asterisks
- Created by encoding variants in TEI
Analyze Texts with Token X

TokenX: a text visualization, analysis, and play tool

Customized for the Willa Cather Archive.
About TokenX on the Cather Archive.
See a List of Texts Used in Text Analysis Project

choose:

Search word frequency*: [ ] execute

*Enter a list of words separated by a vertical bar (beef|bicycle|happin*)

-or-

Sample file:** [ ] execute

**Choose from all of Cather’s works of fiction not currently protected by copyright

Welcome to TokenX! Please select a file to analyze.
Re-focus Attention Using Token X

TokenX: a text visualization, analysis, and play tool

Customized for the Willa Cather Archive.
About TokenX on the Cather Archive.
See a List of Texts Used in Text Analysis Project
File: /data/cather/cat.0006.2_samples.xml

Choose word(s) to replace with image(s):
- train --- > Bicycle
- town --- > Flag
- she --- > Willa Cather

Replace words with images

THE SCULPTOR’S FUNERAL

A group of the townspeople stood on the station siding of a little Kansas, awaiting the coming of the night, which was already twenty minutes overdue. The snow had fallen thick over everything; in the pale starlight the line of bluffs across the wide, white meadows south of the made soft, smoke-coloured curves against the clear sky. The men on the siding stood first on one foot and then on the other, their hands thrust deep into their trousers pockets, their overcoats open, their shoulders screwed up with the cold; and they glanced from time to time toward the southeast, where the railroad track wound along the river shore. They conversed in low tones and moved about restlessly, seeming uncertain as to what was expected of them. There was but one of the company who looked as though he knew exactly why he was there; and he kept conspicuously apart, walking to the far end of the platform, returning to the station door, then pacing up the track again, his chin sunk in the high collar of his overcoat, his burly shoulders drooping forward, his gait heavy and dogged. Presently he was approached by a tall, spare, grizzled man clad in a faded Grand Army suit, who shuffled out from the group and advanced with a certain deference, craning his neck forward until his back made the angle of a jack-knife three-quarters open.

"I reckon he's a-goin' to be pretty late agin to-night, Jim," he remarked in a squeaky falsetto. "S'pose it's the snow?"