Introducing TEI

What is Text-Encoding Initiative (TEI), anyway?
TEI is a set of guidelines for encoding texts in a fashion that enables a variety of uses in digital scholarship and humanities computing. TEI is also the name for the consortium (www.tei-c.org) which maintains the guidelines, and the broader community of practice in the humanities and social sciences which use them.

What XML?
XML, or the Extensible Markup Language is a text-based specification for storing information in a structured fashion that can be easily read directly by humans and processed by computers. It’s a metalanguage, used to describe (and create) other languages. One of XML’s (and therefore TEI’s) greatest strengths is that it allows metadata to exist inside and around unstructured data. Each XML element begins with an opening tag, which comprises two brackets surrounding the element name, like this: <TEI>. Elements end with a closing tag, which is the same as the opening but with a backslash inserted before the name: </TEI>. Certain tags, such as <lb> (line break) follow a different format, because they don’t contain any content: <lb/>. This is actually a shorthand form of <lb></lb>.

Computing systems represent and process XML data as trees, as a hierarchical network of nodes, in something resembling the diagram below. This hierarchy can be described using terms borrowed from genealogy: for example, the main <ab> tag which contains most of the content in the figure below has three siblings, <milestone>, <ab>, and <ab>. Its parent is <div>, and grandparent is <body>. Any of the tags from <div> through <TEI> are considered ancestors.
Overview

Nellie McElroy Ledger Website: http://tiny.cc/nellie

Reference materials are accessible on the TEI & XML page of the Digital Humanities subject guide, at http://libguides.lib.rochester.edu/dh/tools/xml

If you are having trouble accessing the ledger website, it is most likely because you are connected to something other than the UR_Connected WiFi network.

This packet includes a number of handouts, such as

- A brief introduction to XML and TEI
- An example of a page from the ledger, fully encoded in TEI
- TEI Encoding Examples: different approaches to encoding a phrase in TEI, depending on the interest of a variety of scholars
- Well-Formedness in TEI: How to spot common encoding errors in TEI
- Cheat Sheet: A template of the different tags you will be using in this workshop, with instructions.
Good Shepherd Philadelphia Pa. Mrs Hawkins interceded and brought her to this city to make her home with her. Attended the Women's Court and then went down stairs to the other Court. Left station at 6-30.

Saturday October 11th. Reported for duty at 8. A. M. Attended the Womens Court and then went down stairs to the other court. Mrs Carrie Smith & Mrs McIntyre called regarding letters that Gertrude Johnson had written. Did duty on Main St. "Baker Theater" from 2 - 4-30. Play was "Sarah Paden in Kindling" - Suggestive all the way through. More than half the audience was girls from 12 - 17 years of age. Back to the station - Chief Quigley came up to my office and I reported work done. Left station at 7 o clock.
Well-Formedness in XML and TEI

(after Syd Bauman, Julia Flanders, and the Women Writers Project)

In text encoding, accuracy in your work is needed in four general areas:

- Well-formed: code must follow the basic rules of XML, use opening and closing tags, and not (mis)use any reserved characters such as the ampersand.
- Valid: TEI comprises a set of guidelines which names all of the available tags, stipulate which tags may appear in which, and to an extent which types of data (e.g., text, number, URL) they are allowed to contain.
- Sensible: The tags used make sense according to the editorial guidelines of the project.
- Accurate: Assertions made about the contents are factual, e.g., dates identified, handwriting deciphered correctly.

```xml
<placeName>Women's Court</placeName>

<milestone unit="line" rend="red"/>

<span rend="red">Molly Hawkins lived the sporting life at <placeName>Gibson Street</placeName> Theater</span>

stairs to the other Court.<lb>Left station at 6-30.

<placeName>Baker Theater</persName>

Carrie Smith & Mrs Taylor

<persName>Macy Halliday[/persName>

<placeName><addrLine>Main St</addrLine></placeName>

<persName><date>Gertrude</date> Johnson</persName>

<date when="2013-10-21">October 21</date>
```

Well-formed!

Ill-formed: @unit attribute missing a quotation mark

Ill-formed: <span> overlaps <placeName>, rather than containing it.

Ill-formed: \l<lb> is missing forward-slash (it's an empty element)

Ill-formed: start and end tags mismatch.

Ill-formed: start and end tags mismatch (XML is case sensitive!)

Ill-formed: & character has special meaning in XML. Ampersands must be encoded as &amp;

Ill-formed: closing tag contains square bracket.

Well-formed XML, but not valid TEI: <addrLine> is not allowed inside <placeName>

Well-formed XML, and valid TEI, but not sensible: Gertrude is not a date.

Well-formed XML, and valid TEI, but not true: Nellie was not referring to a date within the year 2013.
The quick brown fox jumped over the lazy dog.

The fox, doggedly nimble, vaulted the beagle, apt to bimble.

The fox, doggedly nimble, vaulted the beagle, apt to bimble.

The quick brown fox jumped over the lazy dog.

The quick brown fox jumped over the lazy dog.

The quick brown fox jumped over the lazy dog.

The quick brown fox jumped over the lazy dog.
key to the encoding examples

1. Encodes the visual representation of each character on the page. Similar approaches have been used in the Chymistry of Isaac Newton (Indiana University, webapp1.dlib.indiana.edu/newton), which encodes an alchemical writing system created by Newton, and in working with non left-to-right writing systems, such as the Westminster Leningrad Codex (Groves Center, tanach.us).

2. Encoding the grammatical structure, one of a number of models used in computational linguistics.

3. A poem, with two lines grouped together. Similar approaches have been used in the Oxford Text Archive (ota.ox.ac.uk).

4. Encoded as a medieval manuscript. Similar approaches are used in numerous projects, such as Sharing Ancient Wisdoms from King's College London (ancientwisdoms.ac.uk).

5. Using TEI to facilitate side-by-side translation into one or more languages, at the word, sentence, or paragraph level. The Vincent Van Gogh Museum (vangoghletters.org) used this approach to publish an interactive digital archive of his letters to his brother Theo and others.

6. Nominal encoding, where the mentions by name of persons, places, and things are disambiguated and linked to a searchable index. Notable projects utilizing this approach include the Seward Family Digital Archive (sewardproject.org).

7. Capturing and storing optical character recognition (OCR) data from mass digitization projects, such as the University of Michigan's Making of America project (quod.lib.umich.edu/m/moagrp).
Cheat Sheet

persName
This tag identifies proper names of people. It is not used for other types of nouns, such as mother, sister, or chief.

EXAMPLE
<persName ref="p017">Miss Graham</persName>

sic
Use this to indicate that a word is mispelled in the ledger.

EXAMPLE
Did duty on <sic>Suth</sic> Main Street.

ab
Enclose each day's ledger entry within a single <ab> tag (example 1). This tag is also used to enclose the blocks of text written in the margins of the page. Notice that that the @type attribute indicates that the <ab> tag appears in the margin, and the @subtype attribute indicates which margin.

EXAMPLE ONE
<ab> Reported for duty at 8 A.M. <lb/>
Attended the ... Back to station <lb/> and reported to the chief and <lb/> went home at 6-30. </ab>

EXAMPLE TWO
<ab type="margin" subtype="left">
<date when="1913-11-25" type="head" rend="red">
Tuesday <lb/>
Nov. 25. </date>
</ab>

placeName
This tag identifies proper names of places, and also street addresses. It is not used for other types of nouns, such as park, house, or police station.

EXAMPLES
<placeName>Gordon Theater</placeName>
<placeName>66 Front Street</placeName>

unclear
Use this to indicate that you are having difficulty reading Nellie's handwriting, and/or are unsure of what a word or phrase may be, or how it is spelled. Provide your best guess in the transcription. Reviewing the subject terms cataloged so far may be helpful.

EXAMPLE
Claude's sister <unclear>Meg</unclear> showed me her new bicycle.

&amp;
In XML-based languages, & must be written as &amp;.

date
Dates written on the ledger (example 1), and date-based headings (often in red) in the margin, in which case, a @type attribute and a @rend attribute, shown in ex. 2. Notice also that line breaks can be included within dates.

EXAMPLE ONE
Mary born <date when="1913-06-15">June 15 1913</date>

EXAMPLE TWO
<ab type="margin" subtype="left">
<date type="head" rend="red" when="1913-11-25">
Tuesday <lb/> Nov. 25. </date>
</ab>