SCHOLARLY EDITIONS: TEI TEXT ENCODING AND PUBLISHING

Michelle Dalmau, @mdalmau
Head of Digital Collections Services
Indiana University Libraries

Anna Kijas, @anna_kijas
Senior Digital Scholarship Librarian
Boston College Libraries

ARL Digital Scholarship Institute, January 8-12, 2018: #arldsi18w

GitHub: ARL Digital Scholarship

http://bit.ly/2B9u5EM

PART 1

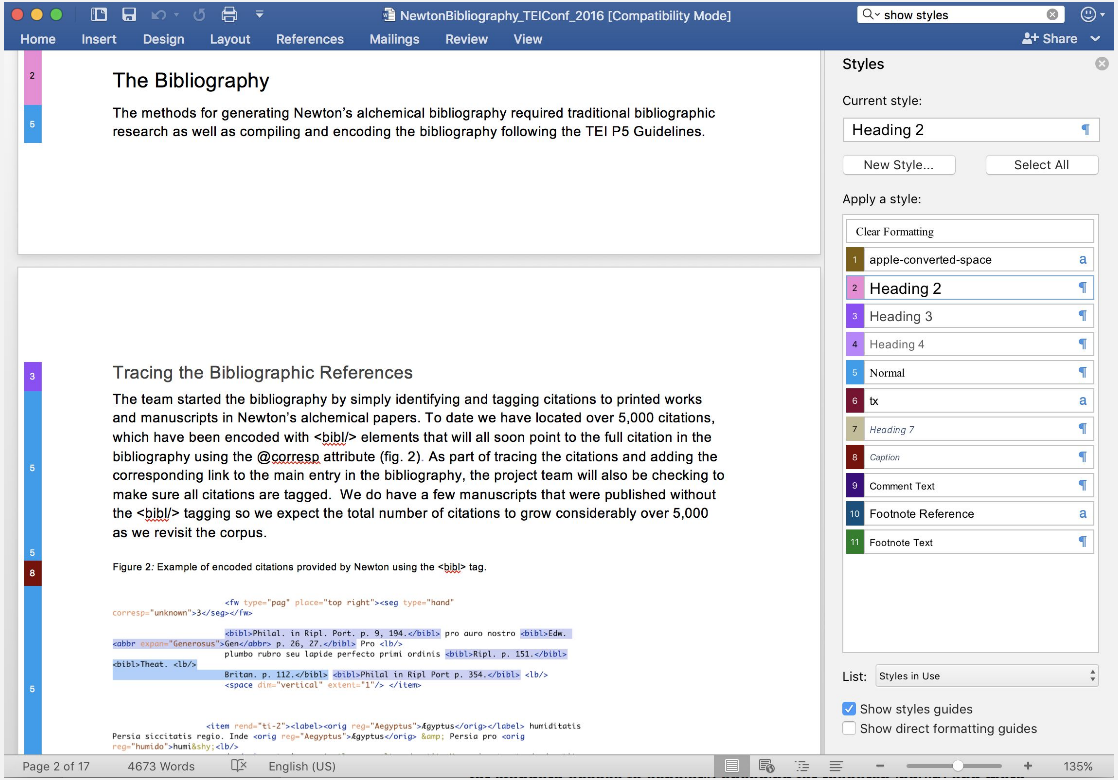
- Why encode?
- What is encoding?
- When to encode?

TEXT ENCODING OVERVIEW, OR, WHY MARKUP TEXTS?

- Store Information
  - Access
  - Preservation
- Share Information
  - Discovery (Searching & Browsing)
  - Interoperability & Portability
  - Harvesting & Repurposing
- Analyze Information
  - Linguistic Analysis
  - Concordances
  - Cluster Analysis
- Visualize Information
  - Interactive timelines
  - Map-based interfaces

ENCODED / MARKED TEXTS

"[T]here is no such thing as an unmarked text, and the markup systems laid upon documents to facilitate computerized analyses are marking orders laid upon already marked up material." (McGann, "Rethinking Textuality")



The Bibliography

The methods for generating Newton's alchemical bibliography required traditional bibliographic research as well as compiling and encoding the bibliography following the TEI P5 Guidelines.

Tracing the Bibliographic References

The team started the bibliography by simply identifying and tagging citations to printed works and manuscripts in Newton's alchemical papers. To date we have located over 5,000 citations, which have been encoded with <bibl/> elements that will all soon point to the full citation in the bibliography using the @corresp attribute (fig. 2). As part of tracing the citations and adding the corresponding link to the main entry in the bibliography, the project team will also be checking to make sure all citations are tagged. We do have a few manuscripts that were published without the <bibl/> tagging so we expect the total number of citations to grow considerably over 5,000 as we revisit the corpus.

Figure 2: Example of encoded citations provided by Newton using the <bibl> tag.

```
                              <fw type="pag" place="top right"><seg type="hand"
corresp="unknown">3</seg></fw>

                              <bibl>Philal. in Ripl. Port. p. 9, 194.</bibl> pro auro nostro <bibl>Edw.
<abbr expan="Generosus">Gen</abbr> p. 26, 27.</bibl> Pro <lb/>
                              plumbo rubro seu lapide perfecto primi ordinis <bibl>Ripl. p. 151.</bibl>
<bibl>Theat. <lb/>
                              Britan. p. 112.</bibl> <bibl>Philal in Ripl Port p. 354.</bibl> <lb/>
                              <space dim="vertical" extent="1"/> </item>


                         <item rend="ti-2"><label><orig reg="Aegyptus">Ægyptus</orig></label> humiditatis
Persia siccitatis regio. Inde <orig reg="Aegyptus">Ægyptus</orig> &amp; Persia pro <orig
reg="humido">humi&shy;<lb/>
```

WHAT IS ENCODING/MARKUP?

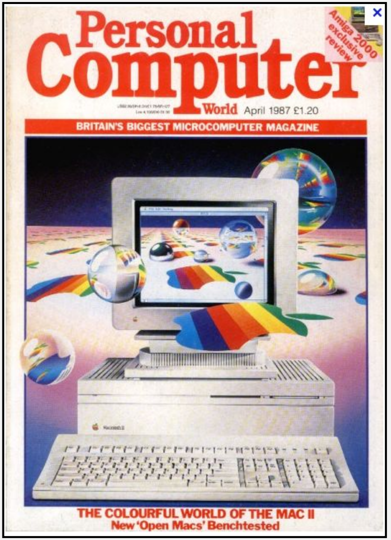
- XHTML for Web Pages
  - Hypertext Markup Language (expressed in XML)
- EAD for Finding Aids
  - Encoded Archival Description (EAD Tag Library)
- JATS for Journal Articles
  - Journal Article Tag Suite (JATS Tag Library)
- **TEI for Texts and Facsimile Images**
  - Guidelines for the Text Encoding Initiative (TEI Guidelines)

REPRESENTING THE TEXT WITH MARKUP

- Structural Features
  - Text divisions (chapters, sections, etc.), paragraphs, lists, tables, line groups, lines, etc.
- Content & Context
  - Metadata for the electronic and for the source document
  - References to people, places, events, organizations, etc. within the text (phrase-level)
  - Thematic and interpretive annotation
- Formatting & Design
  - Bold, italics, small case, indentations, color, dimensions, binding, watermarks, and other features of the material document

WHAT IS THE TEXT ENCODING INITIATIVE?

- TEI is:
  - a formally constituted organization, the TEI Consortium
  - a scholarly community—with an annual conference, open- access journal, and active email discussion list
  - a text encoding standard produced by that organization, TEI's Guidelines for Electronic Te Encoding and Interchange


Personal Computer World April 1987 £1.20
Amiga 2000 exclusive review
BRITAIN'S BIGGEST MICROCOMPUTER MAGAZINE
THE COLOURFUL WORLD OF THE MAC II
New 'Open Macs' Benchtested

For our purposes, TEI refers to the technical text encoding standard.

1987, Pretty in Pink
1987, birth of the personal computer and the TEI

WHEN DO WE USE THE TEI?

- Digital Scholarly Editions
  - Facsimile
  - Genetic
  - Diplomatic / Normalized
- Enhanced Discovery
  - Document-centric navigation and searching
- Deriving different views of texts

ACTIVITY 1: 20 MINUTES

Part 1: In groups, observe how the markup impacts the interface and functionality of each site; determine approaches to digital edition-making **(8 mins)**

Part 2: Re-convene as a whole and discuss observations and findings (**12 mins**)

Group 1
Chymistry of Isaac Newton
http://chymistry.org

Group 2
Swinburne Project
http://swinburneproject.org

Group 3
Willa Cather Archive
https://cather.unl.edu

Group 4
Shelley-Godwin Archive
http://shelleygodwinarchive.org

Group 5
Colonial Despatches
http://bcgenesis.uvic.ca

PART 2

- What are the TEI Guidelines?
- What is the TEI workflow?
- What is document analysis?

TEI GUIDELINES: QUICK OVERVIEW

- Text Encoding Initiative (TEI) / *Guidelines for Electronic Text Encoding and Interchange (TEI)*
- The TEI *Guidelines* "are addressed to anyone who works with any text in electronic form. They provide means of representing those features of a text which need to be identified explicitly in order to **facilitate processing of the text by computer programs**" (Sperberg-McQueen).
- TEI provides **elements**, **attributes**, and other mechanisms for encoding prose, poetry, drama, dictionaries, critical apparatus, linguistic corpora, and other scholarly and non-scholarly texts.

TEI GUIDELINES: QUICK OVERVIEW

The TEI Guidelines:

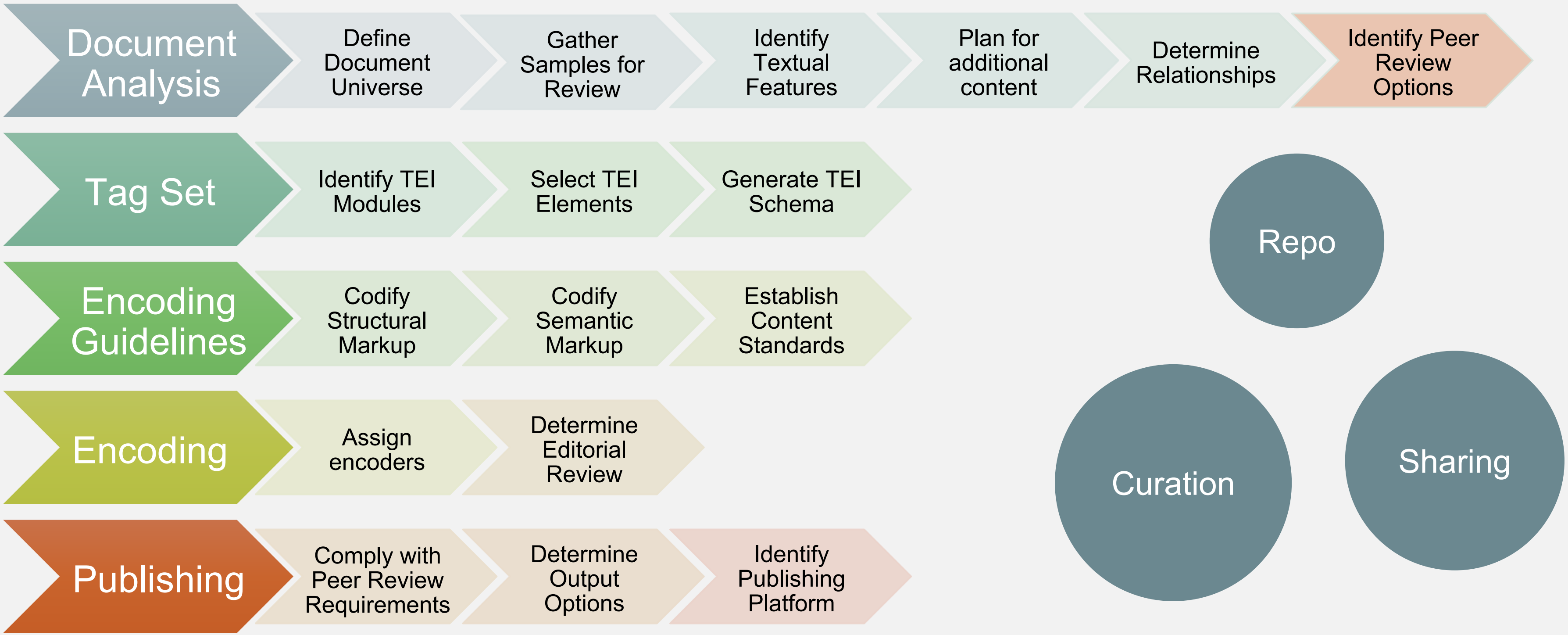
Can be applied lightly or heavily

Are designed as a set of modules/mechanisms that can be selected as needed:

core:	Elements common to all TEI documents
figures:	Tables, formulae, music notation, and figures
linking:	Linking, segmentation, and alignment
msdescription:	Manuscript description
namesdates:	Names and dates

Can be tailored

TEI WORKFLOW



Document Analysis: Define Document Universe → Gather Samples for Review → Identify Textual Features → Plan for additional content → Determine Relationships → Identify Peer Review Options

Tag Set: Identify TEI Modules → Select TEI Elements → Generate TEI Schema

Encoding Guidelines: Codify Structural Markup → Codify Semantic Markup → Establish Content Standards

Encoding: Assign encoders → Determine Editorial Review

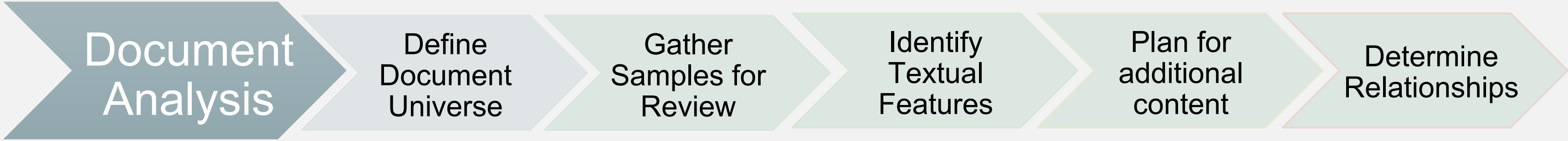
Publishing: Comply with Peer Review Requirements → Determine Output Options → Identify Publishing Platform

Repo

Curation

Sharing

DOCUMENT ANALYSIS



Document Analysis

Define Document Universe

Gather Samples for Review

Identify Textual Features

Plan for additional content

Determine Relationships

Document Universe	Sample Documents	Textual Features	Additional Content	Relationships
• One or Many? • What is it now and what should it be? • What is/are the genre? • Are the documents similar or different? • What do you know about the documents? • How many versions?	• Recognize the typical • Identify the atypical • Search for the unexpected (or leave room to account for it later)	• What is the level of representation? • How is the text structured and how is content presented? • What are your editorial interventions? • Appearance? • What parts of the documents will be omitted?	• Will additional content apply at the document level (or level of encoding) or at the phrase-level? • Annotations or glosses? • Introduction? • Commentary? • Translation? • Prosopography? • Subject analysis?	• How are the documents to be encoded related? • How are the parts of a document related?

ACTIVITY 2: 15 MINUTES

Part 1: **Individually**, conduct document analysis on the the excerpt of *O Pioneers;* **annotate** the handout. (**6 mins**)

Part 2: **Discuss** observations and findings (**9 mins**)

- Page 1 of handout contains sample questions to aid in document analysis.
- Annotate the document:
  - How is the text ordered? Sketch an outline.
  - X-out content that does not need to be encoded.
  - What are salient features of the texts (i.e., dialogue, preludes, etc.)?
  - Note structural elements (i.e., chapter headings, paragraphs, etc.)

PART 3

- What is XML?
- How to encode using TEI?

QUICK INTRODUCTION TO XML

XML, or eXtensible Markup Language, is a **non-proprietary meta language** for creating markup languages suited for different tasks, domains, and disciplines.

An XML markup language consists of "tags" used to define the structure and other features of a text.

XHTML:

```
    <p>(paragraph of text)</p>
    <img src="buffy.jpg"/>
    <a href="http://www.indiana.edu">Indiana University</a>
```

TEI:

```
    <sp who="#rosamond"> (speech) </sp>
    <lg> (line group, stanza) </lg>
    <salute>Dear Fred,</salute>
```

XML KEY TERMS

Elements are the basic, named structural units of an XML document (nouns of encoding)

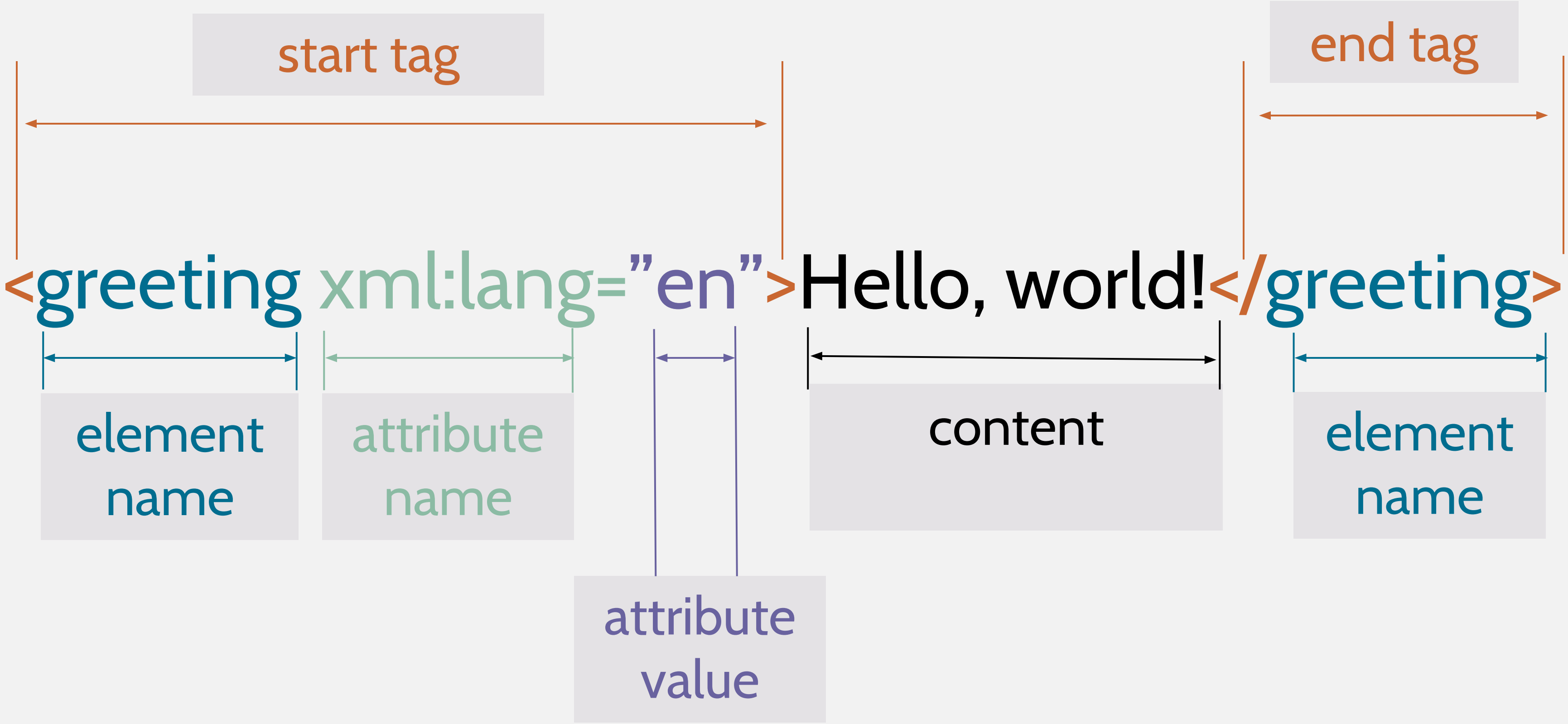
- <title>The Odyssey</title>

Attributes are name/value pairs (name="value") associated with elements (adjectives of encoding)

- <creator type="author">Homer</creator>
- An element may have multiple attributes

DTDs (Document Type Definitions) and **Schemas** define the rules that govern a particular type of XML document. They declare elements and attributes and the allowable content for those elements and attributes (grammar rules)

XML: ANATOMY OF AN ELEMENT



start tag

end tag

```
<greeting xml:lang=”en”>Hello, world!</greeting>
```

element name

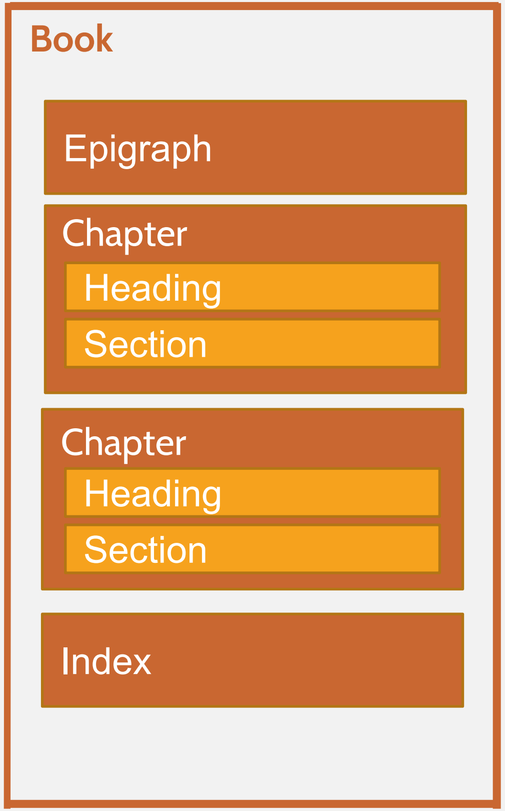
attribute name

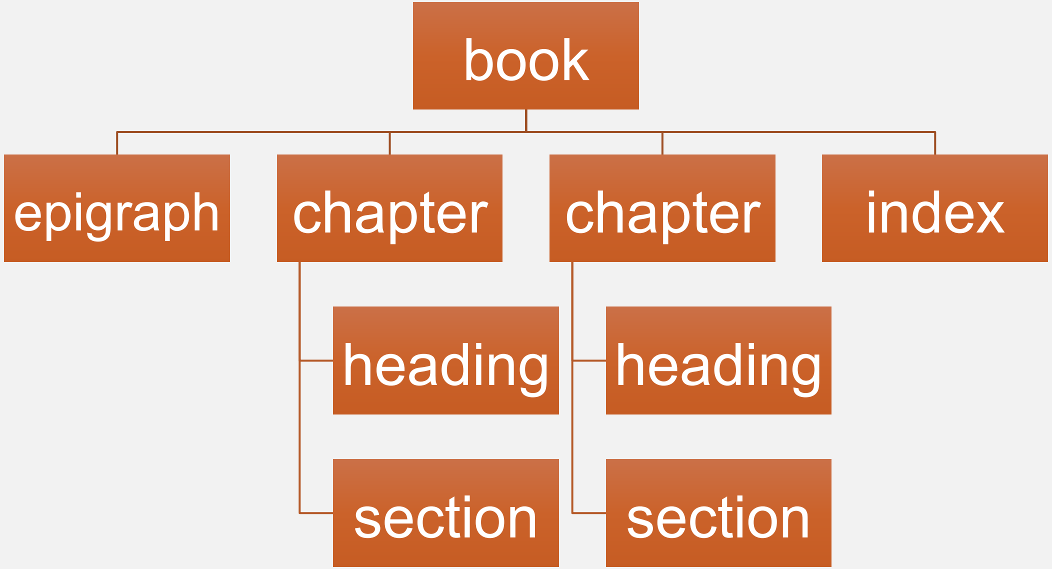
attribute value

content

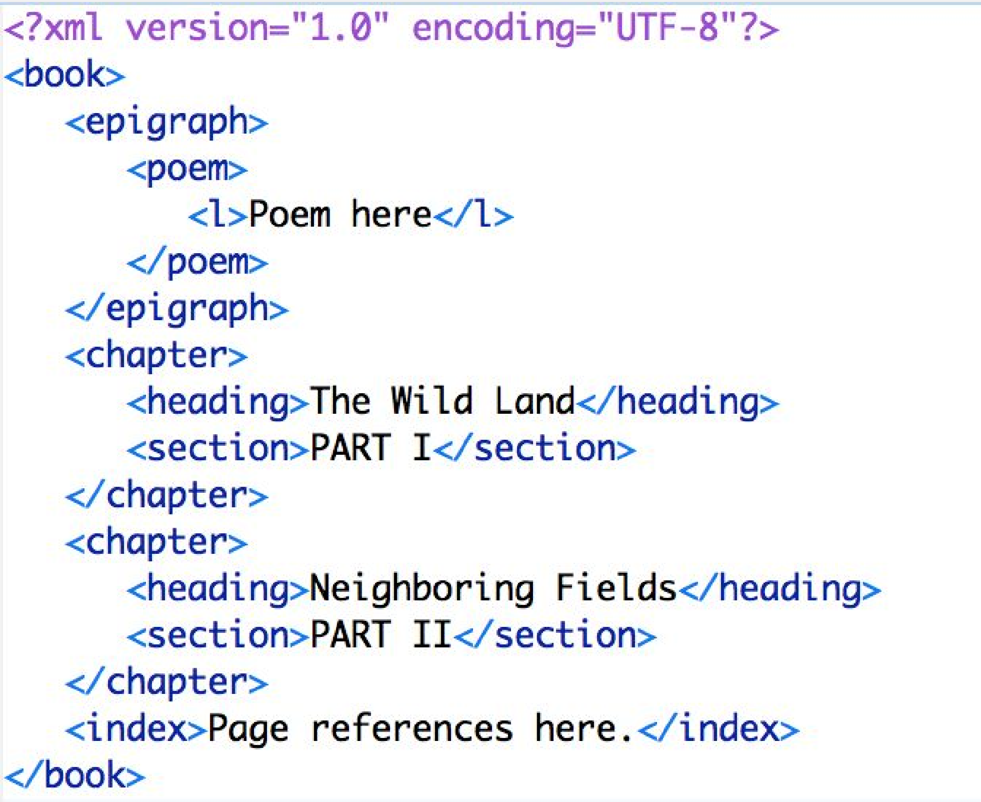
element name

XML REPRESENTATION: BOXES & TREE





XML REPRESENTATION: MARKUP



```xml
<?xml version="1.0" encoding="UTF-8"?>
<book>
   <epigraph>
      <poem>
         <l>Poem here</l>
      </poem>
   </epigraph>
   <chapter>
      <heading>The Wild Land</heading>
      <section>PART I</section>
   </chapter>
   <chapter>
      <heading>Neighboring Fields</heading>
      <section>PART II</section>
   </chapter>
   <index>Page references here.</index>
</book>
```

XML: WELL-FORMED AND VALID

All XML documents need to be well-formed according to some basic rules:

- Open and close all tags/elements
- Tags/elements may not overlap
- Attribute values must be quoted

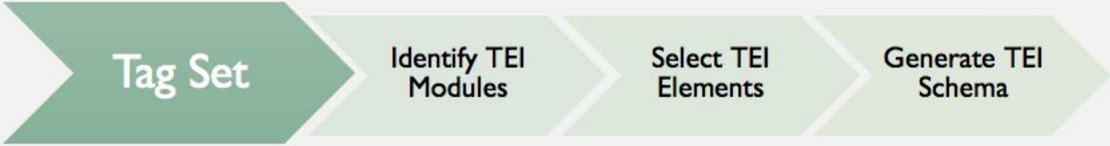
XML documents should be valid according to a DTD or Schema:

- Use the appropriate elements & attributes
- Adhere to the “grammar rules” (e.g., allowable attributes for elements)

Software programs help reinforce these principles

- XML Editors like Oxygen

TEI TAG SET



Tag Set

Identify TEI Modules

Select TEI Elements

Generate TEI Schema

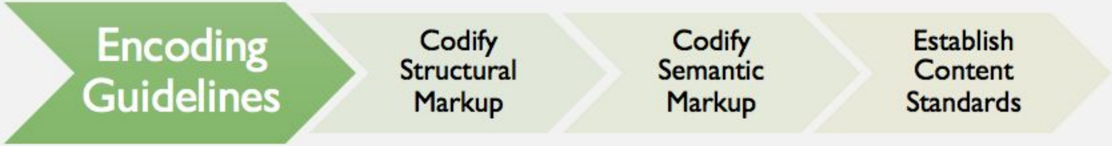
P5 Tag/Element Set:

http://www.tei-c.org/release/doc/tei-p5-doc/en/html/REF-ELEMENTS.html

Listing of the tag set with examples and relevant links to prose documentation

Document Analysis Annotation	TEI Tag	Notes
poem	<epigraph>	Before Part I
toc reveals novel in 5 parts/sections	<div type=”part”>	The Wild Land, Neighboring Fields, Winter Memories, etc.
paragraphs	<p>	Format body of text

TEI P5 GUIDELINES



TEI P5 Guidelines:
http://www.tei-c.org/release/doc/tei-p5-doc/en/html/index.html

Prose documentation with examples

Victorian Women Writers Project Guidelines:
https://wiki.dlib.indiana.edu/x/HQDIBg

Prose documentation tailored for the VWWP project

TEI P5: BASIC COMPONENTS

Encoding
Assign encoders
Determine Editorial Review

<TEI>: The root element of a TEI document

- <teiHeader>: The metadata header for a TEI document. Includes bibliographic, technical, administrative, and other metadata about the digital file and the analog source, if one exists.
- <text>: The text itself, e.g., the title page and chapters of a novel, the acts and scenes of a drama, the books or cantos of a long poem. The <text> element is further subdivided into:
  - <front>: Front matter, e.g, the title page(s), table of contents, potentially a preface or dedication
  - <body>: The main body of a document, excluding front and back matter
  - <back>: Back matter, e.g., indices, appendices

TEI HEADER

File Description **<fileDesc>** includes Source Description **<sourceDesc>**

Bibliographic description of the electronic and source files

Encoding Description **<encodingDesc>**

Documents relationship between the electronic texts and the source(s) from which it is derived

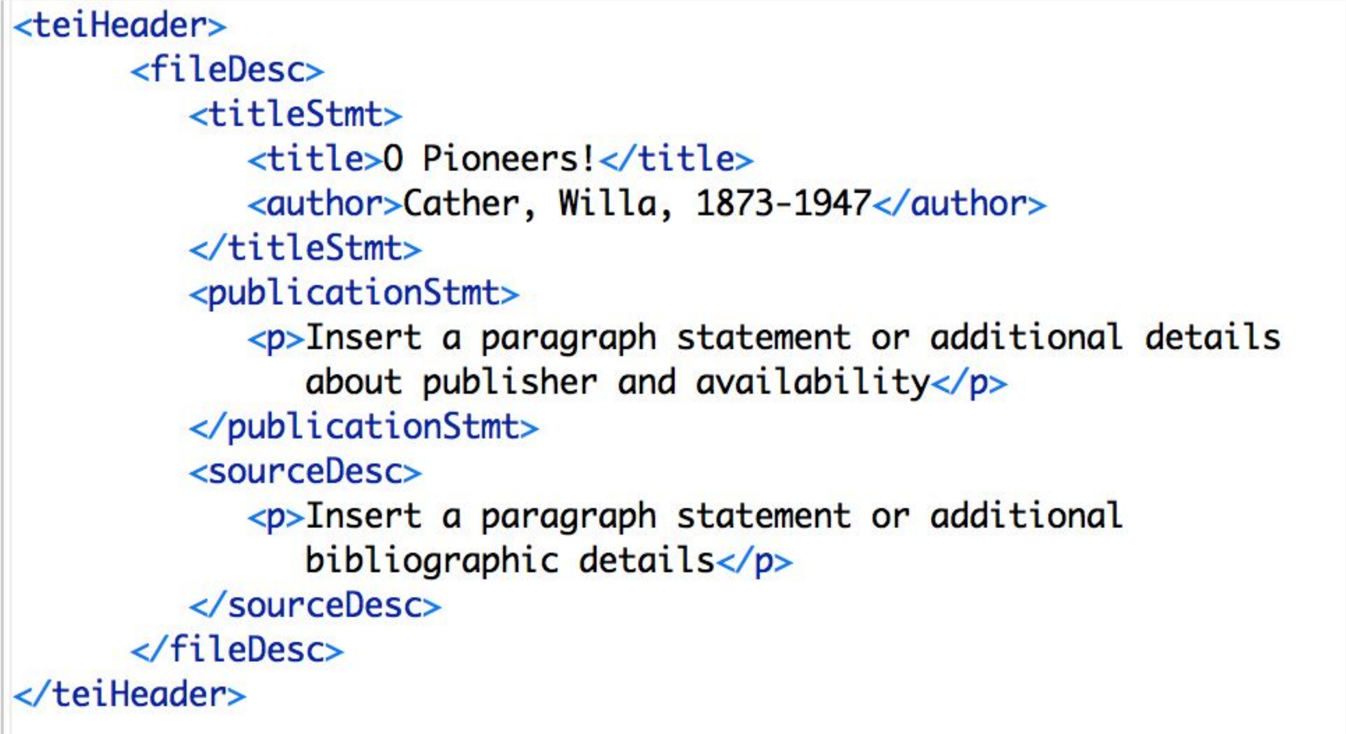
Profile Description **<profileDesc>**

Documents non-bibliographic aspects of the text such as languages, names, etc.

Revision Description **<revisionDesc>**

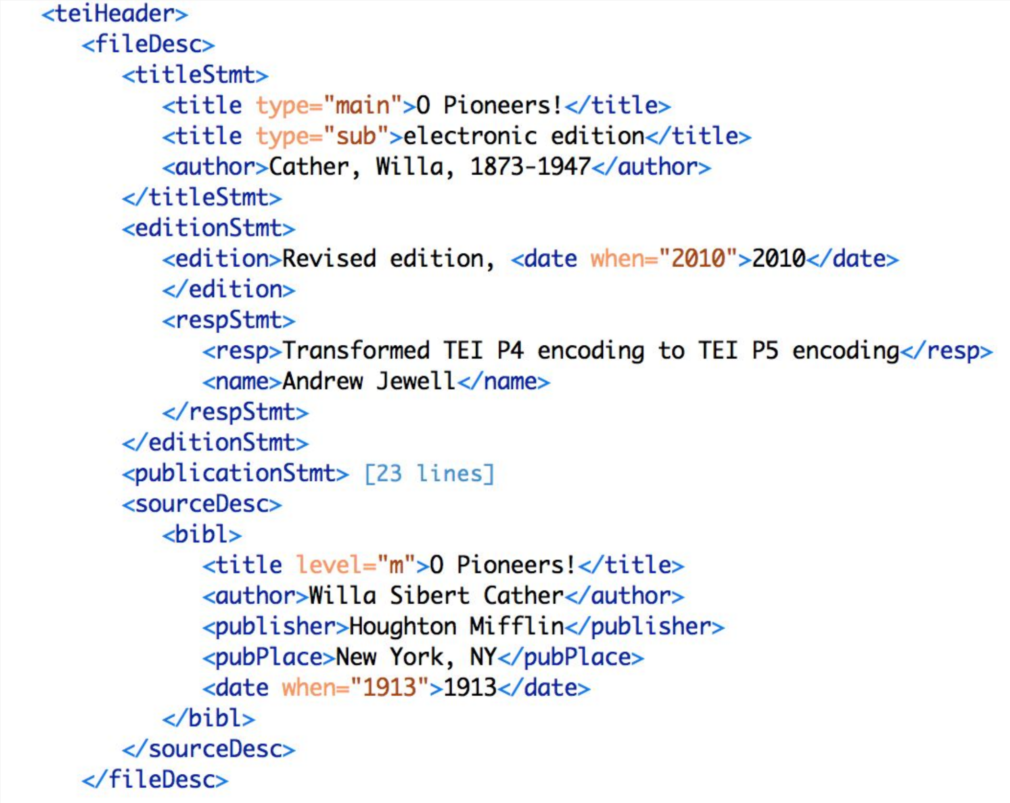
Documents changes made to the file (usually by editors)

TEI HEADER (REQUIRED ELEMENTS)



```
<teiHeader>
      <fileDesc>
         <titleStmt>
            <title>O Pioneers!</title>
            <author>Cather, Willa, 1873-1947</author>
         </titleStmt>
         <publicationStmt>
            <p>Insert a paragraph statement or additional details
               about publisher and availability</p>
         </publicationStmt>
         <sourceDesc>
            <p>Insert a paragraph statement or additional
               bibliographic details</p>
         </sourceDesc>
      </fileDesc>
</teiHeader>
```

TEI HEADER (SOURCE DESCRIPTION)



```
<teiHeader>
   <fileDesc>
      <titleStmt>
         <title type="main">O Pioneers!</title>
         <title type="sub">electronic edition</title>
         <author>Cather, Willa, 1873-1947</author>
      </titleStmt>
      <editionStmt>
         <edition>Revised edition, <date when="2010">2010</date>
         </edition>
         <respStmt>
            <resp>Transformed TEI P4 encoding to TEI P5 encoding</resp>
            <name>Andrew Jewell</name>
         </respStmt>
      </editionStmt>
      <publicationStmt> [23 lines]
      <sourceDesc>
         <bibl>
            <title level="m">O Pioneers!</title>
            <author>Willa Sibert Cather</author>
            <publisher>Houghton Mifflin</publisher>
            <pubPlace>New York, NY</pubPlace>
            <date when="1913">1913</date>
         </bibl>
      </sourceDesc>
   </fileDesc>
```

TEI P5: BASIC MARKUP

<div>: (division) is used for basic structural divisions of a text, e.g, volumes, chapters, sections, cantos, tables of contents, indices, appendices, etc. The **@**type attribute may be used to designate the type of the division.

```
<div type="chapter">…</div>
<div type="section">…</div>
<div type=”contents">…</div>
<div type="canto">…</div>
```

<head>: (heading) contains any type of heading, for example the title of a section, or the heading of a list, figure, table, etc.

<p>: (paragraph)

<pb>: (page break) marks the boundary between one page of a text and the next

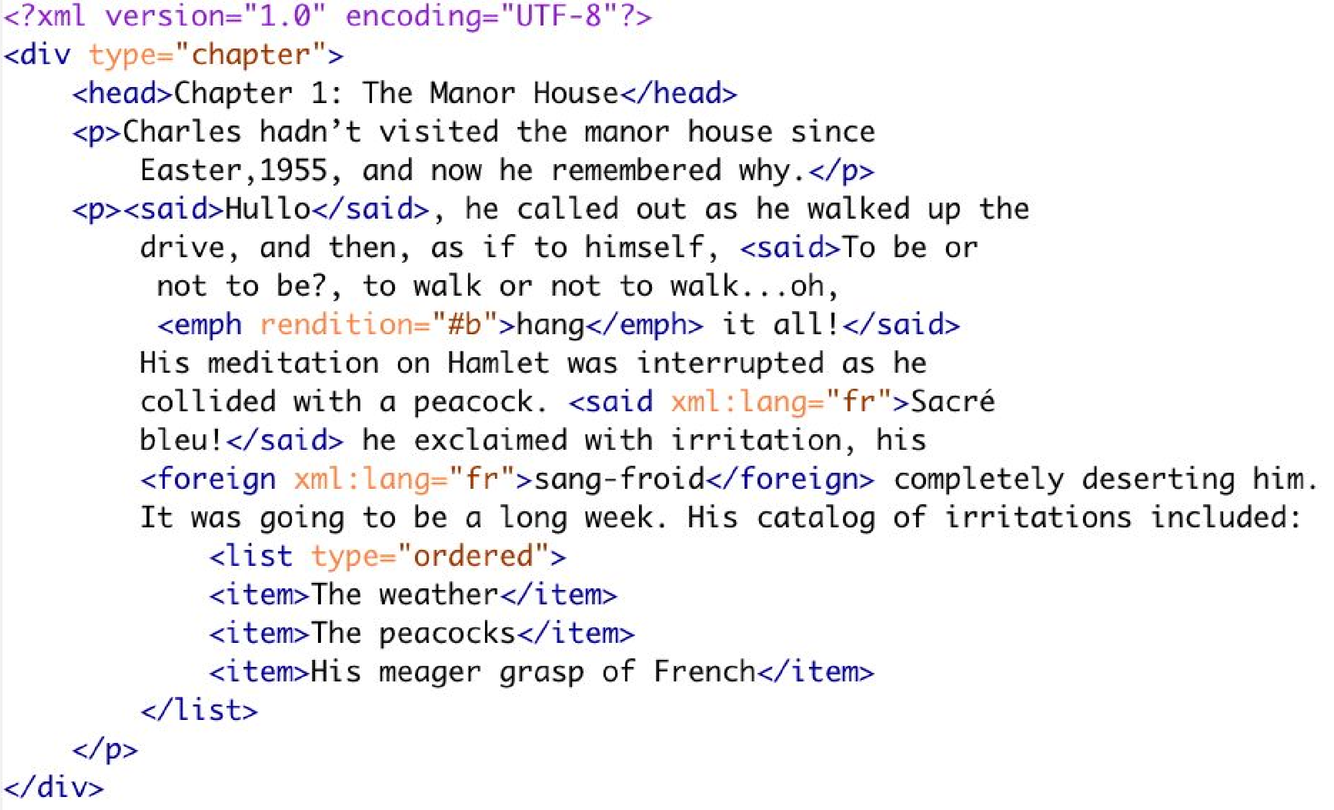
TEI P5: BASIC MARKUP: PROSE

Chapter 1: The Manor House

Charles hadn’t visited the manor house since Easter, 1955, and now he remembered why. “Hullo”, he called out as he walked up the drive, and then, as if to himself, “To be or not to be?, to walk or not to walk…oh, **hang** it all!” His meditation on Hamlet was interrupted as he collided with a peacock. “Sacré bleu!” he exclaimed with irritation, his sang-froid completely deserting him. It was going to be a long week. His catalog of irritations included:

1. The weather
2. The peacocks
3. His meager grasp of French

TEI P5: BASIC MARKUP: PROSE



```
<?xml version="1.0" encoding="UTF-8"?>
<div type="chapter">
    <head>Chapter 1: The Manor House</head>
    <p>Charles hadn’t visited the manor house since
        Easter,1955, and now he remembered why.</p>
    <p><said>Hullo</said>, he called out as he walked up the
        drive, and then, as if to himself, <said>To be or
         not to be?, to walk or not to walk...oh,
         <emph rendition="#b">hang</emph> it all!</said>
        His meditation on Hamlet was interrupted as he
        collided with a peacock. <said xml:lang="fr">Sacré
        bleu!</said> he exclaimed with irritation, his
        <foreign xml:lang="fr">sang-froid</foreign> completely deserting him.
        It was going to be a long week. His catalog of irritations included:
            <list type="ordered">
            <item>The weather</item>
            <item>The peacocks</item>
            <item>His meager grasp of French</item>
        </list>
    </p>
</div>
```

TEI P5: Basic Components of Personography

<listPerson>: groups descriptions about people that are found within these elements:

- **<person>**: provides information about an identifiable individual. Each person must have a unique xml:id.
    - <person xml:id=”pers_emil_bergson”>

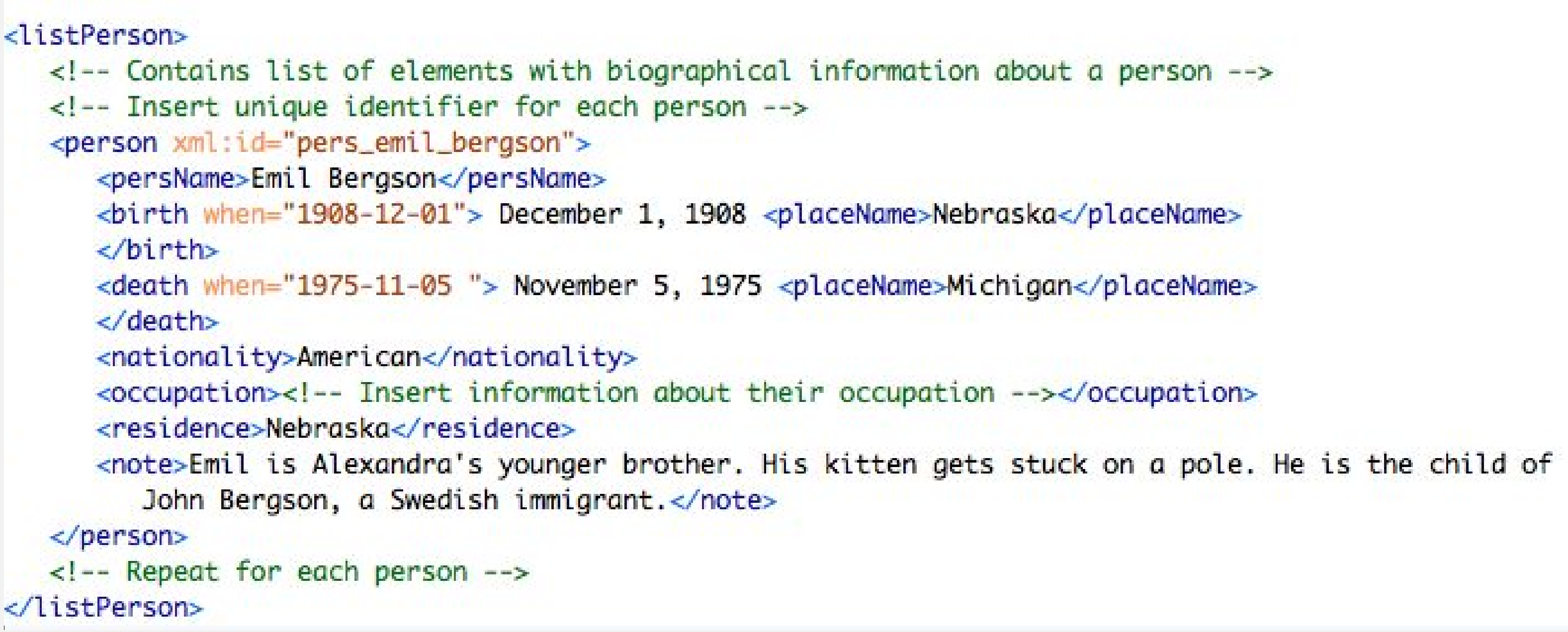
- **<persName>**: contains the name of a person, can be further broken down into forenames, surnames, honorifics, etc.

Each person can have additional elements that describe **personal characteristics**, such as <nationality>, <sex>, <age>, <occupation>, <residence>, etc. or **personal events**, such as <birth> and <death>.

Attributes such as @when, @notBefore, @notAfter, @from, or @to can be used with datable events.

- <birth when=”1908-12-01”>December 1, 1908</birth>

TEI P5: Personography



```
<listPerson>
   <!-- Contains list of elements with biographical information about a person -->
   <!-- Insert unique identifier for each person -->
   <person xml:id="pers_emil_bergson">
      <persName>Emil Bergson</persName>
      <birth when="1908-12-01"> December 1, 1908 <placeName>Nebraska</placeName>
      </birth>
      <death when="1975-11-05 "> November 5, 1975 <placeName>Michigan</placeName>
      </death>
      <nationality>American</nationality>
      <occupation><!-- Insert information about their occupation --></occupation>
      <residence>Nebraska</residence>
      <note>Emil is Alexandra's younger brother. His kitten gets stuck on a pole. He is the child of
         John Bergson, a Swedish immigrant.</note>
   </person>
   <!-- Repeat for each person -->
</listPerson>
```

ACTIVITY 3: 30 MINUTES

Part 1: **Encode** sample pages from *O Pioneers!* (**30 mins**)

- Metadata
- Text and Body
- Personography

Go to GitHub for instructions:
http://bit.ly/2B9u5EM

1. Download opioneers-excerpt.xml file from Box
2. Launch Oxygen XML Editor
3. File => Open => opioneers-excerpt.xml
4. Save file to your Desktop; name file: **opioneers-excerpt.xml**

(*Brief review Oxygen XML Editor and the TEI Schema*)

Begin encoding **all together**; Anna/Michelle will walkthrough the encoding with the class

Break: 30 Minutes!

PART 4

- How to publish TEI-encoded texts?

PUBLISHING SYSTEMS & PLATFORMS

Publishing

Comply with Peer Review Requirements

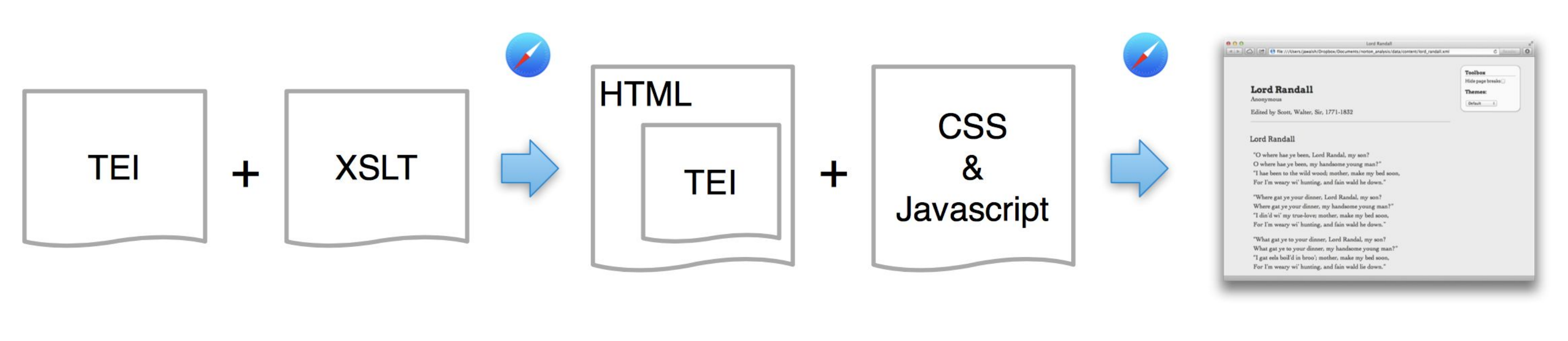
Determine Output Options

Identify Publishing Platform

- eXtensible Text Framework (XTF) by California Digital Library
- TAPAS: TEI Archiving, Publishing, and Access Service
- TEI Publisher
- TEI Boilerplate

- Other Open Source Systems:
  - Drupal, Omeka, Islandora, Open Journal Systems, etc.

XSLT vs TEI Boilerplate


TEI
+
XSLT
HTML
TEI
+
CSS
&
Javascript
Lord Randall
file:///Users/jawalsh/Dropbox/Documents/norton_analysis/data/content/lord_randall.xml
Toolbox
Hide page breaks □
Themes:
Default

Lord Randall
Anonymous

Edited by Scott, Walter, Sir, 1771-1832

Lord Randall

"O where hae ye been, Lord Randal, my son?
O where hae ye been, my handsome young man?"
"I hae been to the wild wood; mother, make my bed soon,
For I'm weary wi' hunting, and fain wald he down."

"Where gat ye your dinner, Lord Randal, my son?
Where gat ye your dinner, my handsome young man?"
"I din'd wi' my true-love; mother, make my bed soon,
For I'm weary wi' hunting, and fain wald he down."

"What gat ye to your dinner, Lord Randal, my son?
What gat ye to your dinner, my handsome young man?"
"I gat eels boil'd in broo'; mother, make my bed soon,
For I'm weary wi' hunting, and fain wald lie down."

PUBLISHING WITH TAPAS

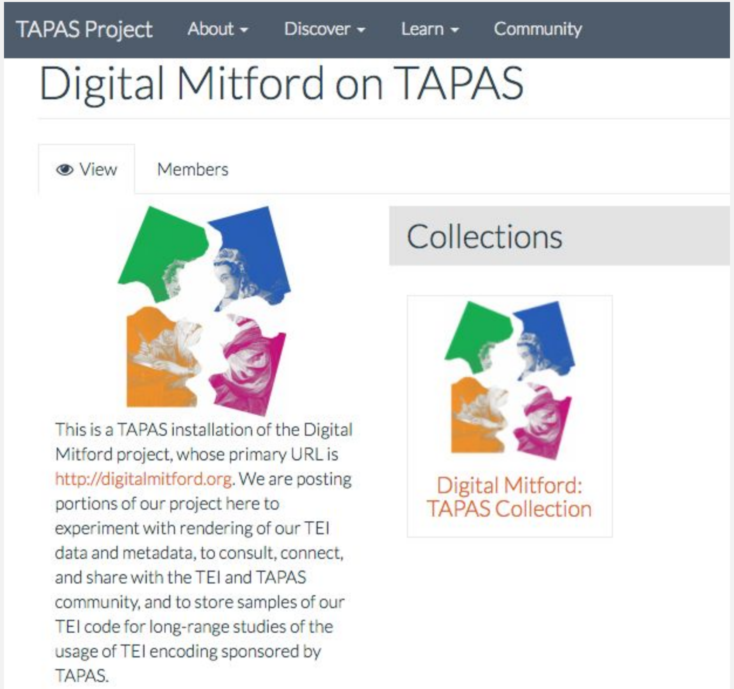
- **Records**: individual metadata record associated with a TEI file that can be independently uploaded or added to a project collection (paid account).

- **Projects**: TEI records can be associated with a project page.

- **Collections**: TEI files can be organized within collections by topic or theme. Files, such as personographies can be associated across collections.

- **Reading interface**: immediately renders TEI files using a TAPAS Generic or TEI Boilerplate stylesheet. Also has a raw XML view.

TAPAS PROJECTS

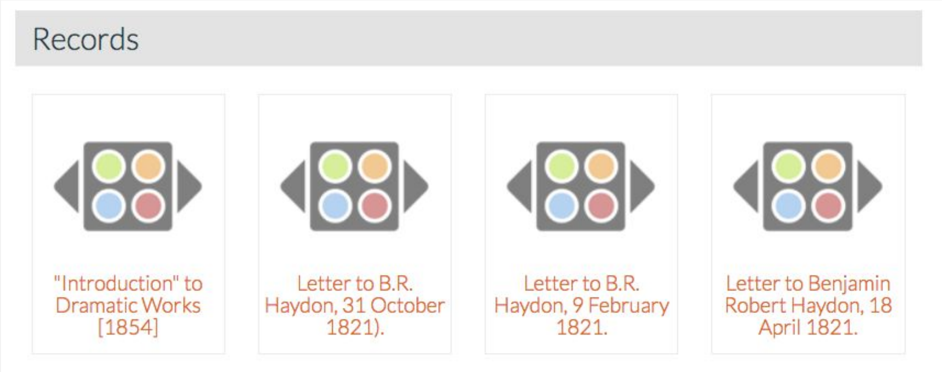


Digital Mitford on TAPAS

This is a TAPAS installation of the Digital Mitford project, whose primary URL is http://digitalmitford.org. We are posting portions of our project here to experiment with rendering of our TEI data and metadata, to consult, connect, and share with the TEI and TAPAS community, and to store samples of our TEI code for long-range studies of the usage of TEI encoding sponsored by TAPAS.

Collections

Digital Mitford: TAPAS Collection



Records

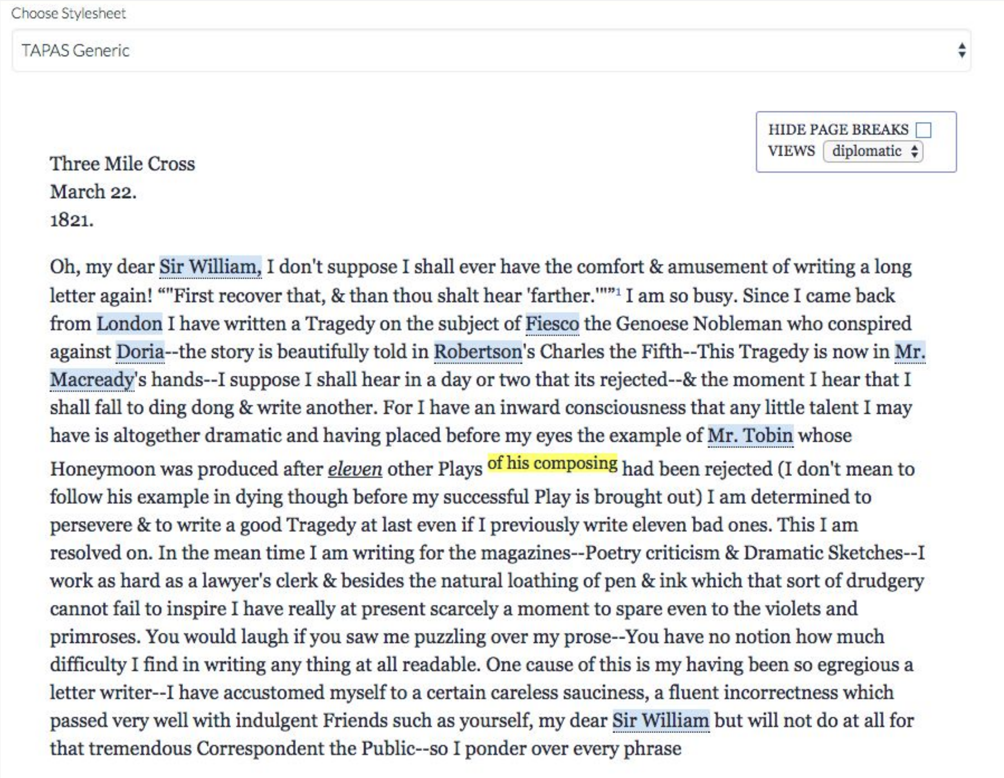
"Introduction" to Dramatic Works [1854]

Letter to B.R. Haydon, 31 October 1821).

Letter to B.R. Haydon, 9 February 1821.

Letter to Benjamin Robert Haydon, 18 April 1821.

TAPAS PROJECTS



Choose Stylesheet

TAPAS Generic

HIDE PAGE BREAKS □

VIEWS diplomatic

Three Mile Cross
March 22.
1821.

Oh, my dear Sir William, I don't suppose I shall ever have the comfort & amusement of writing a long letter again! “"First recover that, & than thou shalt hear 'farther.'"”[1] I am so busy. Since I came back from London I have written a Tragedy on the subject of Fiesco the Genoese Nobleman who conspired against Doria--the story is beautifully told in Robertson's Charles the Fifth--This Tragedy is now in Mr. Macready's hands--I suppose I shall hear in a day or two that its rejected--& the moment I hear that I shall fall to ding dong & write another. For I have an inward consciousness that any little talent I may have is altogether dramatic and having placed before my eyes the example of Mr. Tobin whose Honeymoon was produced after *eleven* other Plays of his composing had been rejected (I don't mean to follow his example in dying though before my successful Play is brought out) I am determined to persevere & to write a good Tragedy at last even if I previously write eleven bad ones. This I am resolved on. In the mean time I am writing for the magazines--Poetry criticism & Dramatic Sketches--I work as hard as a lawyer's clerk & besides the natural loathing of pen & ink which that sort of drudgery cannot fail to inspire I have really at present scarcely a moment to spare even to the violets and primroses. You would laugh if you saw me puzzling over my prose--You have no notion how much difficulty I find in writing any thing at all readable. One cause of this is my having been so egregious a letter writer--I have accustomed myself to a certain careless sauciness, a fluent incorrectness which passed very well with indulgent Friends such as yourself, my dear Sir William but will not do at all for that tremendous Correspondent the Public--so I ponder over every phrase

View letter: http://www.tapasproject.org/digital-mitford/files/letter-sir-william-elford-22-march-1821

ACTIVITY 4: 15 MINUTES

Part 1: Publish *opioneers-excerpt.xml* with TAPAS (5 minutes)

Part 2: Discussion: Compare views (TAPAS Generic and TEI Boilerplate) using the reading interface from the Willa Cather Archive version (10 mins)

Go to GitHub for instructions: http://bit.ly/2B9u5EM

Part 1: Publish your file in TAPAS

- Select the “Willa Cather O’Pioneers” collection”
  - Add and complete a new record
  - Upload your file

- Ta-da! You have "published" a TEI/XML file in TAPAS!

Part 2: Discussion

- Compare your published file with the version from the Willa Cather Archive