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, July 30-August 3, 2018: #arldsi18s

GitHub: ARL Digital Scholarship
https://bit.ly/2K5v9wQ

PART 1

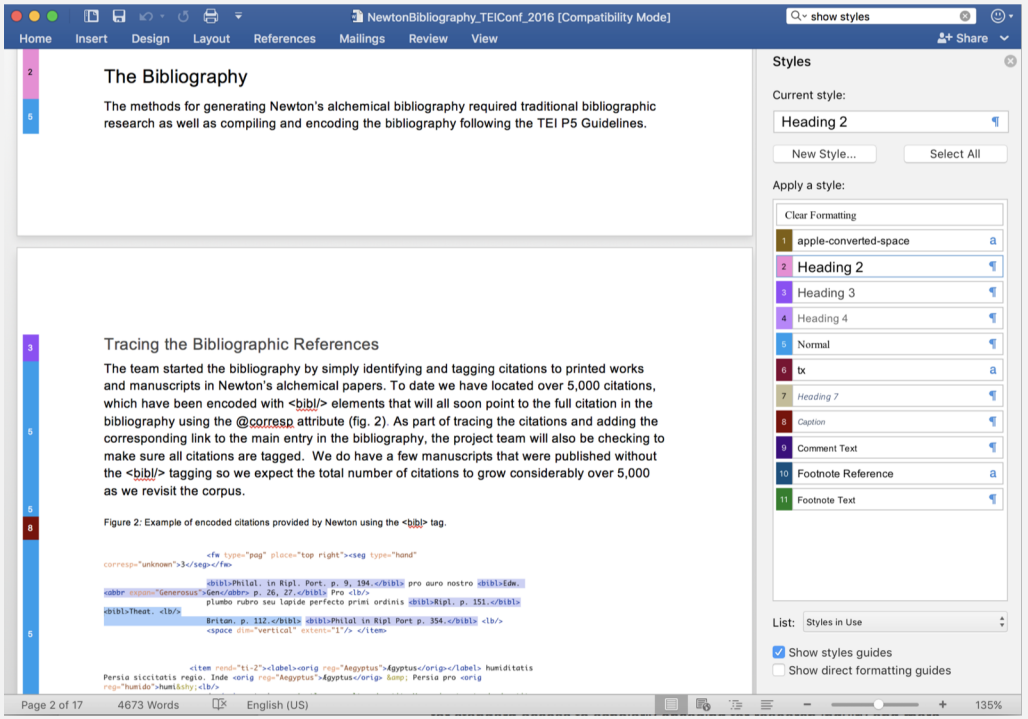
- Why encode?
- What is encoding?
- When to encode?

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")



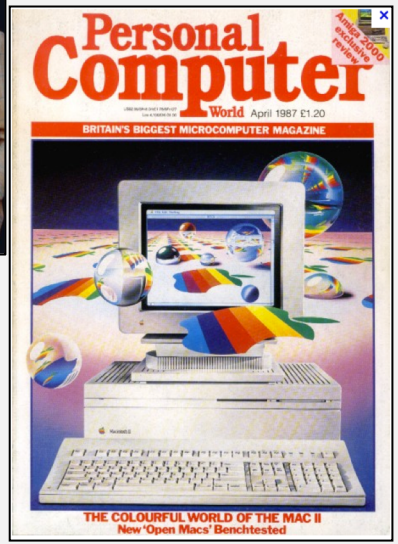
REPRESENTING THE TEXT WITH MARKUP

- 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 Text Encoding and Interchange

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
Map of Early Modern London
https://mapoflondon.uvic.ca

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

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

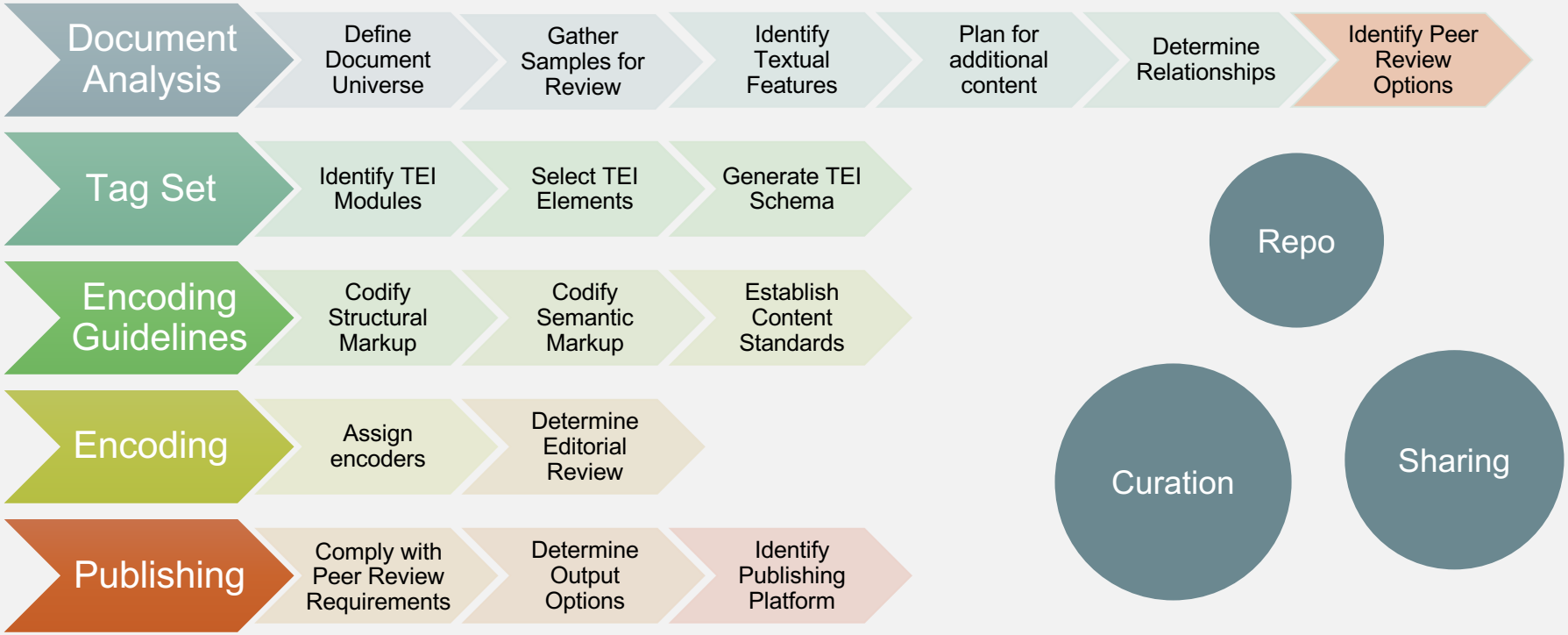
PART 2

- 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 WORKFLOW



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.)
  - Who are the characters in the story?

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)

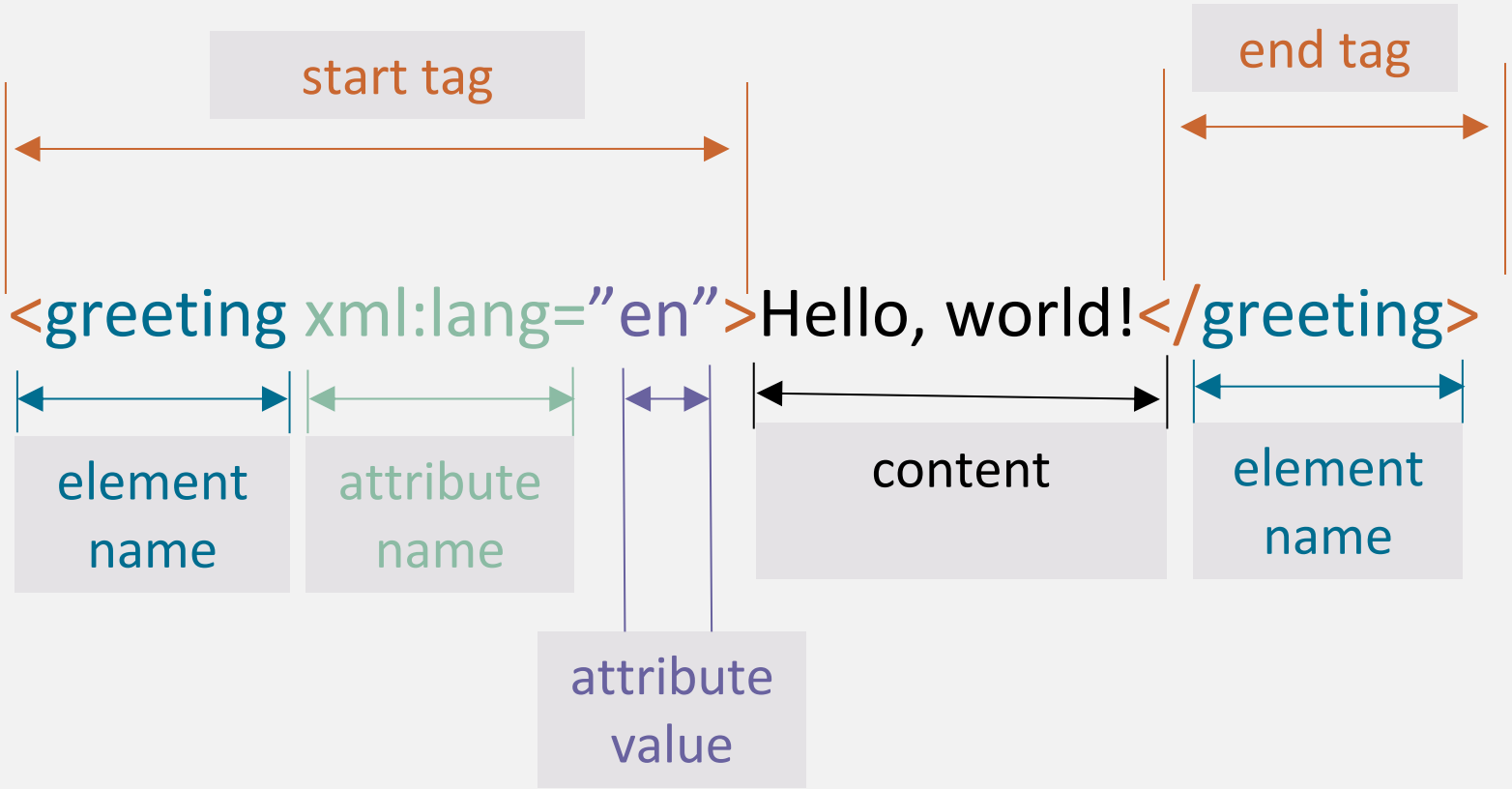
- <title>The Odyssey</title>

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

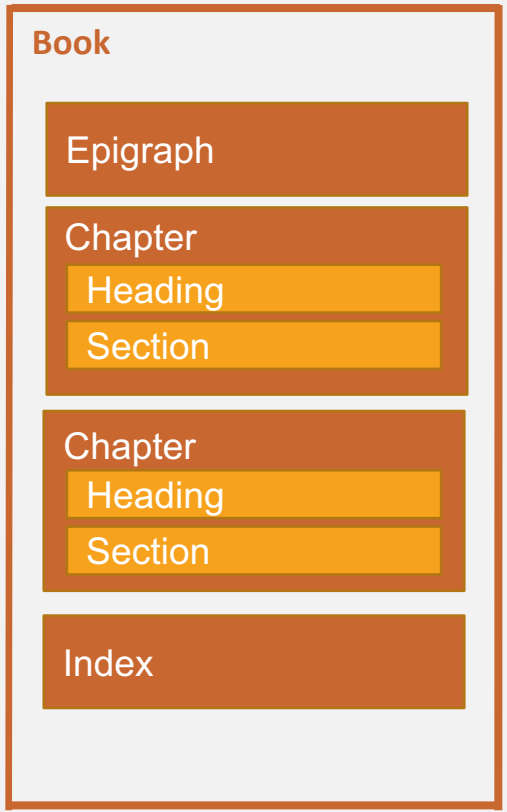
- <creator type="author">Homer</creator>
- An element may have multiple attributes

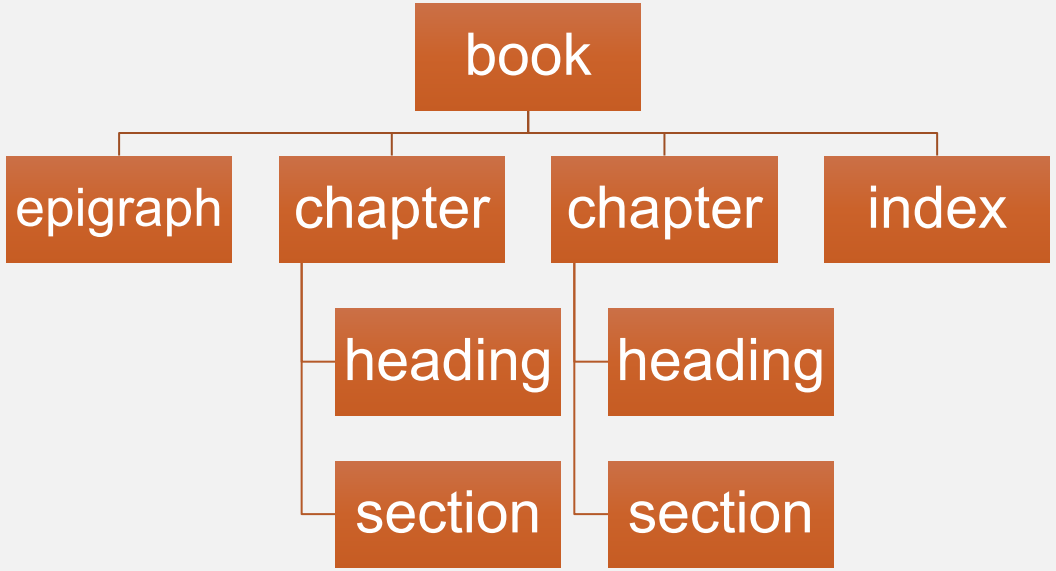
DTDs (Document Type Definitions) and **Schemas** define the rules that govern a particular type of XML document. They declare elements and attributes and the allowable content for those elements and attributes (grammar rules).

XML: ANATOMY OF AN ELEMENT



XML REPRESENTATION: BOXES AND TREES





XML REPRESENTATION: MARKUP

```
<?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

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		The Wild Land, Neighboring Fields, Winter Memories, etc.
paragraphs	<p>	Format body of text

TEI P5 GUIDELINES

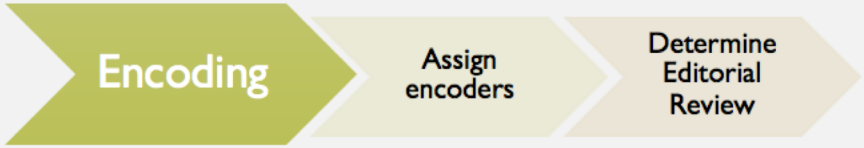
TEI P5 Guidelines:
http://www.tei-c.org/release/doc/tei-p5-doc/en/html/index.html

Prose documentation with examples

Shelley-Godwin Archive
http://shelleygodwinarchive.org/about/#encodingthesga

Prose documentation tailored for the SGA project

TEI P5: BASIC COMPONENTS



<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, prosopography, 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: 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,
   </person>
   <!-- Repeat for each person -->
</listPerson>
```

ACTIVITY 3: 35 MINUTES

Pre-break: Review Oxygen XML Editor (5 mins)

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

- Metadata (5 mins)
- Text and Body (10 mins)
- Personography (15 mins)

Go to GitHub for instructions:
https://bit.ly/2K5v9wQ

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.

Make sure XML files are valid before end of activity.

Break: 30 Minutes!

ACTIVITY 3: 35 MINUTES

Pre-break: Review Oxygen XML Editor (5 mins)

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

- Metadata (5 mins)
- Text and Body (10 mins)
- Personography (15 mins)

Go to GitHub for instructions:
https://bit.ly/2K5v9wQ

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.

Make sure XML files are valid before end of activity.

PART 4

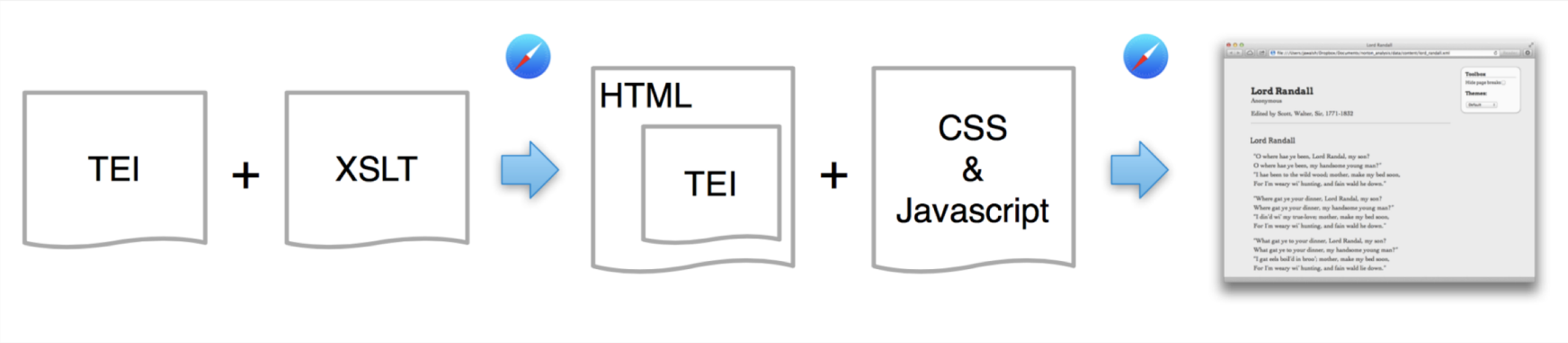
- How to publish TEI-encoded texts?

PUBLISHING SYSTEMS & PLATFORMS

- 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

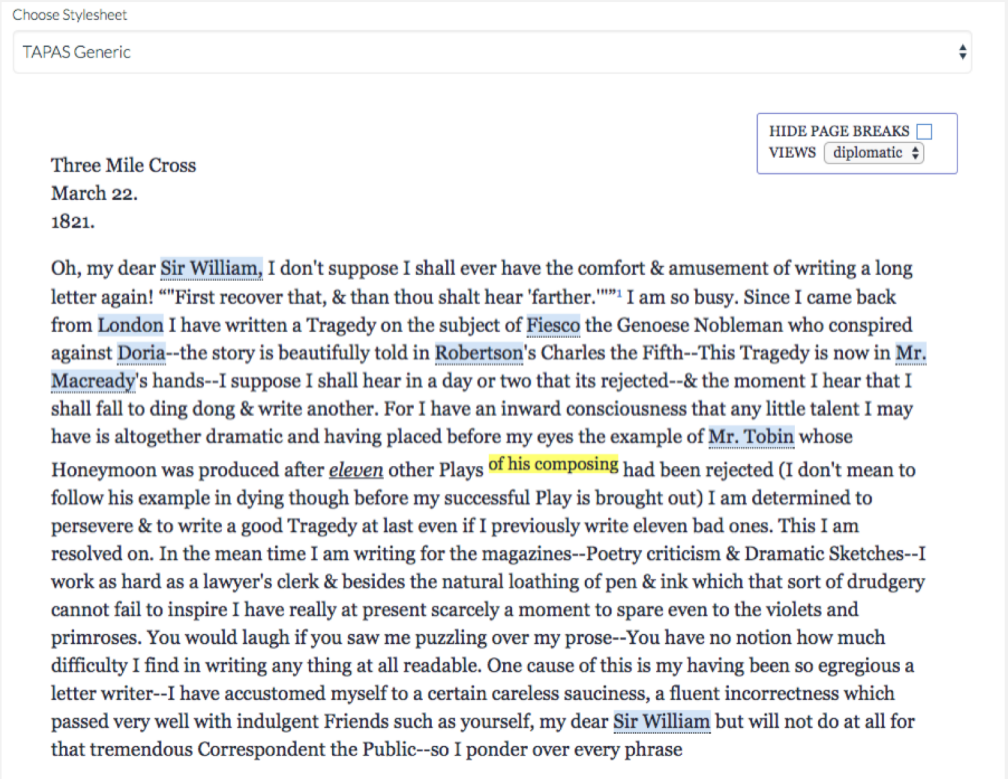


PUBLISHING WITH TAPAS

- **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

TAPAS PROJECTS



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:
https://bit.ly/2K5v9wQ

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:

https://cather.unl.edu/0017.html