

SCHOLARLY EDITIONS: TEI TEXT ENCODING AND PUBLISHING

Leigh Bonds, [@Leigh_Bonds](#)
Digital Humanities Librarian
Ohio State University Libraries

Anna Kijas, [@anna_kijas](#)
Senior Digital Scholarship Librarian
Boston College Libraries

ARL Digital Scholarship Institute, July - August , 2019: [#arldsi19s](#)

We would like to acknowledge Michelle Dalmau, Head, Digital Collections Services and Associate Librarian for Digital Collections Services and Library Technologies at Indiana University Libraries, who created the original slides.

These slides have been updated and modified for use in subsequent institutes. (Dalmau: June 2017; Dalmau/Kijas: January and July 2018; Kijas/Bonds: July 2019).

GitHub: ARL Digital Scholarship

bit.ly/2XjFsUG

LEARNING GOALS

- Understand use cases for text encoding; recognize why libraries are especially well-situated to support text encoding
- Review the workflow for TEI-encoding; learn about the various stages with special emphasis on the importance of document analysis
- Gain familiarity with the XML and TEI; learn about core TEI features
- Learn publishing options for TEI-encoded documents with a focus on the TAPAS repository and publishing platform

PART 1

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

ACTIVITY 1: 15 MINUTES

Part 1: Get into pairs and analyze characteristics.

Part 2: Reconvene as a whole and discuss analyses.

Cookie Analysis

Examine the cookie provided and discuss with your partner

- what characteristics are unique to the cookie?
- what characteristics distinguish it from others?

Make a list of the characteristics discussed and contribute to the collective list.

HTML compared with TEI XML

```
<div type="act">
  <head>Atto Terzo</head>
  <div type="scene">
    <head>Scena Prima</head>
    <stage>Querciuola, Costanzo, Brigida
      <lb/>del Capitano</stage>
    <sp><speaker>Quer.</speaker>
      <p>Ah ah ah ah, non ui potrei mai dire quanto uoi state
        bene: me parete un magnan naturale, e ui giuro che a
        pena <unclear>ai</unclear> riconosco, e mi parete quasi
        un <gap/>.</p>
    </sp>
    <sp><speaker>Gost.</speaker>
      <p>Può far mio padre, ch'io habbia d'andar cosi dinanzi
        alla mia innamorata.</p>
    </sp>
    <sp><speaker>Quer.</speaker>
      <p>Che credete che sia? Quando le donne uengon lisciate
        dinanzia lorquasti per questo no piaccian loro? E non è
        forse peggio'l carbon che'l solimata? anzi meglio, che so
        pur tigne un poco il uiso fuera, non guasta i denti dentro,
        e non corrompe il fiato.</p>
    </sp>
    <sp><speaker>Gost.</speaker>
      <p>Hor su che ho afar? di uia.</p>
    </sp>
    <sp><speaker>Quer.</speaker>
      <p>La prima cosa, perche uoi non siate conosciuto
        bisognache contrasacciate la uoce a guisa di questi
        magnani, con dire in un tuon mezo fioco. Chi vuol donne,
        acconciar chiaui in toppe, e toppe rotte? prouate un poco
        se sapete dire. Tenete su queste toppe.</p>
    </sp>
    [...and the rest of the scene]
  </div>
  [...and the other scenes in the act]
</div>
```

```
<div class="act">
  <h1>Atto Terzo</h1>
  <div class="scene">
    <h2>Scena Prima</h2>
    <p>Querciuola, Costanzo, Brigida
      <br/>del Capitano</p>
    <p>Quer.</p>
    <p>Ah ah ah ah, non ui potrei mai dire quanto uoi state
      bene: me parete un magnan naturale, e ui giuro che a
      pena [ai?] riconosco, e mi parete quasi
      un .</p>

    <p>Gost.</p>
    <p>Può far mio padre, ch'io habbia d'andar cosi dinanzi
      alla mia innamorata.</p>

    <p>Quer.</p>
    <p>Che credete che sia? Quando le donne uengon lisciate
      dinanzia lorquasti per questo no piaccian loro? E non è
      forse peggio'l carbon che'l solimata? anzi meglio, che so
      pur tigne un poco il uiso fuera, non guasta i denti dentro,
      e non corrompe il fiato.
    </p>
    <p>Gost.</p>
    <p>Hor su che ho afar? di uia.</p>

    <p>Quer.</p>
    <p>La prima cosa, perche uoi non siate conosciuto
      bisognache contrasacciate la uoce a guisa di questi
      magnani, con dire in un tuon mezo fioco. Chi vuol donne,
      acconciar chiaui in toppe, e toppe rotte? prouate un poco
      se sapete dire. Tenete su queste toppe.</p>

    [...and the rest of the scene]
  </div>
  [...and the other scenes in the act]
</div>
```

Source: Syd Bauman, Julia Flanders, and
the Women Writers Project (2007).
https://www.wwp.northeastern.edu/outreach/seminars/current/presentations/overview/overview_newer_tutorial_06.xhtml

Representing Text with TEI XML



ATTO TERZO

SCENA PRIMA.

Querciuela, Gostanzo, Brigida
del Capitano.

Quer. *El ah ah ah, non ui potrei
mai dire quanto ui fiato be
ne: mi parete un magnan natu-
rale, e ui giuro che a pena
riconosco, e mi parete qua-
si un martirio.*

Gost. *Può far mio padre, ch'io habbia d'andar
cofi dimenchi alla mia innamorta.*

Quer. *Che credete che sia? Quando le donne ven-
gon lasciate dimenchi a i lor guisti per que-
sto non piaccion loro? E non è forse peggio l'
carbon che l' salmasa? anzi meglio, che se
pur tigne un poco il nifo fuori, non guasta i
denti dentro, e non corrompe il fiato.*

Gost. *Hor sù che ho a far? di mia.*

Quer. *La prima cosa, perche noi non fate conosci-
to bisogne che coverasfacciate la uoce a guisa
di questi magnani: con dire in un tuon me-
zo fioco. Chi vuol donne, acconciar chiau in
toppe e soppe rotte, state a udir come dico
io, oh, chi vuol donne acconciar chiau in
toppe, e soppe rotte? prouate un poco se sa-
rete dire. Tenete su queste toppe.*

[...and the rest of the scene]

[...and the other scenes in the act]

```
<div type="act">
  <head>Atto Terzo</head>
  <div type="scene">
    <head>Scena Prima</head>
    <stage>Querciuela, Costanzo, Brigida
      <lb/>del Capitano</stage>
    <sp><speaker>Quer.</speaker>
    <p>Ah ah ah ah, non ui potrei mai dire quanto uoi state
      bene: me parete un magnan naturale, e ui giuro che a
      pena <unclear>ai</unclear> riconosco, e mi parete quasi
      un <gap/>.</p>
    </sp>
    <sp><speaker>Gost.</speaker>
    <p>Può far mio padre, ch'io habbia d'andar cosi dinanzi
      alla mia innamorta.</p>
    </sp>
    <sp><speaker>Quer.</speaker>
    <p>Che credete che sia? Quando le donne uengon lasciate
      dinanzia lorquasi per questo no piaccion loro? E non è
      forse peggio l' carbon che'l solimata? anzi meglio, che so
      pur tigne un poco il uiso fuora, non guasta i denti dentro,
      e non corrompe il fiato.</p>
    </sp>
    <sp><speaker>Gost.</speaker>
    <p>Hor su che ho afar? di uia.</p>
    </sp>
    <sp><speaker>Quer.</speaker>
    <p>La prima cosa, perche uoi non siate conosciuto
      bisognache contrasacciate la uoce a guisa di questi
      magnani, con dire in un tuon mezo fioco. Chi vuol donne,
      acconciar chiau in toppe, e toppe rotte? prouate un poco
      se sapete dire. Tenete su queste toppe.</p>
    </sp>
    [...and the rest of the scene]
  </div>
</div>
[...and the other scenes in the act]
```

Source: Syd Bauman, Julia Flanders, and the
Women Writers Project (2007).
https://www.wwp.northeastern.edu/outreach/seminars/current/presentations/overview/overview_newer_tutorial_O6.xhtml

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

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

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 2: 15 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 (7 mins)

Group 1

Chymistry of Isaac Newton

<http://chymistry.org>

Group 2

Willa Cather Archive

<https://cather.unl.edu>

Group 3

Shelley-Godwin Archive

<http://shelleygodwinarchive.org>

Group 4

Folger Digital Texts

<https://www.folgerdigitaltexts.org/>

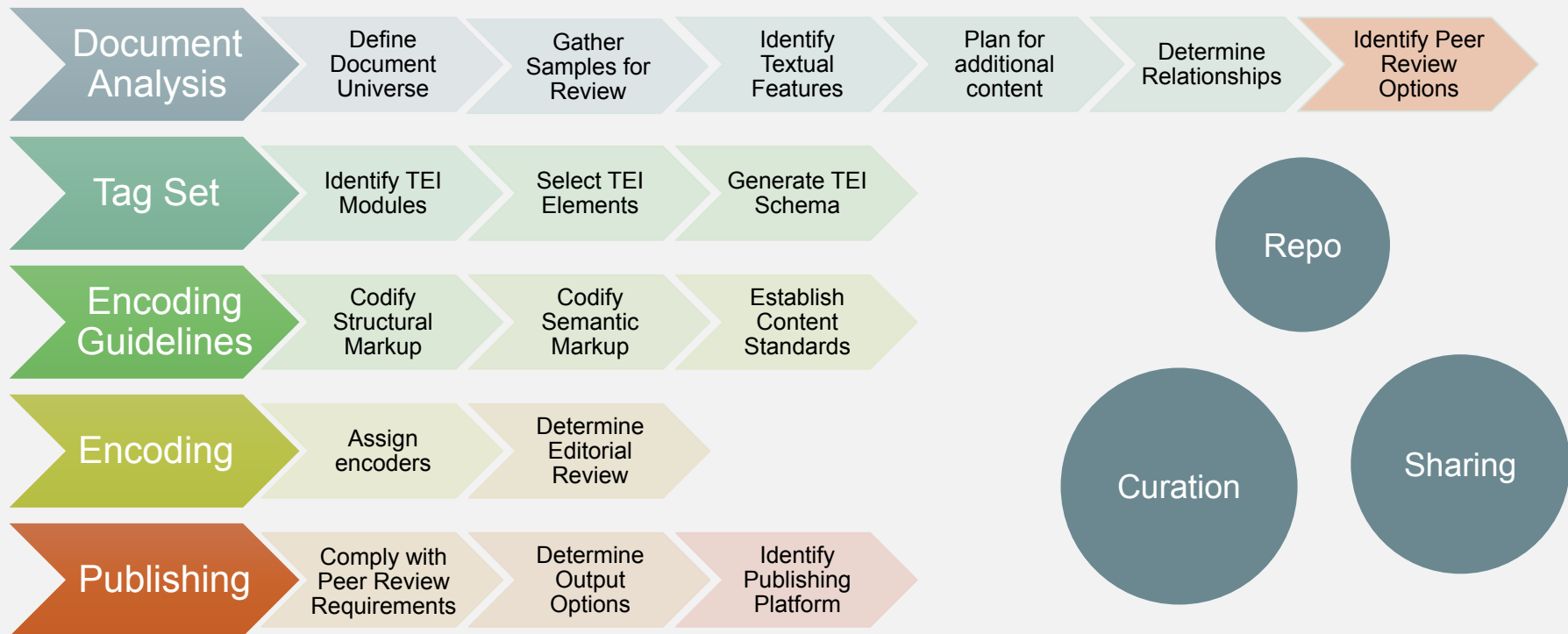
PART 2

- 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
<ul style="list-style-type: none">• 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?	<ul style="list-style-type: none">• Recognize the typical• Identify the atypical• Search for the unexpected (or leave room to account for it later)	<ul style="list-style-type: none">• 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?	<ul style="list-style-type: none">• 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?	<ul style="list-style-type: none">• How are the documents to be encoded related?• How are the parts of a document related?

ACTIVITY 3: 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>`

``

`Indiana University`

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**)

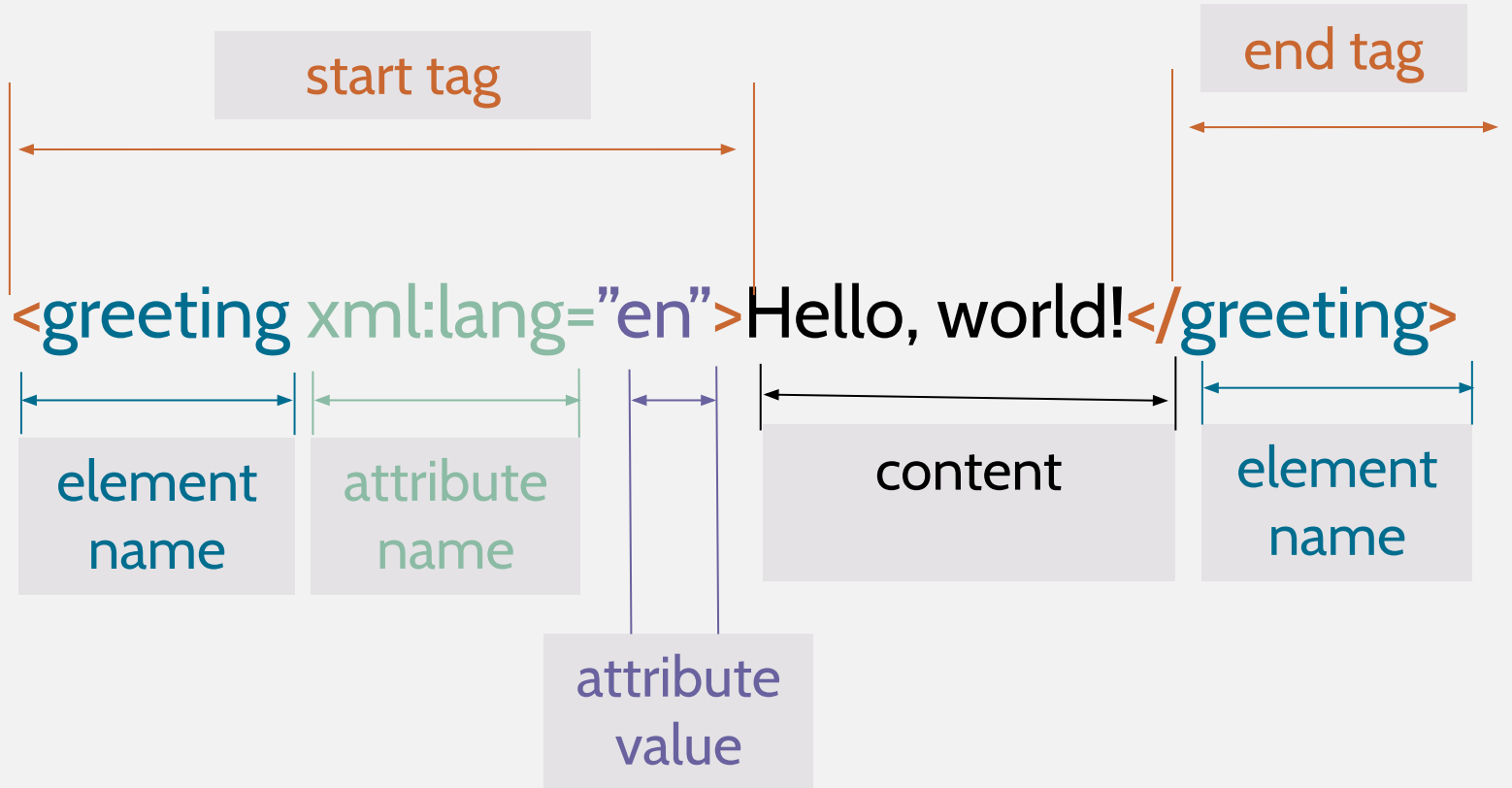
- `<title>The Odyssey</title>`

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

Book

Epigraph

Chapter

Heading

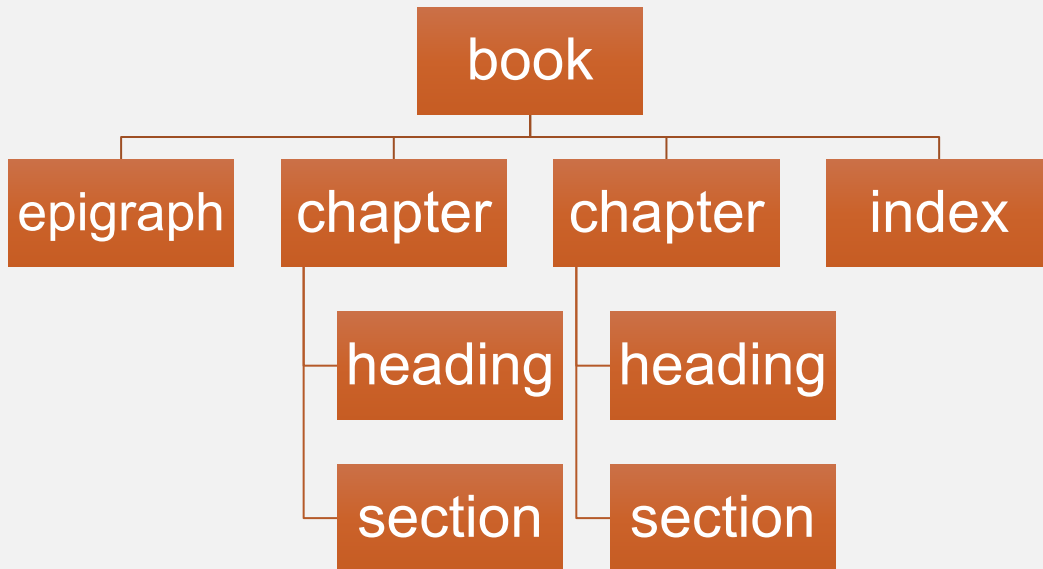
Section

Chapter

Heading

Section

Index



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	<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

Shelley-Godwin Archive

<http://shelleygodwinarchive.org/about/#encodingthesga>

Prose documentation tailored for the SGA 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, prosopography, etc.

Revision Description **<revisionDesc>**

Documents changes made to the file (usually by editors)

TEI HEADER (REQUIRED ELEMENTS)

```
<teiHeader>
  <fileDesc>
    <titleStmt>
      <title>0 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">0 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">0 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>
```


Break: 30 Minutes!

ACTIVITY 4: 35 MINUTES

Review Oxygen XML Editor (5 mins)

Encode sample pages from *O Pioneers!* (25 mins)

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



Go to GitHub for instructions:

bit.ly/2XjFsUG

1. Download opioneers-excerpt.xml file
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/Leigh will walk through the encoding with the class.

Make sure XML files are valid before end of activity.

PART 4

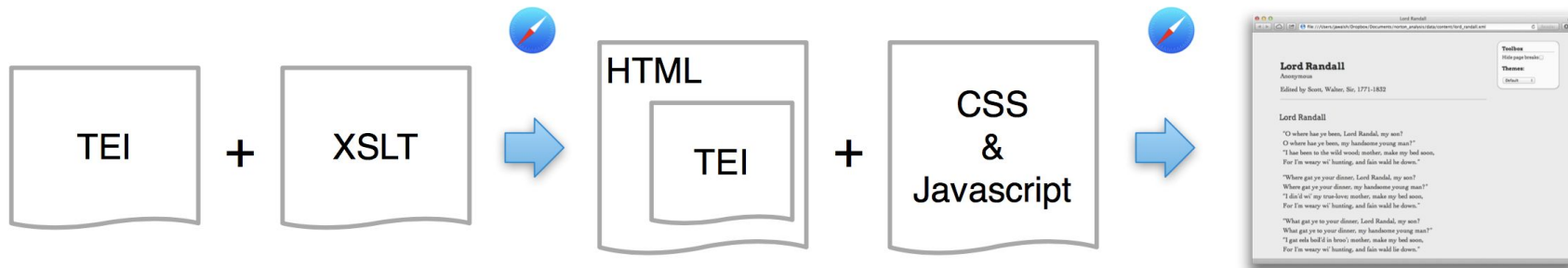
- How to publish TEI-encoded texts?

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 Project](#) [About ▾](#) [Discover ▾](#) [Learn ▾](#) [Community](#)

Digital Mitford on TAPAS

 View

Members



Collections



Digital Mitford:
TAPAS Collection

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.

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 5: 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:
bit.ly/2XiFsUG

Part 1: Publish your file in TAPAS

- Login to 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/OO17.html>

DISCUSSION & WRAP-UP

- Questions, further discussion?
- How do workshop participants imagine using these new digital scholarship skills?
- Who at their institution could help with this type of digital scholarship work?