# 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: #arldsi19

Technologies at Indiana University Libraries, who created the original slides.

We would like to acknowledge Michelle Dalmau, Head, Digital Collections

Services and Associate Librarian for Digital Collections Services and Library

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>Ouer.</speaker>
   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/>.
  </sp>
  <sp><speaker>Gost.</speaker>
   Può far mio padre, ch'io habbia d'andar cosi dinanzi
    alla mia innamorta.
  </sp>
  <sp><speaker>Quer.</speaker>
   Che credete che sia? Quando le donne uengon lisciate
    dinanzia lorguasti 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.
  <sp><speaker>Gost.</speaker>
   Hor su che ho afar? di uia.
  </sp>
  <sp><speaker>Ouer.</speaker>
   La prima cosa, perche uoi non siate conosciuto
    bisognache contrasacciate la uoce a guisa di guesti
    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.
  </sp>
   [...and the rest of the scene]
[...and the other scenes in the act]
</div>
```

```
<div class="act">
 <h1>Atto Terzo</h1>
  <div class="scene">
 <h2>Scena Prima</h2>
  Querciuola, Costanzo, Brigida
         <br/>del Capitano
  Quer.
  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 .
  Gost. 
  Può far mio padre, ch'io habbia d'andar così dinanzi
    alla mia innamorta.
  Ouer.
  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.
 Gost. 
  Hor su che ho afar? di uia.
  Quer.
  ∠p>La prima cosa, perche uoi non siate conosciuto
    bisognache contrasacciate la uoce a guisa di guesti
    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.
   [...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



```
<div type="act">
 <head>Atto Terzo</head>)
  <div type="scene">
 <head>Scena Prima</head>)
  <stage>Querciuola, Costanzo, Brigida
          <lb/>del Capitano</stage>
  <sp><speaker>Ouer.</speaker>
   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/>.
  </sp>
   <sp><speaker>Gost.</speaker>
   Può far mio padre, ch'io habbia d'andar cosi dinanzi
    alla mia innamorta. 
  <sp><speaker>Quer.</speaker>
   Che credete che sia? Quando le donne uengon lisciate
    dinanzia lorguasti 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.
  </sp>
   <sp><speaker>Gost.</speaker>
   Hor su che ho afar? di uia.
  </sp>
   <sp><speaker>Ouer.</speaker>
   La prima cosa, perche uoi non siate conosciuto
    bisognache contrasacciate la uoce a guisa di guesti
    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. 
  </sp>
   [...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 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/

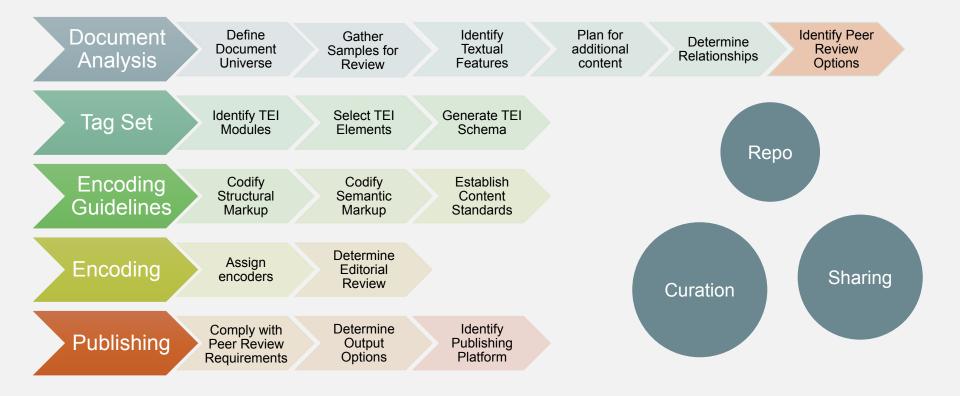
# PART 2

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

#### **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.

# XML KEY TERMS

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

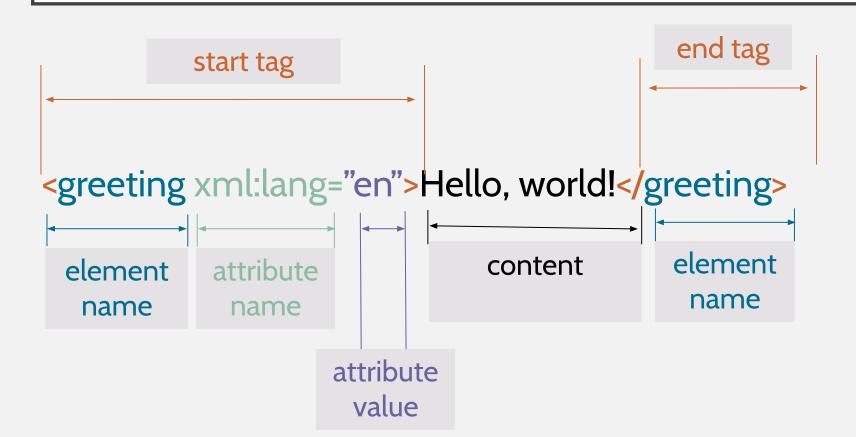
- <title>The Odyssey</title>

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

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

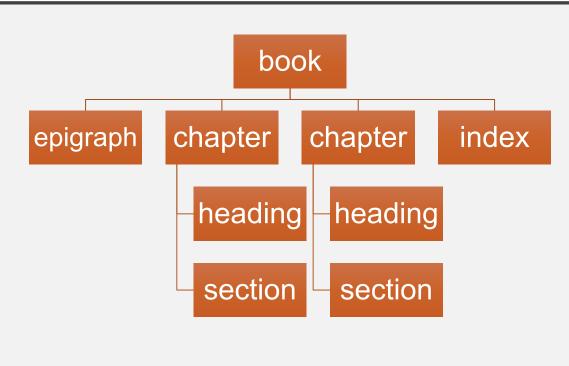
**DTDs** (Document Type Definitions) and **Schemas** define the rules that govern a particular type of XML document. They declare elements and attributes and the allowable content for those elements and attributes (grammar rules).

# 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**

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></epigraph>	Before Part I
toc reveals novel in 5 parts/sections	<div type="part"></div>	The Wild Land, Neighboring Fields, Winter Memories, etc.
paragraphs		Format body of text

# **TEI P5 GUIDELINES**

Encoding Guidelines

Codify Structural Markup Codify Semantic Markup Establish Content Standards

# **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 < encoding Desc >

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 < revision Desc>

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
        </titleStmt>
        <publicationStmt>
           Insert a paragraph statement or additional details
              about publisher and availability
        </publicationStmt>
        <sourceDesc>
           Insert a paragraph statement or additional
              bibliographic details
        </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 lewell</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>
    Charles hadn't visited the manor house since
       Easter, 1955, and now he remembered why. 
    <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>
</div>
```

# Break: 30 Minutes!

#### **ACTIVITY 4: 30 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?

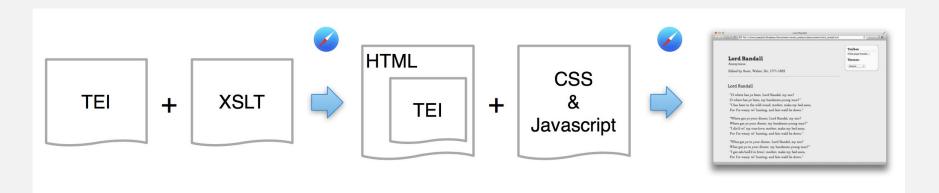
# PUBLISHING SYSTEMS & PLATFORMS

**Publishing** 

Comply with Peer Review Requirements Determine Output Options Identify Publishing Platform

- <u>eXtensible Text Framework</u> (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 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/2XjFsUG

### 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/0017.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?