

Introduction to Text Encoding

Dipa Desai and Emily Sullivan

HIST 7251

Professor Jessica Parr

Spring 2024



Northeastern University
NULab for Texts, Maps, and Networks

*Feel free to ask questions at any point
during the presentation!*

Workshop Agenda

- Metadata Concept Review
- Overview of TEI Guidelines + Encoding Structures
 - Navigating the TEI Guidelines
- Text Encoding Project Examples
- Class Activity: Discussion of <teiHeader>
- TEI + Metadata Mapping

Find all of the course materials at:

<https://bit.ly/sp24-parr-hist7251-tei>



Oxygen XML Editor

You can open and examine XML files in the text editor of your choice (Atom, Visual Code Editor, etc.) but an excellent platform for text encoding is Oxygen XML Editor.

- To download, visit:
https://www.oxygenxml.com/xml_editor/download_oxygenxml_editor.html?os
- Select version 26. Choose the option that matches your computer and follow the steps to download and install Oxygen.
- When you are prompted, paste the license key for Northeastern from the email circulated before class.



Review: Metadata Concepts



Northeastern University
NULab for Texts, Maps, and Networks

*Feel free to ask questions at any point
during the presentation!*

Types of Metadata

Dublin Core provides descriptive metadata.

DACS is a content standard. Content standards are guides to structure textual values in metadata.

Today we will talk about **TEI**, an example of a markup language that provides structural metadata.

Metadata Type	Example Properties	Primary Uses
Descriptive metadata	Title Author Subject Genre Publication date	Discovery Display Interoperability
Technical metadata	File type File size Creation date/time Compression scheme	Interoperability Digital object management Preservation
Preservation metadata	Checksum Preservation event	Interoperability Digital object management Preservation
Rights metadata	Copyright status License terms Rights holder	Interoperability Digital object management
Structural metadata	Sequence Place in hierarchy	Navigation
Markup languages	Paragraph Heading List Name Date	Navigation Interoperability

Source: [NISO's Metadata Primer](#)



Types of Metadata Standards

There are four central kinds of metadata standards based on what they are regularizing and describing for consistency:

- **Structure standards:** sets of metadata elements defined for a structural purpose (known as schemes or schemas).
- **Content standards:** rules for the input data in elements that describe textual data.
- **Value standards:** types of standards that restrain or narrow the possibilities of input for terms to reduce variation (known as controlled vocabularies).
- **Format standards:** technical specifications for how to encode metadata for machine readability and processing (known as data formats or encoding standards).



Key Metadata Vocabulary

- **Namespace:** a descriptor for a **container** of information, used in reference with metadata to indicate the specific **vocabulary** a file structure or data system is using and can stand in for a **URI** or **IRI** (uniform resource identifier or internationalized resource identifier).
- **Schema:** the organization or structure for data, a dataset, or database that acts as a model or representation of information.
- **eXtensible Markup Language (XML):** a markup language and file format for storing, transmitting, and reconstruing data to be human and machine readable.



Levels of Archival Description

Item: the smallest intellectually indivisible archival unit.

File Unit: an organized suite of items grouped together for use or archival arrangement.

Series: file units and/or items grouped together based on similar filing, function, activity, form, or some other relationship that ties them together.

Record Groups or Collections: archival documents grouped together based on similar provenance, acquisition, organizational content, form, or name of collector.

See the [National Archives' Archival Materials and Related Elements](#) for more information.



Overview: XML, TEI, and <teiHeader>



Northeastern University
NULab for Texts, Maps, and Networks

*Feel free to ask questions at any point
during the presentation!*

TEI History & Guidelines



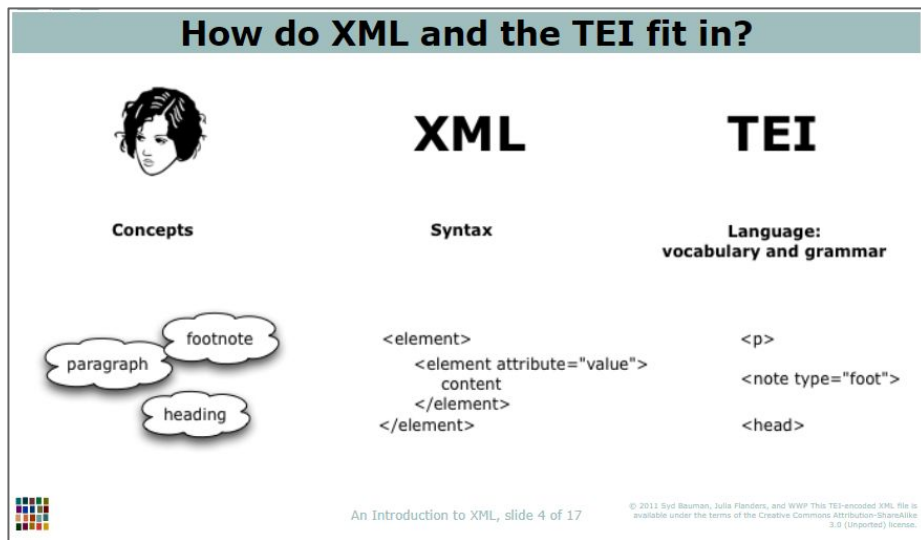
TEI (Text Encoding Initiative) was developed in the late 1980s to address concerns for digital document longevity, interoperability, and re-use.

TEI uses the XML markup language to encode meanings and structures of text. This allows documents to be read by humans and machines and it increases access to digitized text.

Let's look at the [TEI Guidelines](#).



How are XML and TEI related?



XML (eXtensible Markup Language) is a markup language and file format for storing, transmitting, and reconstruing data to be human and machine readable. It is composed of nested *elements*, a container for a type of information identified by the element name.

TEI is a language consisting of guidelines for representing texts in digital form through schemas.

TEI is highly customizable and projects can create their own elements with a customized schema.

[Introduction to XML](#), Julia Flanders and Syd Bauman (2018)



XML Elements

XML documents are associated with *schemas*, sets of rules about the structure of an XML document consisting of rules and constraints.

XML documents are structured around *elements* that each have a beginning and end tag that “wraps” around the information it contains and describes.

[next](#) [prev](#) [first](#)

XML Elements

Text is divided into *elements* (the “nouns” of the encoding — *content objects*).

- elements by *start-tags* and *end-tags*
`<heading>Wines</heading>`
- start-tags by `< ... >`
`<heading>`
- end-tags by `</ ... >`
`</heading>`
- special case: short-hand for an element with no content
`<anchor/>` = `<anchor></anchor>`

element (“the name element”)

`<name>Matthew P. Damon</name>`

start-tag content end-tag

[Introduction to XML](#), Julia Flanders and Syd Bauman (2018)



XML Attributes

next prev first

XML Attributes

attribute ("the type attribute")

```
<name type="person">Matt Damon</name>
```

attribute name attribute value

- any number of attributes can be specified on a given start- (or empty-) tag
- but **only one** with a given name!
- order does not matter
- whitespace can be adjusted to make it look good to humans

Sometimes it is necessary to encode additional information **elements** themselves which is done using *attributes*.

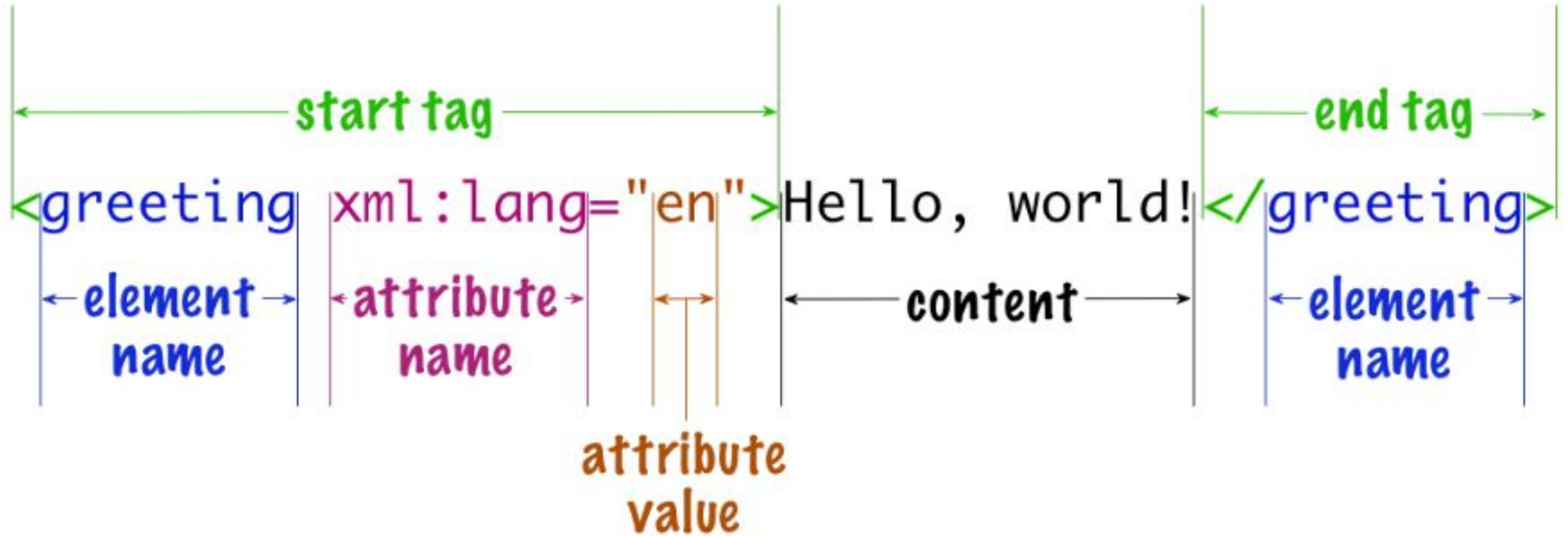
Attributes (often identified with the @ symbol in prose) offer more information about an element, and are contained in the start tag:

< element attribute="value">

[Introduction to XML](#), Julia Flanders and Syd Bauman (2018)



Anatomy of an Element



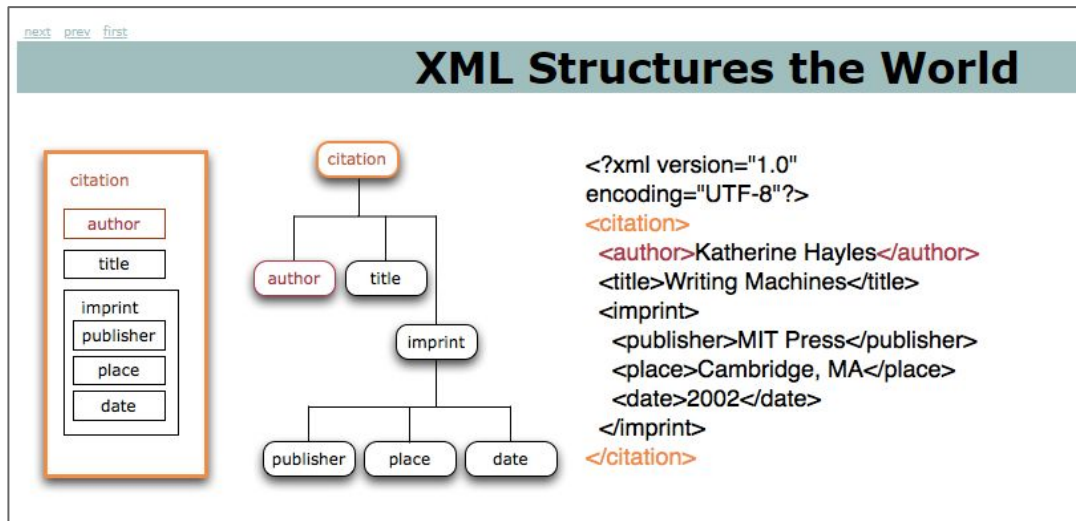
[Introduction to XML](#), Julia Flanders and Syd Bauman (2018)



Nested Structure of TEI

XML functions with nested elements under a *root* element. For TEI, this is the <TEI> element.

Elements nested under an element are called *child* elements to the *parent* element.



[Introduction to XML](#), Julia Flanders and Syd Bauman (2018)

The hierarchical relationships for elements are represented by what an element “**May contain**” or is “**Contained by**” in the TEI Guidelines.



TEI Infrastructure

Module name	Formal public identifier	Where defined
analysis	Analysis and Interpretation	17 Simple Analytic Mechanisms
certainty	Certainty and Uncertainty	21 Certainty, Precision, and Responsibility
core	Common Core	3 Elements Available in All TEI Documents
corpus	Metadata for Language Corpora	15 Language Corpora
dictionaries	Print Dictionaries	9 Dictionaries
drama	Performance Texts	7 Performance Texts
figures	Tables, Formulae, Figures	14 Tables, Formulae, Graphics, and Notated Music
gaiji	Character and Glyph Documentation	5 Characters, Glyphs, and Writing Modes
header	Common Metadata	2 The TEI Header
iso-fs	Feature Structures	18 Feature Structures
linking	Linking, Segmentation, and Alignment	16 Linking, Segmentation, and Alignment
msdescription	Manuscript Description	10 Manuscript Description
namesdates	Names, Dates, People, and Places	13 Names, Dates, People, and Places
nets	Graphs, Networks, and Trees	19 Graphs, Networks, and Trees
spoken	Transcribed Speech	8 Transcriptions of Speech
tagdocs	Documentation Elements	22 Documentation Elements
tei	TEI Infrastructure	1 The TEI Infrastructure
textcrit	Text Criticism	12 Critical Apparatus
textstructure	Default Text Structure	4 Default Text Structure
transcr	Transcription of Primary Sources	11 Representation of Primary Sources
verse	Verse	6 Verse

The TEI encoding schema consists of *modules*, sets of particular XML elements and attributes.

Each element has assigned *classes*, sets of elements grouped together by attributes and model (or element location).

A TEI *schema* can be composed of any number of modules, but always should include: **tei**, **core**, **header**, and **textstructure**.

All XML documents require a defined or associated schema, which in TEI is easily customizable for project purposes.



TEI Elements



Northeastern University
NULab for Texts, Maps, and Networks

*Feel free to ask questions at any point
during the presentation!*

<teiHeader>

The <teiHeader> has five major child elements including:

- <fileDesc> includes a full bibliographic description of the electronic TEI/XML file. This element is required in the <teiHeader>.
- <encodingDesc> documents the relationship between the electronic text (the TEI/XML file) and the source(s) from which it was derived
- <profileDesc> provides a description of non-bibliographic aspects of a text (languages, subjects, etc)
- <xenoData> provides a container element for non-TEI metadata
- <revisionDesc> is the changelog of the file summarizing major revisions



<teiHeader>

<TEI> root element

File metadata in <teiHeader>

File description in <fileDesc>

Encoding description in
<encodingDesc>

Non-bibliographic information
in <profileDesc>

File changes in <revisionDesc>

```
TEI.2 text front div1 p
1 <!DOCTYPE TEI.2 SYSTEM "http://docsouth.unc.edu/dtds/teixlite.dtd">
2 <TEI.2>
3 <teiHeader type="text" status="new">
4 <fileDesc>
5 <titleStmt> [21 lines]
27 <editionStmt> [3 lines]
31 <extent>ca. 550K</extent>
32 <publicationStmt> [9 lines]
42 <sourceDesc default="NO"> [19 lines]
62 </fileDesc>
63 <encodingDesc>
64 <projectDesc default="NO"> [4 lines]
69 <editorialDecl default="NO"> [14 lines]
84 <classDecl> [6 lines]
91 </encodingDesc>
92 <profileDesc>
93 <langUsage default="NO"> [2 lines]
96 <textClass default="NO"> [13 lines]
110 </profileDesc>
111 <revisionDesc>
112 <change> [7 lines]
120 <change> [7 lines]
128 <change> [7 lines]
136 <change> [7 lines]
144 </revisionDesc>
145 </teiHeader>
146 <text>
```

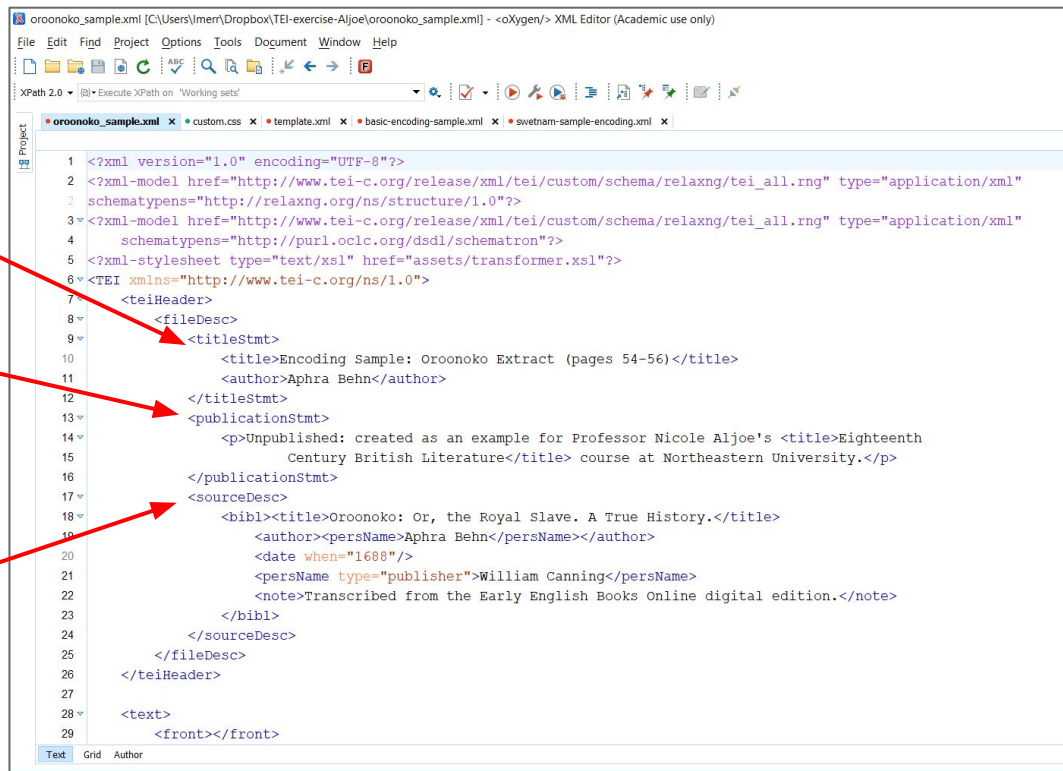


<fileDesc>

encoded document title
statement in <titleStmt>

publication statement
about encoded document
<publicationStmt>

bibliographic information
on source document
<sourceDesc>



```
1 <?xml version="1.0" encoding="UTF-8"?>
2 <?xml-model href="http://www.tei-c.org/release/xml/tei/custom/schema/relaxng/tei_all.rng" type="application/xml"
3 schematypens="http://relaxng.org/ns/structure/1.0"?>
4 <?xml-model href="http://www.tei-c.org/release/xml/tei/custom/schema/relaxng/tei_all.rng" type="application/xml"
5 schematypens="http://purl.oclc.org/dsdl/schematron"?>
6 <?xml-stylesheet type="text/xsl" href="assets/transformer.xsl"?>
7 <TEI xmlns="http://www.tei-c.org/ns/1.0">
8   <teiHeader>
9     <fileDesc>
10       <titleStmt>
11         <title>Encoding Sample: Oroonoko Extract (pages 54-56)</title>
12         <author>Aphra Behn</author>
13       </titleStmt>
14       <publicationStmt>
15         <p>Unpublished: created as an example for Professor Nicole Aljoe's <title>Eighteenth
16           Century British Literature</title> course at Northeastern University.</p>
17       </publicationStmt>
18       <sourceDesc>
19         <bibl><title>Oroonoko: Or, the Royal Slave. A True History.</title>
20           <author><persName>Aphra Behn</persName></author>
21           <date when="1688"/>
22           <persName type="publisher">William Canning</persName>
23           <note>Transcribed from the Early English Books Online digital edition.</note>
24         </bibl>
25       </sourceDesc>
26     </fileDesc>
27   </teiHeader>
28   <text>
29     <front></front>
```



<titleStmt>

Element that groups information about the title of a work and those responsible for its content.

<title>

<author>

<sponsor> or <funder>

information for an organization or individual

```
TEI  teiHeader  fileDesc  titleStmt
1  <?xml version="1.0" encoding="UTF-8"?>
2  <?xml-model href="http://www.w3.org/2001/XMLSchema" type="application/xml" schematypens="http://relaxng.org/ns/structure/1.0"?>
3  <?xml-model href="http://www.w3.org/2001/XMLSchema" type="application/xml" schematypens="http://purl.oclc.org/dsdl/schematron"?>
4  <!-- $Id: jones-m.misellanies.xml 41384 2020-07-16 19:34:08Z syd $ -->
5  <TEI xmlns="http://www.w3.org/2001/XMLSchema" xmlns:xi="http://www.w3.org/2001/XInclude"
6    xml:lang="en">
7    <teiHeader xml:id="TR00146.hdr">
8      <fileDesc>
9        <titleStmt>
10         <title type="main">Miscellanies in Prose and Verse, 1750</title>
11         <author>
12           <persName type="person-female" ref="p:mjones.isr">Jones, Mary</persName>
13         </author>
14         <sponsor>Brown University</sponsor>
15         <sponsor>Northeastern University</sponsor>
16         <funder>U.S. National Endowment for the Humanities</funder>
17       </titleStmt>
18       <xi:include href="http://www.w3.org/2001/XInclude" xpointer="element(WWPedition)"/>
19     <publicationStmt>
20       <publisher>Northeastern University Women Writers Project</publisher>
21     </address>
22       <addrLine>SL 213</addrLine>
23       <addrLine>Northeastern University</addrLine>
24       <addrLine>360 Huntington Avenue</addrLine>
25       <addrLine>Boston, MA 02115-5005</addrLine>
26       <addrLine>USA</addrLine>
27       <addrLine>url:mailto:wwp@neu.edu</addrLine>
28       <addrLine>url:http://www.wwp.northeastern.edu/</addrLine>
29     </address>
30     <idno type="WWP">TR00146</idno>
31     <idno type="URL">https://www.wwp.northeastern.edu/texts/jones-m.misellanies.html</idno>
32     <xi:include href="http://www.w3.org/2001/XInclude" xpointer="element(WWPavailability)"/>
33     <date when="2007-07-31"/>
34   </publicationStmt>
```



<publicationStmt>

Element that groups information about the publication or distribution of an *electronic* text. May include an availability statement regarding usage or rights (including licenses).

<publisher>

<address>

<idno> includes an identifier number of an item, person, etc.

```
TEI  teiHeader  fileDesc  titleStmt
1  <?xml version="1.0" encoding="UTF-8"?>
2  <?xml-model href="http://www.w3.org/2001/XMLSchema-instance" type="application/xml" schematypens="http://relaxng.org/ns/structure/1.0"?>
3  <?xml-model href="http://www.w3.org/2001/XMLSchema-instance" type="application/xml" schematypens="http://purl.oclc.org/dsdl/schematron"?>
4  <!-- $Id: jones-m.miscellanies.xml 41384 2020-07-16 19:34:08Z syd $ -->
5  <TEI xmlns="http://www.w3.org/2001/XMLSchema-instance" xmlns:xsi="http://www.w3.org/2001/XInclude"
6    xml:lang="en">
7    <teiHeader xml:id="TR00146.hdr">
8      <fileDesc>
9        <titleStmt>
10         <title type="main">Miscellanies in Prose and Verse, 1750</title>
11         <author>
12           <persName type="person-female" ref="p:mjones.isr">Jones, Mary</persName>
13         </author>
14         <sponsor>Brown University</sponsor>
15         <sponsor>Northeastern University</sponsor>
16         <funder>U.S. National Endowment for the Humanities</funder>
17       </titleStmt>
18       <xsi:include href="http://www.w3.org/2001/XInclude" xpointer="element(WWPedition)"/>
19     </fileDesc>
20     <publicationStmt>
21       <publisher>Northeastern University Women Writers Project</publisher>
22       <address>
23         <addrLine>SL 213</addrLine>
24         <addrLine>Northeastern University</addrLine>
25         <addrLine>360 Huntington Avenue</addrLine>
26         <addrLine>Boston, MA 02115-5005</addrLine>
27         <addrLine>USA</addrLine>
28         <addrLine>url:mailto:wwp@neu.edu</addrLine>
29         <addrLine>url:http://www.wwp.northeastern.edu/</addrLine>
30       </address>
31       <idno type="WWP">TR00146</idno>
32       <idno type="URL">https://www.wwp.northeastern.edu/texts/jones-m.miscellanies.html</idno>
33       <xsi:include href="http://www.w3.org/2001/XInclude" xpointer="element(WWPavailability)"/>
34       <date when="2007-07-31"/>
35     </publicationStmt>
36  </TEI>
```



<sourceDesc>

Element that groups information about the source(s) from which an electronic text is derived or generated, usually with bibliographic information.

<biblStruct> is a structured format for bibliographic info

<monogr> is a structured element containing bibliographic info for an item (book or journal) published as an independent item

```
TEI teiHeader fileDesc titleStmt
34      </publicationStmt>
35      <sourceDesc n="OT00146">
36        <biblStruct>
37          <monogr>
38            <author>
39              <persName ref="p:mjones.isr" type="titlePage">Mary Jones</persName>
40              <persName ref="p:mjones.isr" type="regularized">Jones, Mary</persName>
41            </author>
42            <title>Miscellanies in prose and verse</title>
43            <edition>First edition</edition>
44            <idno type="nnc">824.J711</idno>
45            <imprint>
46              <pubPlace>Oxford</pubPlace>
47              <publisher>
48                <persName type="titlePage" ref="p:jdodsley.uwq">Mr. Dodsley</persName><persName
49                  type="regularized" ref="p:jdodsley.uwq">Dodsley, James</persName>, <persName
50                  type="titlePage" ref="p:clements.cak">Mr. Clements</persName><persName
51                  type="regularized" ref="p:clements.cak">Clements, Richard</persName>, and
52                  <persName type="titlePage" ref="p:wfrederic.jvl">Mr. Frederick</persName><persName
53                  type="regularized" ref="p:wfrederic.jvl">Frederick, William</persName>
54              </publisher>
55              <date when="1750">MDCCL</date>
56            </imprint>
57            <extent>
58              <measure unit="page" quantity="460"/>
59            </extent>
60            <extent>
61              <dimensions>
62                <format>octavo</format>
63              </dimensions>
64            </extent>
65          </monogr>
66        </biblStruct>
67      </sourceDesc>
68    </fileDesc>
```



<encodingDesc>

Element that documents the relationship between the electronic text and its source(s).

<projectDesc> describes the aim or purpose of the encoding project

<editorialDecl> provides details about the editorial principles and practices applied during the encoding of a text/texts.

```
63 <encodingDesc>
64   <projectDesc>
65     <p>This text was created as part of a project by the Committee on Institutional
66       Cooperation. Project description and participants are available at the project
67       website at http://www.lettrs.indiana.edu/wright.</p>
68   </projectDesc>
69   <editorialDecl>
70     <p>This electronic text file was created by Optical Character Recognition (OCR), and
71       has been encoding and edited using the recommendations for Level 4 of the TEI in
72       Libraries Guidelines. Digital page images are linked to the text file.</p>
73   </editorialDecl>
74 </encodingDesc>
75 <profileDesc>
76   <langUsage>
77     <language ident="fra">French</language>
78     <language ident="lat">Latin</language>
79   </langUsage>
80 </profileDesc>
81 <revisionDesc>
82   <change who="AEL Data" when="2002-12-03">Finished transcription and SGML encoding</change>
83   <change who="Perry Willett, Indiana University" when="2002-12-04">Finished TEI conversion and
84   <change who="Margaret Hermes, Indiana University" when="2003-01-30">Finished final proofreading
85   <change who="Siobhain Rivera" when="2014-08-04">Converted document from SGML to TEI P5.</change>
86   <change who="Siobhain Rivera" when="2014-08-04">Edited encoding to fix Table of Contents and t
87 </revisionDesc>
88 </teiHeader>
```



<profileDesc>

Element that groups non-bibliographic information about a text.

<language> identifies languages used or contained in a text

<textClass> groups information about the topic of a text using a standard classification schema or thesaurus (LCSH)

<keywords> describe topics in a text with an attribute @scheme to identify the classification source

```
86      </encodingDesc>
87      <profileDesc>
88        <langUsage>
89          <language id="fr">French</language>
90        </langUsage>
91        <textClass>
92          <keywords scheme="lcsch">
93            <list type="simple">
94              <item>Douglass, Frederick, 1818-1895.</item>
95              <item>African Americans -- Maryland -- Biography.</item>
96              <item>African American abolitionists -- Biography.</item>
97              <item>Abolitionists -- United States -- Biography.</item>
98              <item>Slaves -- Maryland -- Biography.</item>
99              <item>Fugitive slaves -- Maryland -- Biography.</item>
100             <item>Slavery -- Maryland -- History -- 19th century.</item>
101             <item>Slavery -- United States -- History -- 19th century.</item>
102             <item>Plantation life -- Maryland -- History -- 19th century.</item>
103             <item>Slaves -- Maryland -- Social conditions -- 19th century.</item>
104             <item>Slaves' writings, American -- Maryland.</item>
105            </list>
106          </keywords>
107        </textClass>
108      </profileDesc>
```



<revisionDesc>

Element that records and summarizes the revision history of a file (like a changelog).

<change> a change or set of changes made to a document during its production or revision of the electronic file.

Uses **@who** and **@when** attributes to identify people involved and dates of changes.

```
63 <encodingDesc>
64   <projectDesc>
65     <p>This text was created as part of a project by the Committee on Institutional
66       Cooperation. Project description and participants are available at the project
67       website at http://www.letrs.indiana.edu/wright.</p>
68   </projectDesc>
69   <editorialDecl>
70     <p>This electronic text file was created by Optical Character Recognition (OCR), and
71       has been encoding and edited using the recommendations for Level 4 of the TEI in
72       Libraries Guidelines. Digital page images are linked to the text file.</p>
73   </editorialDecl>
74 </encodingDesc>
75 <profileDesc>
76   <langUsage>
77     <language ident="fra">French</language>
78     <language ident="lat">Latin</language>
79   </langUsage>
80 </profileDesc>
81 <revisionDesc>
82   <change who="AEL Data" when="2002-12-03">Finished transcription and SGML encoding</change>
83   <change who="Perry Willett, Indiana University" when="2002-12-04">Finished TEI conversion and final editing</change>
84   <change who="Margaret Hermes, Indiana University" when="2003-01-30">Finished final proofreading</change>
85   <change who="Siobhain Rivera" when="2014-08-04">Converted document from SGML to TEI P5.</change>
86   <change who="Siobhain Rivera" when="2014-08-04">Edited encoding to fix Table of Contents and title page.</change>
87 </revisionDesc>
88 </teiHeader>
```



<text>

The <text> contains a unitary or composite single text. It is outside of the <teiHeader>.

- <front> front matter (preface, etc.)
- <body> body of the text (chapters, etc.)
- <back> back matter (appendices, etc.)
- <div> marks sections or divisions in a text (chapters, letters, etc.)
- <p> marks paragraphs of prose
- <persName> is the element used to mark a person's name
- <placeName> is an element used to mark names of locations



<text>

front matter (preface,
etc.) in <front>

body of the text (chapters,
etc.) in <body>

back matter (appendices,
etc.) in <back>

```
template.xml [C:\Users\lmerri\Dropbox\TEI-exercise-Aljoe\template.xml] - <Oxygen/> XML Editor (Academic use only)
File Edit Find Project Options Tools Document Window Help
XPath 2.0 Execute XPath on "Working sets"
Project
TEI
28 <text>
29 <front>
30 <div type="letter" style="font-style:italic">
31 <opener style="font-style:normal">
32 <salute>To the
33 Right Honourable
34 the
35 <persName style="font-style:italic"><hi style="font-style:normal">Lord</hi> Maitland</persName></salute>.
36 <salute>My Lord,</salute>
37 </opener>
38 <p>
39 <!-- If you are encoding one of the passages from the dedication, delete this comment and begin transcribing y
40 you need for each paragraph in your section. -->
41 </p>
42 <closer style="font-style:normal">
43 <salute><hi style="font-style:italic">My Lord,</hi>
44 Your Lordship's moft oblig'd
45 and obedient Servant,</salute>
46 <signed><persName style="font-style:italic">A. Behn.</persName></signed>
47 </closer>
48 </div>
49 </front>
50 <body>
51 <p><!-- If you are encoding a passage from the body of the document, delete this comment and begin transcribi
52 as you need for each new paragraph in your section. -->
53 </p>
54 </body>
55 <back>
56 <div type="editorial">
57 <interpGrp>
58 <interp xml:id="name_of_your_interpretation">Definition of your interpretation</interp> <!-- For each of t
59 on (make sure you have this correct!) as a value on the @xml:id attribute, and write a quick gloss so you remember whi
60 want to associate with your interpretations, and @ana to point to the relevant <interp>. -->
61 </interpGrp>
62 </div>
63 </back>
Text Grid Author
```



<div> and <p>

<div> marks sections or divisions in a text (chapters, letters, etc.)

<p> marks paragraphs of prose

Specify different types of **<div>** with the **@type** attribute (i.e. type="editorial")

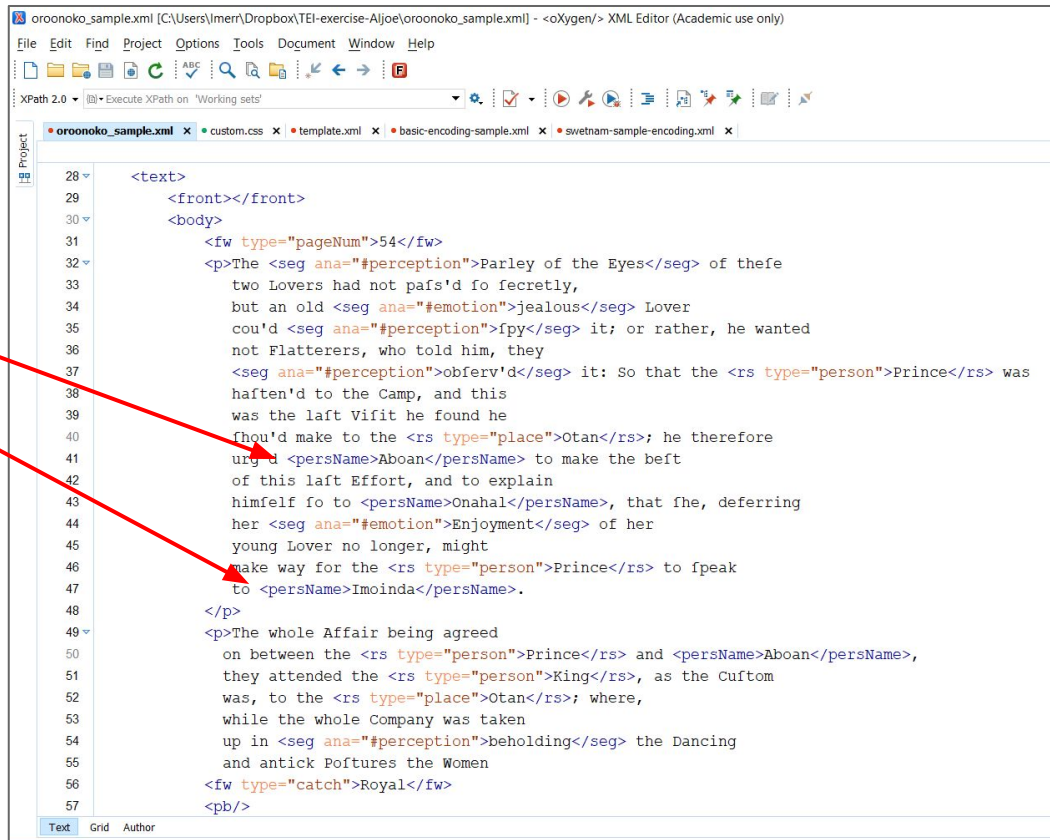
```
template.xml [C:\Users\lmerr\Dropbox\TEI-exercise-Aljoe\template.xml] - <oXygen/> XML Editor (Academic use only)
File Edit Find Project Options Tools Document Window Help
XPath 2.0 Execute XPath on 'Working sets'
oroonoko_sample.xml x custom.css x template.xml x
TEI
28 <text>
29 <front>
30 <div type="letter" style="font-style:italic">
31 <opener style="font-style:normal">
32 <salute>To the
33 Right Honourable
34 the
35 <persName style="font-style:italic"><hi style="font-style:normal">Lord</hi> Maitland</persName></salute>.
36 <salute>My Lord,</salute>
37 </opener>
38 <p>
39 <!-- If you are encoding one of the passages from the dedication, delete this comment and begin transcribing y
40 you need for each paragraph in your section. -->
41 </p>
42 <closer style="font-style:normal">
43 <salute><hi style="font-style:italic">My Lord,</hi>
44 Your Lordship's moft oblig'd
45 and obedient Servant,</salute>
46 <signed><persName style="font-style:italic">A. Behn.</persName></signed>
47 </closer>
48 </div>
49 </front>
50 <body>
51 <p><!-- If you are encoding a passage from the body of the document, delete this comment and begin transcribi
52 as you need for each new paragraph in your section. -->
53 </p>
54 </body>
55 <back>
56 <div type="editorial">
57 <interpGrp>
58 <interp xml:id="name_of_your_interpretation">Definition of your interpretation</interp> <!-- For each of t
59 on (make sure you have this correct!) as a value on the @xml:id attribute, and write a quick gloss so you remember whi
60 want to associate with your interpretations, and $ana to point to the relevant <interp>. -->
61 </interp>
62 </interpGrp>
63 </div>
64 </back>
Text Grid Author
```



Names

<persName> is the element used to mark a person's name

<placeName> is an element used to mark names of locations



```
28 <text>
29 <front></front>
30 <body>
31 <fw type="pageNum">54</fw>
32 <p>The <seg ana="#perception">Parley of the Eyes</seg> of thefe
33 two Lovers had not pafs'd fo fecretly,
34 but an old <seg ana="#emotion">jealous</seg> Lover
35 cou'd <seg ana="#perception">fpy</seg> it; or rather, he wanted
36 not Flatterers, who told him, they
37 <seg ana="#perception">obferv'd</seg> it: So that the <rs type="person">Prince</rs> was
38 haften'd to the Camp, and this
39 was the laft Vifit he found he
40 fhould make to the <rs type="place">Otan</rs>; he therefore
41 urg'd <persName>Aboan</persName> to make the beft
42 of this laft Effort, and to explain
43 himfelf fo to <persName>Onahal</persName>, that fhe, deferring
44 her <seg ana="#emotion">Enjoyment</seg> of her
45 young Lover no longer, might
46 make way for the <rs type="person">Prince</rs> to fpeak
47 to <persName>Imoinda</persName>.
48 </p>
49 <p>The whole Affair being agreed
50 on between the <rs type="person">Prince</rs> and <persName>Aboan</persName>,
51 they attended the <rs type="person">King</rs>, as the Cuftom
52 was, to the <rs type="place">Otan</rs>; where,
53 while the whole Company was taken
54 up in <seg ana="#perception">beholding</seg> the Dancing
55 and antick Poftures the Women
56 <fw type="catch">Royal</fw>
57 <pb/>
```



Activity: TEI Headers



Northeastern University
NULab for Texts, Maps, and Networks

*Feel free to ask questions at any point
during the presentation!*

Text Encoding Projects

[Folgers Shakespeare Library Digital Texts](#): a corpus of Shakespeare's plays, writings, and poetry (HTML, XML, TEI Simple, TXT) completed in 2010 and published online in 2012 by the Folger Shakespeare Library.

[Women Writers Online \(WWO\)](#): a collection of early women's writing in English published by the Women Writers Project at Northeastern with full text transcriptions of texts from 1526–1850.



Text Encoding Projects

[Documenting the American South: North American Slave Narratives](#) a digital collection of autobiographical narratives of self-emancipated and formerly enslaved people published in English up to 1920 hosted by the University of North Carolina at Chapel Hill.

[Wright American Fiction](#) a collection of 2,887 digital texts from influential American authors (including Harriet Beecher Stowe, Mark Twain, and Herman Melville) by the Indiana University Digital Library Program.



Downloading the Example Texts

Individual XML documents from the four different projects can be downloaded from a Google Drive folder at:

<https://bit.ly/HIST7250-XMLTexts>

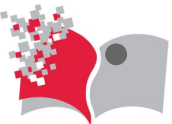
You can preview the files online but we recommend you download the entire folder and open the files in the Oxygen XML Editor.



Oxygen XML Editor

You can open and examine XML files in the text editor of your choice (Atom, Visual Code Editor, etc.) but an excellent platform for text encoding is Oxygen XML Editor.

- To download, visit:
https://www.oxygenxml.com/xml_editor/download_oxygenxml_editor.html?os
- Select version 26. Choose the option that matches your computer and follow the steps to download and install Oxygen.
- When you are prompted, paste the license key for Northeastern from the email circulated before class.



Group Activity: Exploring <teiHeader>

In pairs or small groups, choose two example XML documents from different projects. All files are available on Google Drive at:

<https://bit.ly/HIST7250-XMLTexts>

Look at the overall structure of the document and pay attention to the metadata in the **<teiHeader>** element.

Consider the following:

- Are there any elements that are similar to metadata fields from DACS or Dublin Core?
- Which TEI elements do you notice in the <teiHeader>?
- TEI allows projects to develop specific encoding guidelines and schema for projects. Look at the <encodingDesc> for examples.



TEI + Metadata Mapping



Northeastern University
NULab for Texts, Maps, and Networks

*Feel free to ask questions at any point
during the presentation!*

How do you structure a group of texts in TEI?

As we explored in Dublin Core and DACS, metadata can describe more than one item and often does for archival collections. How is this reflected in the TEI?

- <group> is an element that groups together a sequence of distinct texts that are regarded as a unit for some purpose
- <div> with the **@type** attribute that pairs an element for a division in the text with an attribute to characterize the part.



TEI to DACS: Primary Source Cooperative

[Primary Source Cooperative](#): a digital scholarship project between the **Massachusetts Historical Society** (MHS) and the **Digital Scholarship Group** (DSG) at Northeastern to create digital, documentary editions of primary sources from the nineteenth century.


[Coop Home](#)[Who We Are](#)

Primary Source Cooperative

Bringing American history to life with primary sources.


The purpose of the Primary Source Cooperative at the Massachusetts Historical Society (the Cooperative, MHS) is to publish online the work of editors who are preparing the content of archival and other manuscript records for scholarly and public access.

Search across the editions:




Roger Brooke Taney Papers

The Papers of Roger Brooke Taney, a project based at the University of West Florida in Pensacola, will digitally publish annotated transcriptions of Taney's papers (correspondence, legal documents, etc.). Each online volume will capture a separate aspect of Taney's life and career, including his tenure as Chief Justice of the United States.



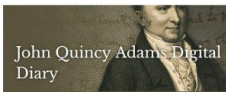
Ellen Swallow Richards Digital Archive

The first woman to graduate from and then teach at MIT.



Catharine Maria Sedgwick Online Letters

During her lifetime, Catharine Maria Sedgwick (1789-1867) became known in the United States as the most significant, experimental, influential, and highly regarded woman writer in the Early National period of American literature. The Catharine Maria Sedgwick Online Letters project is a digital edition of her letters.



John Quincy Adams Digital Diary

One of America's great statesmen, John Quincy Adams' distinguished career in public service spanned six decades and included roles as diplomat, secretary of state, president, and congressman. The John Quincy Adams Digital Diary makes JQA's diary, which spans over 68 years, truly accessible for the first time. [Link to site](#)



Northeastern University
NULab for Texts, Maps, and Networks

Feel free to ask questions at any point during the presentation!

Example: John Quincy Adams Digital Diaries

John Quincy Adams
Digital Diary (JQA) uses
the `<div>` element with
the `@type` attribute for
entries. Each TEI
document contains
entries for one month.

```
<div type="entry" xml:id="jqadiaries-v24-1796-07-03">
  <head>3 July 1796</head>
  <bibl><author>JQA</author><date type="creation" when="1796-07-03"/><editor
    role="transcription">Neal Millikan</editor>
  <subject>Recreation</subject>
</bibl>
<div type="docbody">
  <p><date>3. </date>Third lesson of Italian. From the distribution of my time
    I find myself now always in a hurry. I shall be under the necessity of
    making some alteration.— Short walk before dinner. Long one in the
    Evening.</p>
</div>
</div>
<div type="entry" xml:id="jqadiaries-v24-1796-07-04">
  <head>4 July 1796</head>
  <bibl><author>JQA</author><date type="creation" when="1796-07-04"/><editor
    role="transcription">Neal Millikan</editor></bibl>
  <div type="docbody">
    <p><date>4. </date>Anniversary of Independence. <persRef ref="adams-thomas">
      >My brother</persRef> in celebration of the day devoted it to
      amusement.— I passed it as usual.— Called to see <persRef ref="u">Mll<hi
        rend="superscript">e:</hi> Lorenzi</persRef> who was not at home—
      Attended the debates in the National Assembly. Saw <persRef ref="u">
        >Reuterswerd</persRef> there.— Drew up a Note to present to the
        <persRef ref="hartog-paulus">President of the
        Assembly</persRef>.</p>
  </div>
</div>
```



PSC Document Workflow

WET Transcription

Primary source documents are transcribed in **Word Enhanced Templates (WET)** in Microsoft Word where metadata and semantic information is encoded with “markers.”

WETVAC

Completed WET transcription documents are fed into a **WETVAC**, a drag and drop interface (using JSON and XSLT) that transforms a WET files into XML documents.

XML Document

Transformed **XML** documents are saved and checked for consistency. The documents follow a **customized TEI schema** that is maintained and updated as needed.



PSC Document Workflow

WET Transcription

{{DATE}} 1896-04-10
{{AUTHOR}} richards-ellen
{{RECIPIENT}} atkinson-edward
{{HEAD}} Ellen Swallow Richards to Edward Atkinson
{{EDITOR}} SSS
{{EDITION}} Ellen Swallow Richards Digital Archive
{{TRANSCRIBER}} SSS
{{TRANSCRIPTION-DATE}}
{{SUBJECT}} Science of Nutrition
{{SUBJECT}} Temperance Women
{{SUBJECT}} Restaurants

{{DATELINE}} Boston, April 10 1896
{{SALUTE}} My dear Mr Atkinson
I have been reading {{P:nettle-unknown}}Prof. Nettle's{{ENDP}} article on my way in the car for these two or three days & have just finished it. Thank you for sending it. It is most suggestive & is in the line in which we all have been working. We have said to the Temperance women, do not grumble at the saloon until you put some soup &c. in its place.
It comes to this to teach the children what are good foods & how to cook them. {{P:abel-mary}}Mrs Abel{{ENDP}} & I opposed restaurants as much as you - & we have proved that if the people know how they can live more cheaply. The trouble is they do not want to know how & it must be the business of the state to teach them.
Yes, the bean lozenge has a future, and I presume it will be worked up
{{CLOSE}} Sincerely yours
{{SIGNED}} Ellen H. Richards

WETVAC



XML Document

```
primarysourcecoop_rev1.jasoch  X  primarysourcecoop_rev1.rng  X  WETVAC-DailyTestDoc-2021-07-09.xml  X
processing-instruction
1  <?xml version="1.0" encoding="UTF-8"?>
2  <?xml-model href="http://www.primarysourcecoop.org/publications/pub/schema/primarysourcecoop_rev1.rng" type="text/xml"?>
3  <TEI xmlns="http://www.tei-c.org/ns/1.0" xml:id="WETVAC-DailyTestDoc-AddDate">
4    <teiHeader>
5      <fileDesc>
6        <titleStmt>
7          <title>First Middle Last to Recipfirst Reciplast and Otherfirst Otherlast</title>
8        </titleStmt>
9        <resp><transcribed by /></resp>
10       <name>Transcriber Extraordinaire</name>
11       <note>Transcribed on <date type="transcription" when="2020-01-31"/></note>
12     </teiHeader>
13     <title></title>
14     <publicationInfo>
15       <publisher>Primary Source Cooperative</publisher>
16       <date>2021</date>
17     </publicationInfo>
18     <availability>
19       <p>Online version 1.0</p>
20       <license>Available under Creative Commons license CC BY-NC-SA
21       Attribution--NonCommercial--ShareAlike.</license>
22     </availability>
23     </publicationInfo>
24     <series></series>
25     <editor>Excellent Editor</editor>
26     </series>
27     <sourceDesc>
28       <list><list>
29         <item><date type="creation" when="1896-01-31"/></item></list></list>
30       <item><date type="creation" when="1896-01-31"/></item></list></list>
31     </sourceDesc>
32     <idno></idno>
33     <repository>Massachusetts Historical Society</repository>
34     <collection>Papers of Sontopia Farms</collection>
35     </idno>
36     </idno>
37     </idno>
```



Updating Outdated Metadata Structures

```
40 <text>
41 <body>
42 <div type="doc" xml:id="CMS1800-12-16-toTheodoreSedgwickIF-001">
43 <bibl>
44 <date type="creation" when="1800-12-16"/>
45 <author>Catharine Maria Sedgwick</author>
46 <recipient>TSI</recipient>
47 <head>Catharine Maria Sedgwick to Theodore Sedgwick I</head>
48 <editor>Patricia Kalayjian, Lucinda Damon Bach, Deborah Gussman</editor>
49 <edition>CMSOL</edition>
50 <name type="transcriber">Patricia Kalayjian</name>
51 <date type="transcription" when="2018-08-04"/>
52 <subject>Sedgwick Family Relations</subject>
53 <subject>Illness</subject>
54 <subject>Literature and History</subject>
55 <subject>Press</subject>
56 <subject>Death/Mourning</subject>
57 </bibl>
```

The legacy metadata structure had a **<bibl>** element in a **<div>** element with **@type** attribute for the entire document.

The schema needed to be updated to the general structure for the **<teiHeader>** on the left.

```
TEI text body
1 <?xml version="1.0" encoding="UTF-8"?>
2 <?xml-model href="http://www.primarysourcecoop.org/publications/pub/schema/primarysourcecoop.rng" type="
3 schematypens="http://relaxng.org/ns/structure/1.0"?>
4 <TEI xmlns="http://www.tei-c.org/ns/1.0" xmlns:xsl="http://www.w3.org/2001/XMLSchema"
5 xml:id="CMS1800-12-16-toTheodoreSedgwickIF">
6 <teiHeader>
7 <fileDesc>
8 <titleStmt>
9 <title/>
10 </titleStmt>
11 <editionStmt>
12 <edition>
13 <date/>
14 </edition>
15 </editionStmt>
16 <publicationStmt>
17 <p>unknown</p>
18 </publicationStmt>
19 <sourceDesc>
20 <p>Converted from a Word document by MHS WETVAC XSLT</p>
21 </sourceDesc>
22 </fileDesc>
23 <encodingDesc>
24 <appInfo>
25 <application xml:id="docx-to-tei-via-mhs-xslt" ident="TEI_fromDOCX_via_XSLT" version="0.2">
26 <label>DOCX to TEI</label>
27 </application>
28 <application ident="MHS-WETVAC" version="0.2b">
29 <label>MHS-WETVAC</label>
30 </application>
31 </appInfo>
32 </encodingDesc>
```



Questions?



Northeastern University
NULab for Texts, Maps, and Networks

*Feel free to ask questions at any point
during the presentation!*

TEI Resources

- [TEI Guidelines](#)
- [Introduction to XML](#) by Julia Flanders and Syd Bauman (2018)
- [Oxygen Software Help & User Guides](#)
- [Guide to customizing schema with One Document Does it all](#)
- [TEI Publisher](#) to practice and view demos of TEI encoding
- [Functional Requirements for Bibliographic Records relationships with TEI](#)
- [Gutentag](#): A tool to download Project Gutenberg texts as TEI files



Thank you!

If you have any questions, contact DITI at nulab.info@gmail.com

Developed by Juniper Johnson, Dipa Desai, Sara Morrell, Kasya O'Connor Grant, and Emily Sullivan

Digital Integration Teaching Initiative Research Fellows

Slides, handouts, and data available at <https://bit.ly/sp24-parr-hist7251-tei>

We'd love your feedback! Please fill out a short survey here:

<https://bit.ly/diti-feedback>

Schedule an appointment with us! <https://bit.ly/diti-meeting>

