An Intro to XML and TEI for *Literature in Context*

T. Howe and J. O'Brien
May 2023

What and why XML?

What world are we in?

X**ML** is a **M**arkup **L**anguage/syntax like HT**ML**, is part of the world of data that is stored and made available on the Internet.

and...

Because XML is often used to talk about texts--novels, poems, plays, essays, other forms of writing--we are also ***in the world of literary studies or humanities.***

Digital, Searchable, Accessible Collections of Full Texts

But what *is* XML?

XML stands for e**X**tensible **M**arkup **L**anguage. It is like HTML (HyperText Markup Language), which web pages are in part built out of. Almost everything that you look at in your web browser has at least a little bit of HTML behind it, though it has some other things, too.

XML and HTML are both **"metalanguages"**--or languages about languages. They both offer ***descriptors* of content** so that your browser or another kind of reader can read, "interpret," and display the content in a particular way.

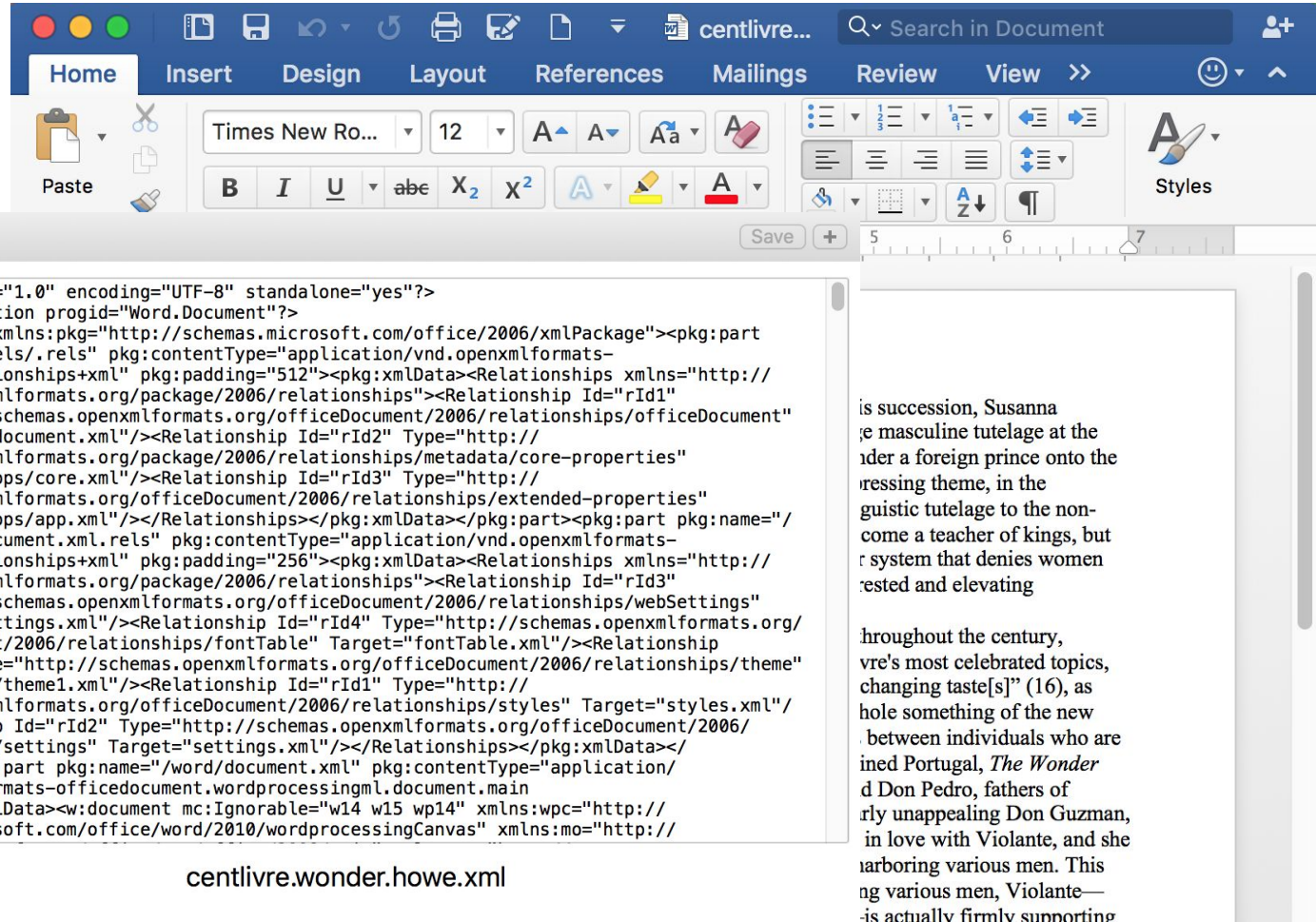
TEI s a specific kind or "flavor" of XML. TEI stands for **Text Encoding Initiative**, which is a "consortium that collectively develops and maintains a standard for the representation of texts in digital form."

Have you ever used Microsoft Word?

Then you have used XML! The new-ish **.docx** filetype has an X in it to indicate this.

If you've ever copy and pasted from a Word document, you might have noticed that there seems to be a whole bunch of gobbeldygook included--which you certainly didn't put there. Usually, we don't see it, because the XML is helping the computer display what you've written in a specific way.

It can indicate things font face and size, page breaks, footnote style, and it even helps make the link between your footnote reference and the actual footnote content--actually, everything about the format and setup of your document.



centlivre.wonder.howe.xml

XML enables text to be “read” by a computer

In conjunction with XSLT (eXtensible Style sheet Language Transformations), which is a computer language--not just a markup syntax--XML enables documents to be read by computers in various ways.

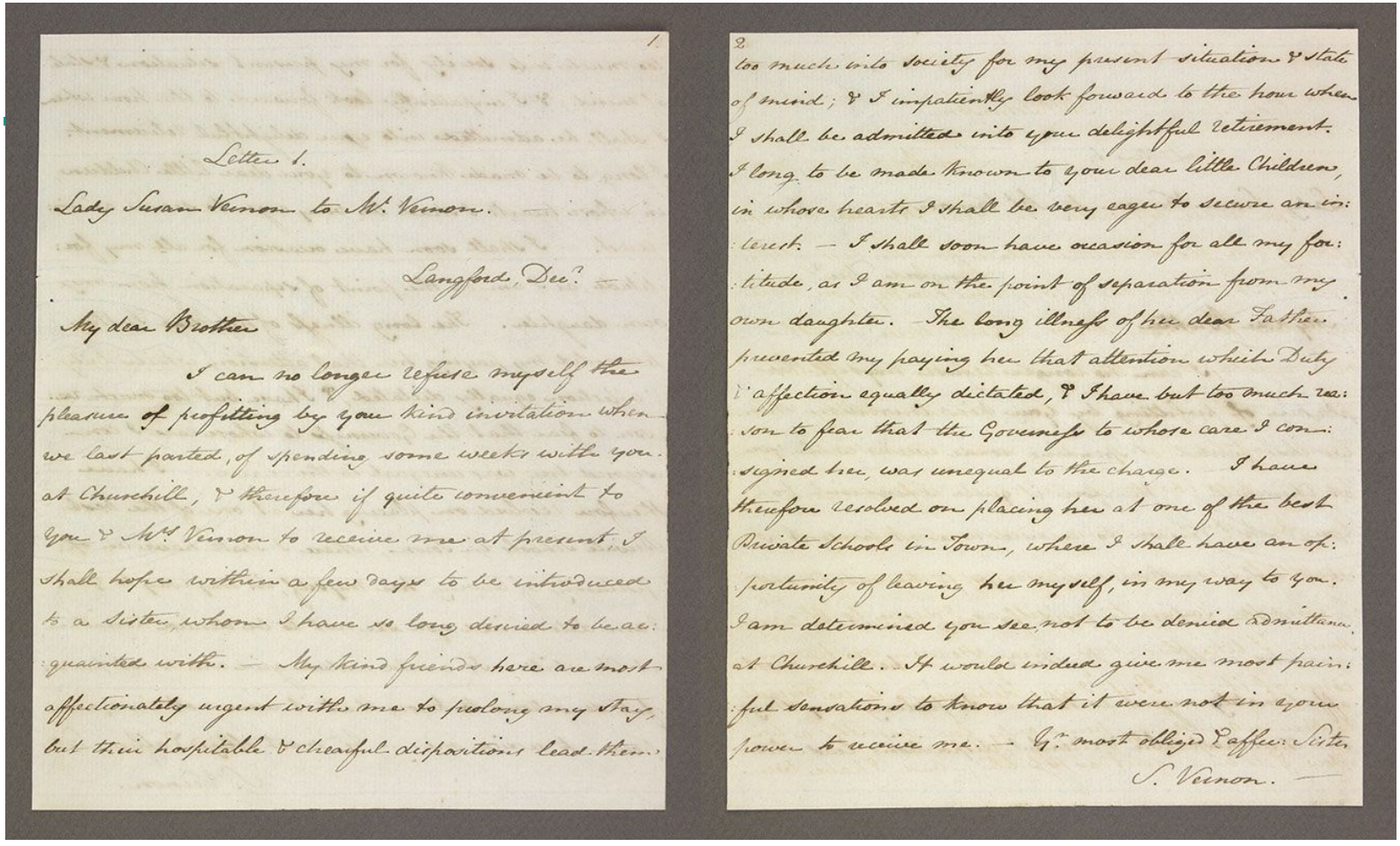
Computers aren’t “smart,” though AI is working on that. For our purposes, we’re using XML to ensure that our texts can be read and used by our database application.

To do that, we have to describe or mark up all the parts of the document so the application or the web browser “understands” what those parts are. We use a standardized version of XML called TEI, so that the text can be read by humans and machines more effectively.

Looking at just the nonspace characters gives us some sense of what a computer *doesn't* understand.

iladysusanvernontomrvernonlangforddecmydearbrothericannolongerrefusemyselfthepleasureofprofitingbyyourkindinvitationwhenwelastpartedofspendingsomeweekswithyouatchurchhillandthereforeifquiteconvenienttoyouandmrsvernontoreceivemeatpresentishallhopewithinafewdaystobeintroduucedtoasisterwhomihavesolongdesiredtobeacquaintedwithmykindfriendsherearemostaffectionatelyurgentwithmetoprolongmystaybuttheirhospitableandcheerfuldispositionsleadthemtoomuchintosocietyformypresentsituationandstateofmindandIimpatientlylookforwardtothehourwhenishallbeadmittedintoyourdelightfulretirementilongtobemadeknowntoyourdearlittlechildreninwhoseheartsishallbeveryeagertosecureaninterestishallsoonhaveneedforallmyfortitudeasiamonthepointofseparationfrommyowndaughterthelongillnessofherdearfatherprevented mypayingherthatattentionwhichdutyandaffectionequallydictatedandihavetoomuchreasontofearthatthegovernesstowhose careiconsignedherwasunequaltothechargeihavethereforeresolvedonplacingheratoneofthebestprivateschoolsintownwhere Ishallhaveanopportunityofleavinghermyselfinmywaytoyouiamdeterminedyouseenottobedeniedadmittanceatchurchhillitwouldindeedgivememostpainfulsensationstoknowthatitwerenotinyourpowertoreceivemeyourmostobligedandaffectionatesistersvernon

Manuscript of *Lady Susan*, in Austen's hand, held by the Pierpont Morgan Library & Museum in New York City.



1

Letter 1.

Lady Susan Vernon to Mr. Vernon.

Langford, Dec.r —

My dear Brother

I can no longer refuse myself the pleasure of profitting by your kind invitation when we last parted, of spending some weeks with you at Churchill, & therefore if quite convenient to you & Mrs. Vernon to receive me at present, I shall hope within a few days to be introduced to a Sister, whom I have so long desired to be acquainted with. — My kind friends here are most affectionately urgent with me to prolong my stay, but their hospitable & chearful dispositions lead them

2

too much into society for my present situation & state of mind; & I impatiently look forward to the hour when I shall be admitted into your delightful retirement. I long to be made known to your dear little Children, in whose hearts I shall be very eager to secure an interest. — I shall soon have occasion for all my fortitude, as I am on the point of separation from my own daughter. The long illness of her dear Father prevented my paying her that attention which Duty & affection equally dictated, & I have but too much reason to fear that the Governess to whose care I consigned her, was unequal to the charge. I have therefore resolved on placing her at one of the best Private Schools in Town, where I shall have an opportunity of leaving her myself, in my way to you. I am determined you see, not to be denied admittance at Churchill. It would indeed give me most painful sensations to know that it were not in your power to receive me. — Yr. most obliged & affec.e Sister.

S. Vernon. —

Letter I

Lady Susan Vernon to Mr. Vernon

Langford, Dec.

My dear Brother,

I can no longer refuse myself the pleasure of profiting by your kind invitation when we last parted of spending some weeks with you at Churchhill, and, therefore, if quite convenient to you and Mrs. Vernon to receive me at present, I shall hope within a few days to be introduced to a sister whom I have so long desired to be acquainted with. My kind friends here are most affectionately urgent with me to prolong my stay, but their hospitable and cheerful dispositions lead them too much into society for my present situation and state of mind; and I impatiently look forward to the hour when I shall be admitted into Your delightful retirement.

I long to be made known to your dear little children, in whose hearts I shall be very eager to secure an interest. I shall soon have need for all my fortitude, as I am on the point of separation from my own daughter. The long illness of her dear father prevented my paying her that attention which duty and affection equally dictated, and I have too much reason to fear that the governess to whose care I consigned her was unequal to the charge. I have therefore resolved on placing her at one of the best private schools in town, where I shall have an opportunity of leaving her myself in my way to you. I am determined, you see, not to be denied admittance at Churchhill. It would indeed give me most painful sensations to know that it were not in your power to receive me.

Your most obliged and affectionate sister,

S. Vernon.

How would a computer know, without being told, what the parts of this letter are, or even that it's a letter? Or that it's meant to be read in a certain way, or displayed on your browser in any particular way?

This is where "structured data" comes into play. XML is a way of structuring--giving shape to--data so that it can be used in more effective ways.

Basically...

XML allows you to create a “metalanguage” to help you “encode” your documents, so that machines can read and display them in different ways.

Let me take a minute to explain...

Structured data: any data that exists in a structured way, typically in a database. Think of a table in a spreadsheet--I have several rows of information, and each row has a label. Those labels tell me what the numbers in the cells of the tables are supposed to represent.

Markup language: a markup language is a way of "marking up" or "annotating" in a way that a machine can understand, so that the machine can be taught what to do with it.

Computers are dumb (right now, anyway), and they have to be told how to do everything. Stuff that is second nature to us--understanding when a word has been misspelled in your native language, for instance, or that something *like this* is italicized--is something that a computer has to be given a set of instructions to understand: <italic>like this</italic>, or <hi rend="italic">or this</hi>.

Letter I

Lady Susan Vernon to Mr. Vernon

Langford, Dec.

My dear Brother,

I can no longer refuse myself the pleasure of profiting by your kind invitation when we last parted of spending some weeks with you at Churchhill, and, therefore, if quite convenient to you and Mrs. Vernon to receive me at present, I shall hope within a few days to be introduced to a sister whom I have so long desired to be acquainted with. My kind friends here are most affectionately urgent with me to prolong my stay, but their hospitable and cheerful dispositions lead them too much into society for my present situation and state of mind; and I impatiently look forward to the hour when I shall be admitted into Your delightful retirement.

I long to be made known to your dear little children, in whose hearts I shall be very eager to secure an interest. I shall soon have need for all my fortitude, as I am on the point of separation from my own daughter. The long illness of her dear father prevented my paying her that attention which duty and affection equally dictated, and I have too much reason to fear that the governess to whose care I consigned her was unequal to the charge. I have therefore resolved on placing her at one of the best private schools in town, where I shall have an opportunity of leaving her myself in my way to you. I am determined, you see, not to be denied admittance at Churchhill. It would indeed give me most painful sensations to know that it were not in your power to receive me.

Your most obliged and affectionate sister,

S. Vernon.

** **
Any division of text

<head>
A heading

<opener>
Opener in a letter, usually has a dateline (with a place and a date) and a salutation

<p>
A paragraph (add more!)

<closer>
Closer in a letter, usually includes a salutation and a signature

</xxxx>
For non-empty elements, this kind of tag concludes or closes the markup

```
<div type="letter" n="1">
  <head type="sub">I.<lb/> <hi rend="italic">Lady Susan Vernon to Mr. Vernon.</hi></head>
  <opener>
    <dateline>
      <placeName>Langford</placeName>, <date>Dec.<lb/></date>
    </dateline>
    <salute>MY DEAR BROTHER,--</salute>
  </opener>
  <p>I can no longer refuse myself the pleasure of profiting by your kind invitation when we last parted of spending some weeks with you at Churchhill, and, therefore, if quite convenient to you and Mrs. Vernon to receive me at present, I shall hope within a few days to be introduced to a sister whom I have so long desired to be acquainted with. My kind friends here are most affectionately urgent with me to prolong my stay, but their hospitable and cheerful dispositions lead them too much into society for my present situation and state of mind; and I impatiently look forward to the hour when I shall be admitted into your delightful retirement.</p>
```

** **
Any division of text

<head>
A heading

<opener>
Opener in a letter, usually has a dateline (with a place and a date) and a salutation

<p>
A paragraph (add more!)

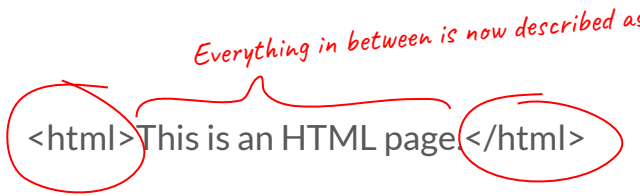
<closer>
Closer in a letter, usually includes a salutation and a signature

</xxxx>
For non-empty elements, this kind of tag concludes or closes the markup

```
<p>I long to be made known to your dear little children, in whose hearts I shall be very eager to secure an interest I shall soon have need for all my fortitude, <pb n="204" facs="pageImages/204.jpg"/>as I am on the point of separation from my own daughter. The long illness of her dear father prevented my paying her that attention which duty and affection equally dictated, and I have too much reason to fear that the governess to whose care I consigned her was unequal to the charge. I have therefore resolved on placing her at one of the best private schools in town, where I shall have an opportunity of leaving her myself in my way to you. I am determined, you see, not to be denied admittance at Churchhill. It would indeed give me most painful sensations to know that it were not in your power to receive me.</p>
<closer>
  <salute>Your most obliged and affectionate sister,<lb/></salute>
  <signed>S. VERNON.</signed>
 </closer>
</div>
```

As in HTML, so too in XML

Basic syntax for both HTML and XML work in the same way. You use “open” and “close” elements--sometimes called tags--to surround parts of the document and thereby describe them for the machine. These always* occur in pairs: one to open, and one to close.



The slash inside the second element indicates that this is the closing part of the tag or element pair--everything between these elements is whatever the element describes it to be.

*Usually. Some elements are “empty” because they don’t contain anything--they just *are*, and they form the beginning and the end together in one. Page beginnings <pb/> are like this, as are hard returns <hr/> or line breaks
.

But how do I know how to describe my texts?

That's where we need what's called a Standardized Vocabulary. One standardized vocabulary of XML is TEI-style encoding.

Remember that TEI is a consortium of researchers--scholars and librarians and computer programmers--"that collectively develops and maintains a standard for the representation of texts in digital form."

TEI: Text Encoding Initiative

TEI: Text Encoding Initiative

The Text Encoding Initiative came together in 1987 to create guidelines for using XML in a consistent way.

TEI is based on XML. It is a particular “metalanguage” that scholars use when they are trying to encode the physical quality, the content, and a variety of other features of a text so that it can be read by computers, searched more effectively, and displayed in a website.

TEI is a standardized version of XML that is used to describe documents that humanists and social scientists find interesting.

There is a guide, online: https://www.tei-c.org/release/doc/tei-p5-doc/en/html/index.html

Digital, Searchable, Accessible Collections of Full Texts

Examples

When you research in the library databases, and you search for a keyword in the abstract of a journal article--you're searching structured data! How does the computer know what an "abstract" is? It doesn't--it just knows that this chunk of text has been called an "abstract" by someone, and it has been told to search only within those things called "abstracts."

- The Women Writer's Project contains a collection of early modern writings by women, all encoded in TEI.
- So does the Colonial Despatches project, from the University of Victoria.
- So does the Swinburne Archive, which houses a digital archive focused on the Victorian poet Algernon Charles Swinburne.
- The Old Bailey Online project contains the legal proceedings (1674-1913) from the Old Bailey, which was London's central criminal court.

The Old Bailey Online

The digitised text can be searched for any character string, but in order to facilitate structured searching and the generation of statistics, the text was also "marked up" in XML. [...To] create meaningful and consistent statistics, certain subcategories of information were also identified, such as types of verdict. The following categories of information have been marked up (* fields can be tabulated statistically):

- * Crime (divided into 9 general categories and 56 specific types)
- Crime date
- Crime location
- Defendant name
- Defendant status or occupational label
- * Defendant gender
- Alias names used by the defendant and the victim
- Defendant location
- Victim name
- Victim status or occupational label
- * Victim gender
- Judges' names
- Jury names
- Other person names
- * Verdicts (divided into 4 general categories and 23 specific types)
- * Punishments (divided into 6 general categories and 26 specific types)
- * Defendant's age (only regularly provided for convicts from 1789)
- Advertisements

Slightly edited from original source.

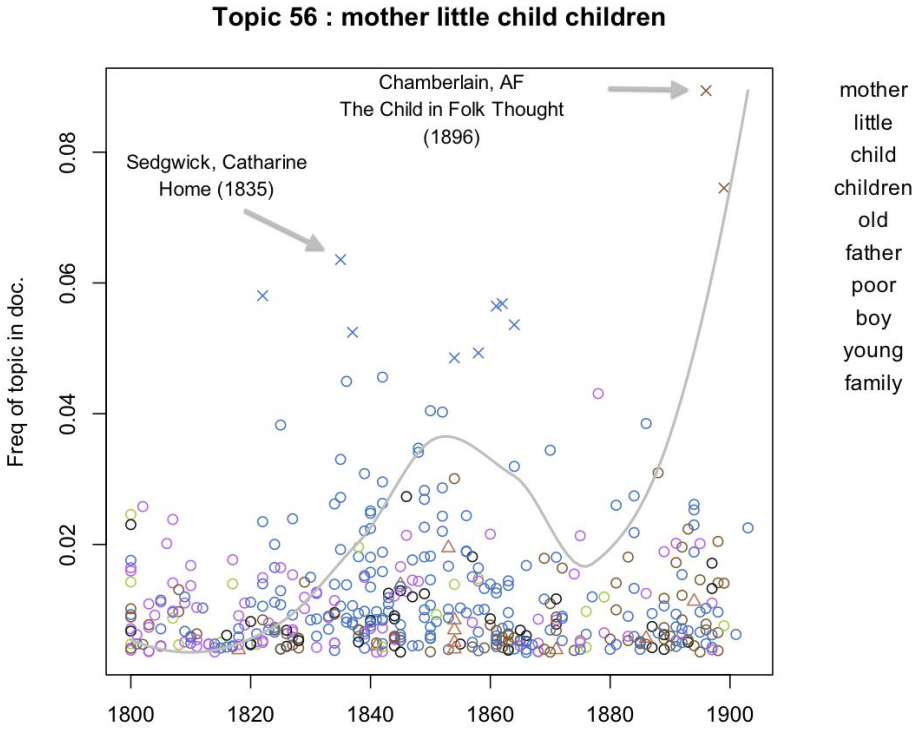
Why Might This Be Useful?

To be able to search for “innocent” verdicts during the early 18th century related to the crime of rape, for instance?

To have searchable full-text access to all the plays of Shakespeare in a reliable electronic edition that is supplemented with page images from the first folio?

To be able to compare the 1712, the 1714, and the 1717 editions of Pope’s *Rape of the Lock* side-by-side on your computer screen?

To be able to search for and then graph language changes in all novels written in English over the course of the 18th and 19th centuries?



Blue/fic, purple/poe, green/drama, black/bio, brown/nonfic, triangle/letters or orations.

[Source Left](), [Source Right]().

FIGURE 9: *British novelistic genres, 1740–1900*

For sources, see 'Note on the Taxonomy of the Forms', page 91.

Big Collections, Small Collections, Individual Documents

How-To XML in TEI for LiC 😊

The Basics

Key Features of TEI: The Basics

Always **an opening and a closing tag.**

Open and close tags surround the text that those tags describes. Tags are also called “**elements**”.

But… sometimes they’re smooshed together:
 ← like that. This happens when nothing in particular is being marked up. It’s called an “empty element.”

Elements can be **<rs type=”explanation”>nested</rs>**, but they must be nested properly.

Elements are cAsE sENSitiVe. <P> =/

Elements, even empty elements, might be further defined by “attributes.” Shorthand is @type.

Key Features of TEI: The Basics

Most TEI documents contain key parts, nested inside other parts in an organized way.

```
<TEI>

  <teiHeader>
  Description of the document
  </teiHeader>

<text>
  <front>
  Front matter
  </front>

  <body>
   Body of the text
  </body>

  <back>
  Back matter
  </back>
</text

</TEI>
```

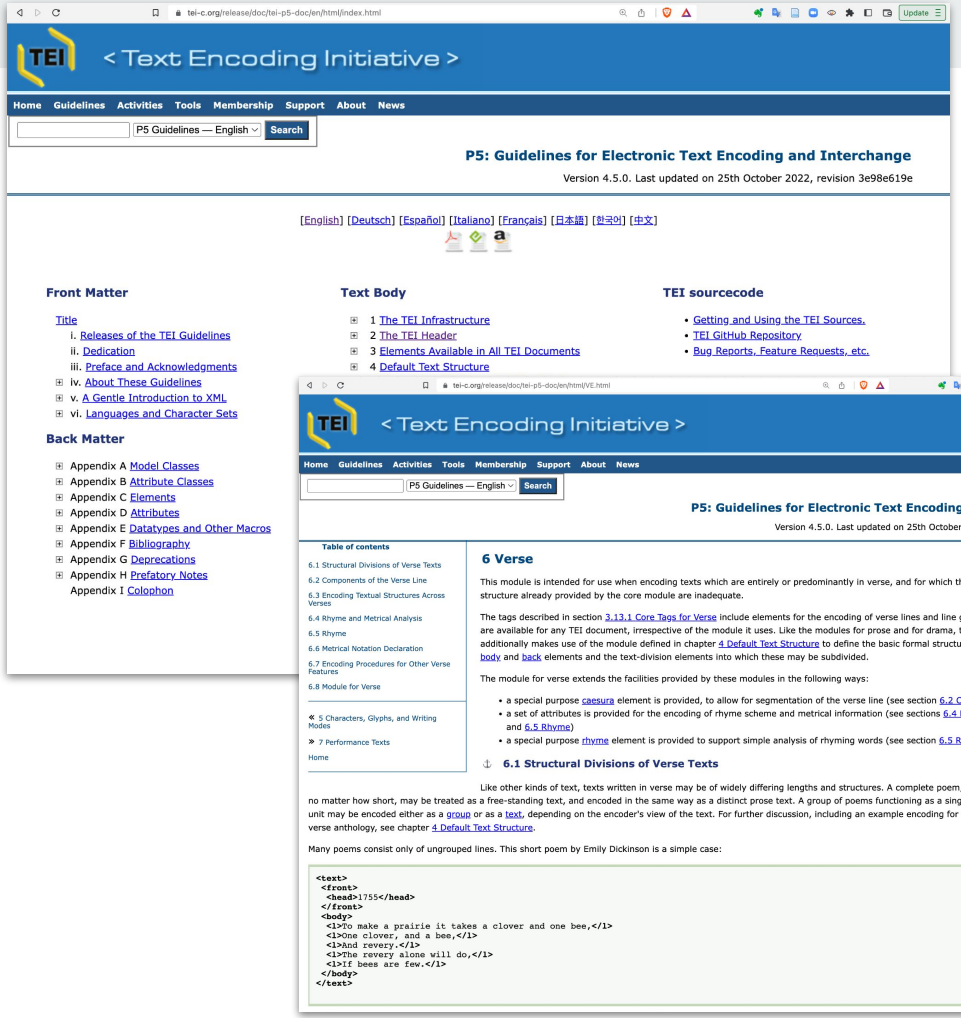
Notice: There should be no characters or bits of text NOT enclosed in markup.

P5 Guidelines

A set of guidelines you can search and read online for TEI-standardized XML.

Also contains examples in varying degrees of complexity.

https://www.tei-c.org/release/doc/tei-p5-doc/en/html/index.html



LiC <teiHeader>

Contains the important metadata about the file, its sources, and its construction. P5

General Structure: fileDesc

```
<teiHeader>
        <fileDesc>
                <titleStmt>
                        <title/>
                        <author/>
                        <editor/>
                        <respStmt/>
                        <sponsor/>
                        <funder/>
                </titleStmt>
                <publicationStmt/>
                <sourceDesc>
                        <biblStruct/>
                </sourceDesc>
        <fileDesc>
```

titleStmt

Contains details about this XML title, including original author/s, general editor, and any student collaborators

```
<titleStmt>
    <title type="main">”An Ode [Rule, Britannia]”</title>
    <author>
      <persName type="lcnaf" key="n79065635">
        <name ref="thomson-arne.xml">
          <forename>James</forename>
          <surname>Thomson</surname>
        </name>
      </persName>
      <persName type="lcnaf" key="n80017895">
        <forename>David</forename>
        <surname>Mallet</surname>
      </persName>
      <persName type="lcnaf" key="n79139208">
        <name ref="thomson-arne.xml">
          <forename>Thomas</forename>
          <surname>Arne</surname>
        </name>
      </persName>
    </author>
    <editor>
      <persName type="orcid" key="0000-0002-7400-4093">
        <name ref="editors.xml#TH">
          <surname>XXX</surname>
          <forename>XXX</forename>
        </name>
      </persName>
    </editor>
    <respStmt>
      <resp>Correction, editorial commentary, and markup</resp>
      <name ref="editors.xml#MUstudstaff"/>
    </respStmt>
    <sponsor/>
    <funder>VIVA</funder>
</titleStmt>
```

publicationStmt

Contains details about the institutional source of this XML publication.

```
<publicationStmt>
   <publisher>Literature in Context</publisher>
   <address>
     <addrLine>Marymount University Department of Literature and Languages</addrLine>
     <addrLine>2807 N. Glebe Road</addrLine>
     <addrLine>Arlington, VA </addrLine>
     <addrLine>22207</addrLine>
     <addrLine>thowe@marymount.edu</addrLine>
     <addrLine>lic.open.anthology@gmail.com</addrLine>
   </address>
   <availability status="free">
     <licence target="http://creativecommons.org/licenses/by-sa/4.0/">Published by
       Literature in Context under a Creative Commons Attribution-ShareAlike 4.0 Unported
       License </licence>
   </availability>
 </publicationStmt>
```

sourceDesc

Contains a description of the source text/”witness” used for this XML edition. Details from this element create the citation and other parts of the digital edition.

```
<sourceDesc>
    <biblStruct>

        <analytic>
          <title>"An Ode [Rule, Britannia]"</title>
        </analytic>

        <monogr>
          <author>
            <name ref="thomson-arne.xml">
              <forename>James</forename>
              <surname>Thomson</surname>
            </name>
          </author>
          <title>Alfred: A Masque</title>
          <imprint>
          <pubPlace>
            <placeName type="tgn" key="7011781">London</placeName>
          </pubPlace>
          <publisher>A. Millar</publisher>
          <date when="1740">1740</date>
          <note/>
        </imprint>
        <extent>44p.; 8º.</extent>
        <biblScope>pp 42-43</biblScope>
      </monogr>
    </biblStruct>
</sourceDesc>
```

General Structure: profileDesc

Contains a description of the creation and setting of the text. See P5 for complete details, and XML template.

```
<profileDesc>
        <langUsage>
                <language/>
        </langUsage>
        <creation/>
        <textDesc n=”xxx”>
                <channel/>
                <constitution/>
                <derivation/>
                <domain/>
                <factuality/>
                <interaction/>
                <preparedness/>
                <purpose/>
        </textDesc
        <settingDesc>
                <setting>
                        <name/>
                        <time/>
                </setting>
        </settingDesc>
<profileDesc>
```

encodingDesc

Contains details of how the text has been encoded, including editorial choices.

Please read the boilerplate, as this describes base expectations.

If you deviate from these expectations, indicate them in the appropriate area of the encodingDesc.

```
<encodingDesc>
    <projectDesc>
      <p>This text is prepared as part of…</p>
    </projectDesc>
    <editorialDecl>
      <interpretation>
        <p>Research informing these annotations draws on…</p>
      </interpretation>
      <normalization>
        <p>Original spelling and capitalization…</p>
      </normalization>
      <hyphenation>
        <p>Hyphenation…</p>
      </hyphenation>
      <segmentation>
        <p>Page breaks…</p>
      </segmentation>
      <correction>
        <p>Materials have been transcribed from and checked against…</p>
      </correction>
    </editorialDecl>
    <tagsDecl>
      <namespace name="http://www.tei-c.org/ns/1.0">
        <tagUsage gi="div">Unnumbered divs used.</tagUsage>
      </namespace>
    </tagsDecl>
    <classDecl>
      <taxonomy xml:id="lcnaf">
        <bibl>Library of Congress Name Authority File</bibl>
      </taxonomy>
      <taxonomy xml:id="lcc">
        <bibl>Library of Congress Classification</bibl>
      </taxonomy>
      <taxonomy xml:id="tgn">
        <bibl>Getty Thesaurus of Geographic Names</bibl>
      </taxonomy>
      <taxonomy xml:id="orcid">
        <bibl>Open Researcher and Contributor ID</bibl>
      </taxonomy>
    </classDecl>
  </encodingDesc>
```

revisionDesc

Contains a record of the changes to the document and who made them.

```
        <revisionDesc>
          <change when="2022-07-22" who="editors.xml#TH">First pass XML file</change>
        </revisionDesc>
</fileDesc>
```

LiC <text><front>

Contains the text content, divided into front matter, body, and back matter. P5

front

Contains front matter, usually a page beginning (titlepage) and the encoded title page. Best practice is to encode the title page for each text, but in some instances, this is not appropriate.

A titlepage has each main line of the title noted in <titlePart/>s separated with <lb/>s. All <titlePage>s must have a <docImprint>. Associating the schema correctly will help you identify the required parts. Here is an example of a valid minimally-encoded title page for a text where it is not approprate/necessary to encode the titlepage.

```
<text>
  <front>
      <pb n="[TP for Alfred: A Masque]" facs="pageImages/TP.png"/>
      <pb n="42" facs="pageImages/42.png"/>
      <titlePage>
        <titlePart> An <hi rend="italic">ODE</hi>. [Rule, Britannia]</titlePart>
        <docImprint/>
      </titlePage>
  </front>
[...]
</text>
```

```
<front>
   <pb n="[TP]" facs="pageImages/256.jpg" />
   <titlePage>
     <docTitle>
          <titlePart><hi rend="italic">FANTOMINA:</hi><lb/></titlePart>
          <titlePart>OR,<lb/></titlePart>
          <titlePart>LOVE in a Maze.<lb/></titlePart>
          <titlePart>BEING A<lb/></titlePart>
          <titlePart><ref target="secret_history_" corresp="secret_history">Secret
History</ref><note xml:id="secret_history"
               target="secret_history_" type="editorial" resp="editors.xml#TH">TEXT
OF ANNOTATION FOR “Secret History”</note><lb/></titlePart>
          <titlePart>OF AN<lb/></titlePart>
          <titlePart>AMOUR<lb/></titlePart>
          <titlePart>Between Two<lb/></titlePart>
          <titlePart>PERSONS OF CONDITION.<lb/></titlePart></docTitle>
          <docAuthor>By Mrs. ELIZA HAYWOOD.<lb/></docAuthor>
          <epigraph>
           <quote>
             <lg>
               <l><hi rend="italic">In Love the Victors from the Vanquish'd
fly</hi>.</l>
               <l><hi rend="italic">They fly that wound, and they pursue that
             dye.</hi></l></lg>
             </quote>
              <bibl><persName>WALLER.</persName><lb/></bibl>
              <lb/>
          </epigraph>
          <docImprint><pubPlace><placeName type="tgn"
key="7011781">London</placeName></pubPlace>:<lb/>
            <publisher>Printed for <persName>D. BROWNE <hi
rend="italic">jun</hi></persName>. at the
                <placeName><hi rend="italic">Black-Swan</hi><lb/> without
<hi rend="italic">Temple-Bar</hi></placeName>, and <persName>S.
CHAPMAN</persName>, <placeName>at<lb/> the <hi
rend="italic">Angel</hi> in <hi
rend="italic">Pallmall</hi></placeName>.</publisher>
           <docDate>M.DCC.XXV.</docDate><lb/></docImprint>
   </titlePage>
</front>
```

LiC <text><body>

After the front matter in the text block comes the body of the text.

P5

body

Contains the main body of the text. Minimally, it must have a . How it is encoded will depend in great part on the nature of the text. Here is the beginning of a poem:

```
 </front>
 <body>
     <div>
         <head type="sub">1.</head>
         <l n="1">When <hi rend="italic">Britain</hi> first, at heaven's command,</l>
         <l n="2" rend="indent">Arose from out the azure main;</l>
         <l n="3"><hi rend="italic">This</hi> was the charter of the land,</l>
         <l n="4" rend="indent">And guardian Angels sung <hi rend="italic">this</hi> strain:</l>
         <l n="5" rend="indent2">"Rule <hi rend="italic"><ref target="britannia_" corresp="britannia">Britannia</ref>,</hi> rule the waves;</l>
         <note xml:id="britannia" target="britannia_" type="editorial" resp="editors.xml#TH" ><graphic
url="notes/The_East_offering_its_riches_to_Britannia_-_Roma_Spiridone,_1778_-_BL_Foster_245.jpg"/>Britannia is a figurative, allegorical representation of
Britain as a female warrior carrying a trident and a shield, often accompanied by a lion….</note>
         <l n="6" rend="indent2">"<hi rend="italic">Britons</hi> never will be slaves."</l>
       </lg>
       …
     </div>
</body>
```

LiC <text><back>

After the main text block comes any back matter. P5

back

Contains any back matter included in the text. Sometimes this is advertising info, cast lists, indices, and so on. Not all texts will have back matter; in that case, just include an empty element:

```
        […]
      </div>
   </body>
   <back/>
</TEI>
```

See templates and schemas for poems, prose pieces, and plays for more guidance. And don't forget using other LiC texts as models, or the value of the P5!

LiC <text><back>

back

Contains any back matter, if necessary.

```
<back>
        <div>
                <p>Sold at the Sign of the Bear in Newgate Street, where Advertisements taken in. </p>
        </div>
 </back>
```

In general, markup is straightforward

Usually, you'll be marking up:

Large textual divisions (note nesting!)

- ...
- act 1
- <opener></opener> text

Small textual divisions

- <head type="main" or "sub">heading text</head>
- paragraph content
- **Letters** have openers, closers, dateLines, salutations, signatures, and so on.
- **Plays** have speeches, speakers, stage directions, and so on.

Page beginnings

- <pb n="342" facs="pageImages/342.jpg"/>

Italics and Indentation

- <hi rend="italic">text italicized</hi>
- <l rend="indent4">Line of poetry</l>

Line numbers (restart by scene; in drama and poetry only)

- <l n="1">Line here</l>
- <lg type="stanza"><l n="13">...</l></lg>

Annotations (not this grant cycle)

Places and people (not this grant cycle--a form of annotation)

- <placeName type="tgn" key="7013425">Birmingham</placeName>
- <name type="lcnaf" key="unique#">Person's Name</name>

LiC texts are also annotated using XML, but that is not something we're doing in this grant cycle.

Usually, you'll be marking up:

Large textual divisions (note nesting!)

- ...
- act 1
- <opener></opener> text

Small textual divisions

- <head type="main" or "sub">heading text</head>
- paragraph content
- **Letters** have openers, closers, dateLines, salutations, signatures, and so on.
- **Plays** have speeches, speakers, stage directions, and so on.

Page beginnings

- <pb n="342" facs="pageImages/342.jpg"/>

Italics and Indentation

- <hi rend="italic">text italicized</hi>
- <l rend="indent4">Line of poetry</l>

Places and people

- <placeName type="tgn" key="7013425">Birmingham</placeName>
- <name type="lcnaf" key="unique#">Person's Name</name>

Line numbers (restart by scene; in drama and poetry only)

- <l n="1">Line here</l>
- <lg type="stanza"><l n="13">...</l></lg>

Annotations (more on this later)

Schema and Templates

Schema is a sort of rubric to help us determine if the XML structure is valid for our purposes.

Our schema is available here, on our GitHub repository:
https://raw.githubusercontent.com/LiteratureInContext/LiC-data/master/schema/LiC_schema_3.rng

It should be added to each TEI file just above the header:

```
<?xml version="1.0" encoding="UTF-8"?>
<?xml-model
href="https://raw.githubusercontent.com/LiteratureInContext/LiC-data/development/schema/LiC_sche
ma_3.rng" type="application/xml" schematypens="http://relaxng.org/ns/structure/1.0"?>
<TEI xmlns="http://www.tei-c.org/ns/1.0" xml:id="behn-rover">
  <teiHeader>
    <!--TEI Header BEGINS-->
```

Other important document identification…

The application uses the document xml:id to identify it and its various associated files. The document xml:id is therefore extremely important. We use the formula “author-short-title.”

That is also what the XML file itself should be saved as: “author-short-title.xml”.

In our AWS server, every text has a folder called “author-short-title”--and that folder contains two sub-folders, “notes” and “pageImages”. These are the storage sites for images in annotations and facsimile page images, respectively. They link to the document on the application site because of the consistency in file naming conventions and the xml:id.

Oxygen and Github

GitHub

GitHub is a version control system for software development, and it is especially useful when development is happening in collaboration.

Every LiC XML document is stored on the GitHub data repository, first in the development branch, where we can work on it. Each LiC collaborator will have a cloned version of this GitHub repository on their local machine where changes are made. Those changes are then pushed into the GitHub repository to be synchronized on everyone else's local working machines.

Then, when it is ready to go to the permanent site, the development branch is pushed through the dev2master branch to the master branch. A series of webhooks make this process work fairly seamlessly.

Oxygen & GitHub

Oxygen is the editing platform we use to edit XML documents. It can be downloaded from https://www.oxygenxml.com/.

There is a GitHub plugin for Oxygen; or, you can use GitHub Desktop, a separate application. In both cases, you will need to create a GitHub account and be added as a collaborator to our GitHub repo: https://github.com/LiteratureInContext/LiC-data/tree/development.

Once you're a collaborator (which gives you rights to collaborate), you clone the repo on your local machine, and from Oxygen or GitHub Desktop, you can pull down recent changes to synchronize your local version, then make your changes, then push them to the development branch. Each time you continue work, you should synchronize your local machine with the GitHub development repository. This ensures we are not working at cross-purposes.

Where to find XML texts and page images?

Where to find XML texts?

EEBO TCP: Early English Books Online Text Creation Partnership
Printed books before 1700 from the ESTC: https://quod.lib.umich.edu/e/eebogroup/

ECCO TCP: Eighteenth-Century Collections Online Text Creation Partnership
Printed books 1701-1800 from the ESTC: https://quod.lib.umich.edu/e/ecco/

Evans TCP: Evans Early American Imprints Text Creation Partnership
Printed books 1470-1820 from the Evans Early American Imprints series: https://quod.lib.umich.edu/e/evans/

UVA eText Center: https://dcs.library.virginia.edu/digital-stewardship-services/etext/

Individual projects, like the Blake Archive

Modern materials: Hand-encoding from plaintext, OCR, &c

What about page images?

It is important that you seek out the earliest authoritative version of the text at hand. The page images should correspond to that version. Sometimes, you have to be a detective! And sometimes, you have to go to a library and secure the page images yourself (or ask a library faculty for help).

HathiTrust, Google Books, and Internet Archive are sometimes good starting points, but it's equally useful to reach out to those who can assist.

Make sure the page images are the best quality possible, full color (vs b/w) when possible. If the text is long, then it is better to secure the title page and the first few pages of text, or another option for giving readers a good sense of what this material object looks like.

EEBO/ECCO to LiC

Coming from EEBO/ECCO-TCP texts...

There are many differences between ECCO/EEBO-TCP texts and LiC texts.

Use the template or an existing LiC text in the same general form as a model:

https://docs.google.com/document/d/1CGWalKEh81L6oknp2DCS4G1Cs22r34xXEDmzlok9JGA/edit?usp=sharing

Use the TEI Guidelines to answer questions, or ask!

Use “Find/Replace” to make upper/lowercase replacement easy, and to delete unnecessary elements (don’t forget the closing slash!). Google “regular expressions” + what you want to do to learn more.

Regular expression in oXygen for content between and including two strings:

https://docs.google.com/document/d/1afPzwsSl4cPWwtZxsBxedV8M5UDUGDClAf-PDMgH2kQ/edit?usp=share_link

Coming from EEBO/ECCO-TCP texts…

TCP texts don't identify places or people; LiC does, for mapping and social networking visualizations.

TCP texts use numbered divs <DIV1><DIV2>; LiC uses unnumbered divs, with attributes as necessary.

TCP texts use internal @facs and they will need to be deleted.

TCP texts use uppercase elements; they must all be lowercase for LiC.

TCP texts use <HI> instead of <hi rend="italic">; they will need to be corrected.

TCP text teiHeader doesn't fully match up with LiC teiHeader. Ignore for now, or update to match LiC teiHeader using the template.