Introduction to Text Encoding

Dipa Desai, Claire Lavarreda
Digital Integration Teaching Initiative
HIST 1357: Data, Surveillance, and Society
Prof. Jess Parr
Spring 2026

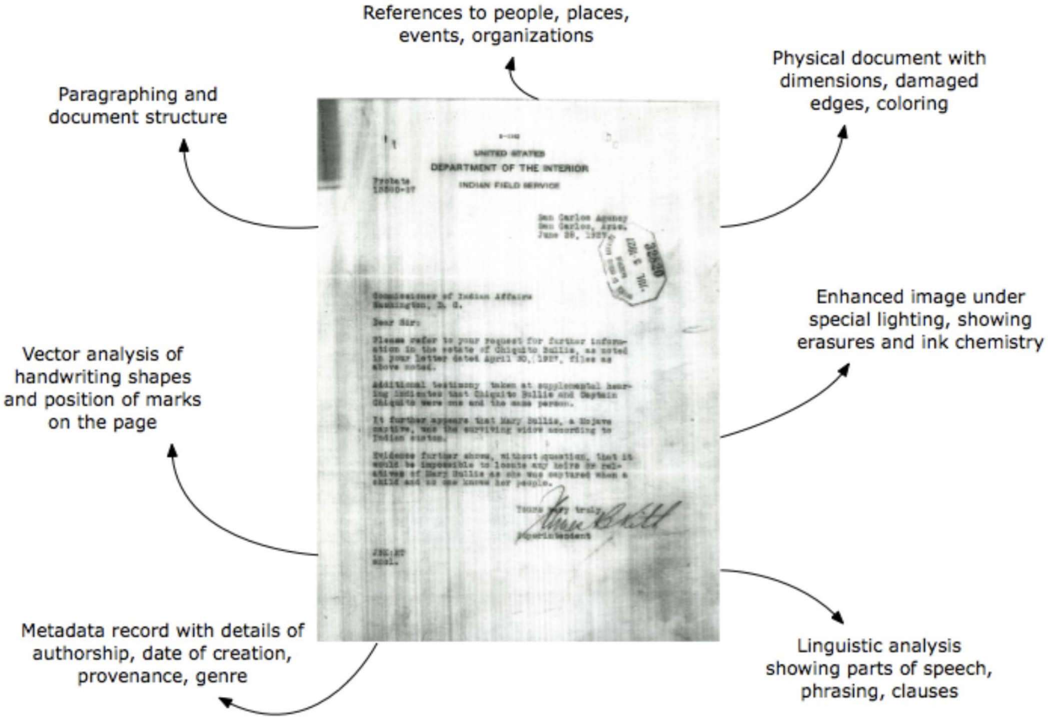
Workshop Agenda

- What is TEI?
- Overview of TEI Guidelines + Encoding Structures
  - Navigating the TEI Guidelines
- Text Encoding Project Examples
- Class Practice: Encoding text

Find all of the course materials at: https://bit.ly/sp26-parr-hist1357-multi

What is text encoding?

Representing research objects


References to people, places, events, organizations
Paragraphing and document structure
Physical document with dimensions, damaged edges, coloring
Enhanced image under special lighting, showing erasures and ink chemistry
Vector analysis of handwriting shapes and position of marks on the page
Linguistic analysis showing parts of speech, phrasing, clauses
Metadata record with details of authorship, date of creation, provenance, genre

UNITED STATES
DEPARTMENT OF THE INTERIOR
INDIAN FIELD SERVICE

Probate
10800-27

San Carlos Agency
San Carlos, Ariz.
June 28, 1927.

Commissioner of Indian Affairs
Washington, D. C.

Dear Sir:

Please refer to your request for further information in the estate of Chiquito Bullis, as noted in your letter dated April 30, 1927, files as above noted.

Additional testimony taken at supplemental hearing indicates that Chiquito Bullis and Captain Chiquito were one and the same person.

It further appears that Mary Bullis, a Mojave captive, was the surviving widow according to Indian custom.

Evidence further shows, without question, that it would be impossible to locate any heirs or relatives of Mary Bullis as she was captured when a child and no one knows her people.

Yours very truly,

Superintendent

JBK:ST
encl.

Text encoding introductory slides from [this presentation] by the Women Writers Project

Background on the TEI

The TEI is:

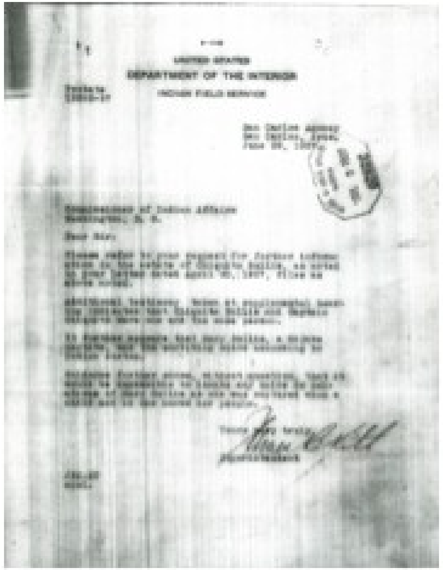
- A markup language (a text encoding language)
- Developed by an international consortium; free and open-source
- Both a community standard and a community research effort

The TEI lets us model texts:

- **Sustainably**—in a plain-text, non-proprietary format
- **Shareably**—using an international community standard adopted by hundreds of projects
- **Articulately**—in a system that provides very fine levels of detail for describing documents
- **Formally**—in a language that both humans and computers can understand and that provides for consistent representation and programmatic retrieval of information

Formalism, Selection, Description

Raw stuff



UNITED STATES
DEPARTMENT OF THE INTERIOR
INDIAN FIELD SERVICE

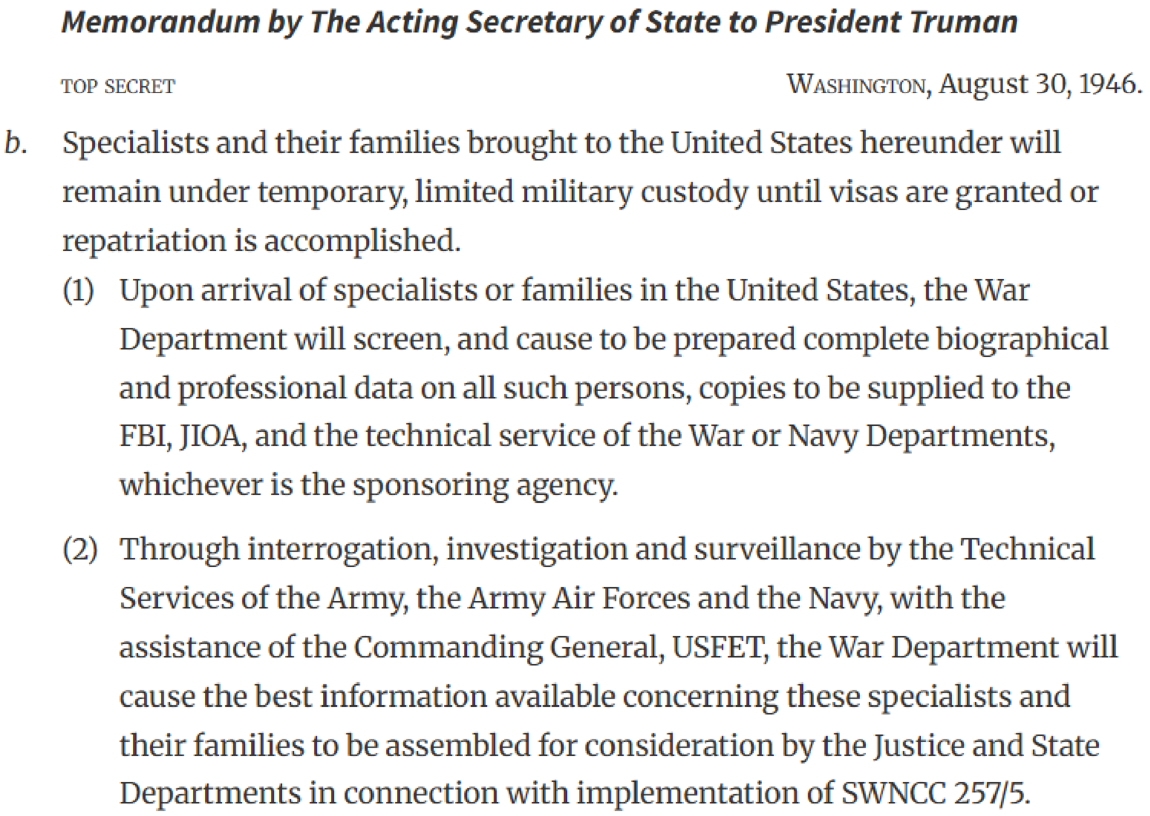
[illegible]

Dear Sir:

[illegible]

Our selection	Our formal description
page size	<dimensions type="page"> <height>200</height> <width>140</width> </dimensions>
text structures	<p>, <salute>, <dateline>
named entities	<persName>, <placeName>
illegible passages	<gap>, <unclear>

Sample text: Memo to President Truman



Memorandum by The Acting Secretary of State to President Truman

TOP SECRET

WASHINGTON, August 30, 1946.

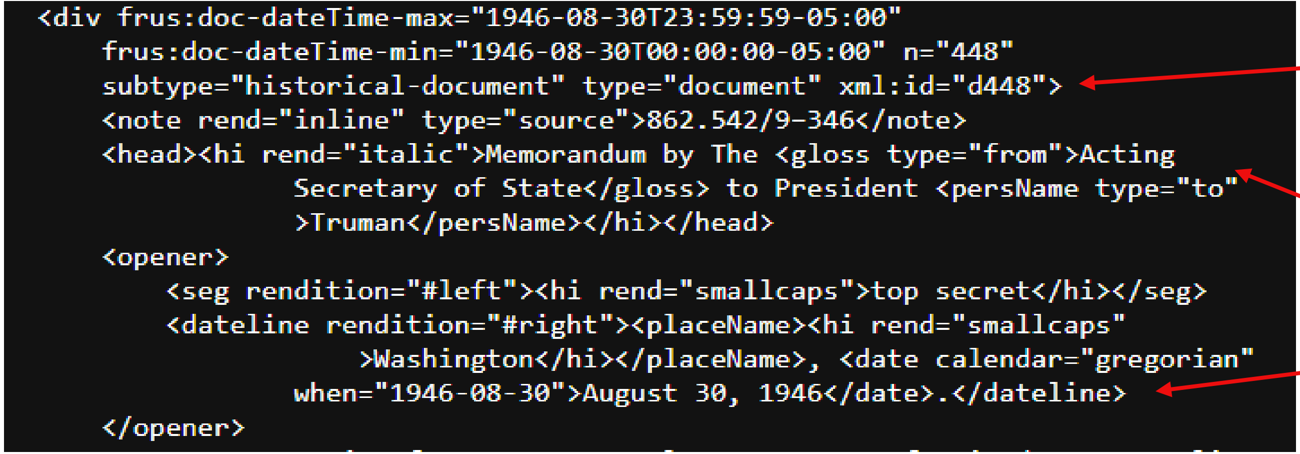
b. Specialists and their families brought to the United States hereunder will remain under temporary, limited military custody until visas are granted or repatriation is accomplished.

(1) Upon arrival of specialists or families in the United States, the War Department will screen, and cause to be prepared complete biographical and professional data on all such persons, copies to be supplied to the FBI, JIOA, and the technical service of the War or Navy Departments, whichever is the sponsoring agency.

(2) Through interrogation, investigation and surveillance by the Technical Services of the Army, the Army Air Forces and the Navy, with the assistance of the Commanding General, USFET, the War Department will cause the best information available concerning these specialists and their families to be assembled for consideration by the Justice and State Departments in connection with implementation of SWNCC 257/5.

Image credit: [Foreign Relations of the U.S.]

Sample encoding

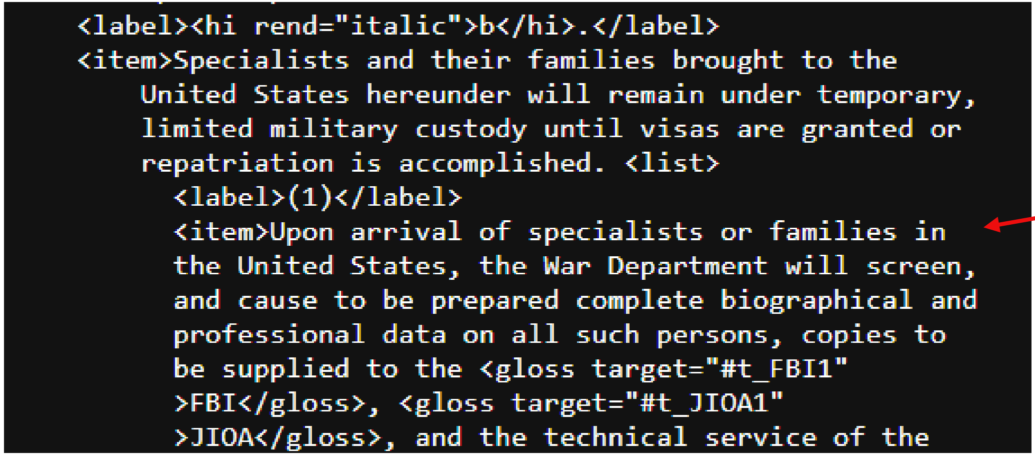


```
<div frus:doc-dateTime-max="1946-08-30T23:59:59-05:00"
    frus:doc-dateTime-min="1946-08-30T00:00:00-05:00" n="448"
    subtype="historical-document" type="document" xml:id="d448">
    <note rend="inline" type="source">862.542/9-346</note>
    <head><hi rend="italic">Memorandum by The <gloss type="from">Acting
                Secretary of State</gloss> to President <persName type="to"
                >Truman</persName></hi></head>
    <opener>
        <seg rendition="#left"><hi rend="smallcaps">top secret</hi></seg>
        <dateline rendition="#right"><placeName><hi rend="smallcaps"
                    >Washington</hi></placeName>, <date calendar="gregorian"
                when="1946-08-30">August 30, 1946</date>.</dateline>
    </opener>
```

Document type

Title

Date



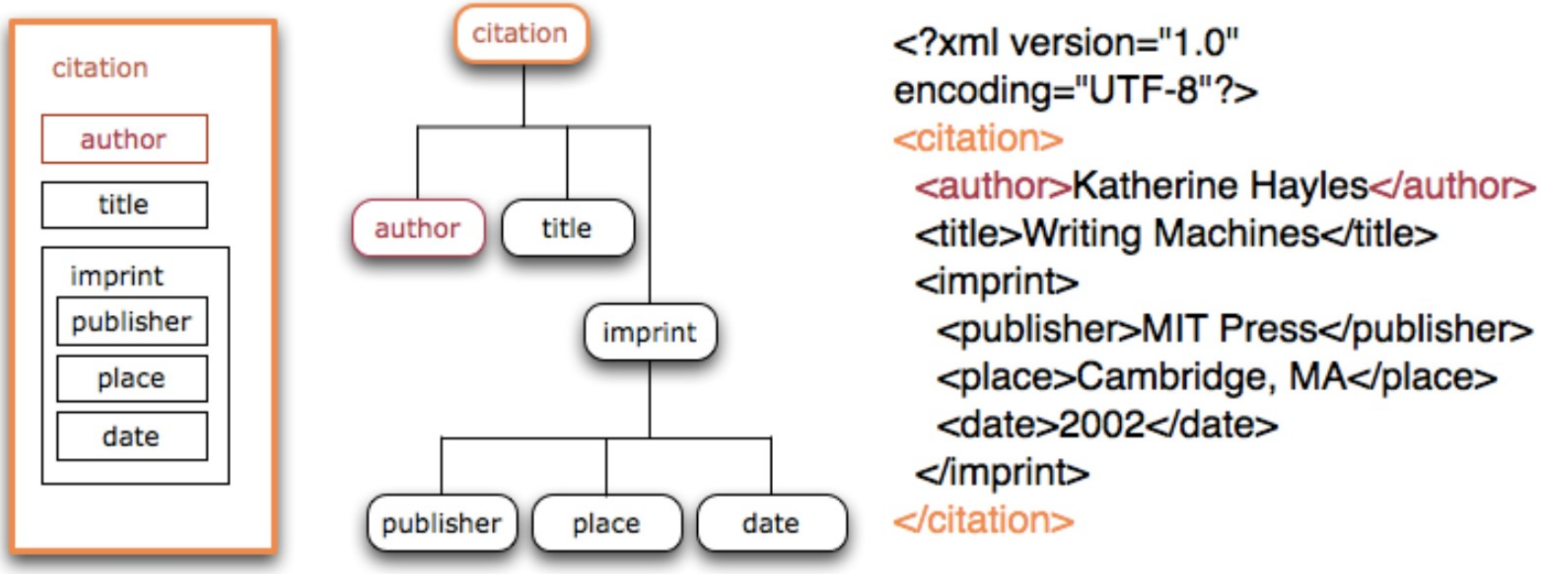
```
<label><hi rend="italic">b</hi>.</label>
<item>Specialists and their families brought to the
    United States hereunder will remain under temporary,
    limited military custody until visas are granted or
    repatriation is accomplished. <list>
      <label>(1)</label>
      <item>Upon arrival of specialists or families in
      the United States, the War Department will screen,
      and cause to be prepared complete biographical and
      professional data on all such persons, copies to
      be supplied to the <gloss target="#t_FBI1"
      >FBI</gloss>, <gloss target="#t_JIOA1"
      >JIOA</gloss>, and the technical service of the
```

Encoded text

Image credit: Foreign Relations of the U.S.

Introduction to XML

XML structures



```
<?xml version="1.0"
encoding="UTF-8"?>
<citation>
 <author>Katherine Hayles</author>
 <title>Writing Machines</title>
 <imprint>
  <publisher>MIT Press</publisher>
  <place>Cambridge, MA</place>
  <date>2002</date>
 </imprint>
</citation>
```

XML introductory slides from this presentation by the Women Writers Project

How do XML and the TEI fit in?

Concepts

footnote

paragraph

heading

XML

Syntax

```
<element>
        <element attribute="value">
              content
        </element>
</element>
```

TEI

**Language:
vocabulary and grammar**

```
<p>

<note type="foot">

<head>
```

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>
```

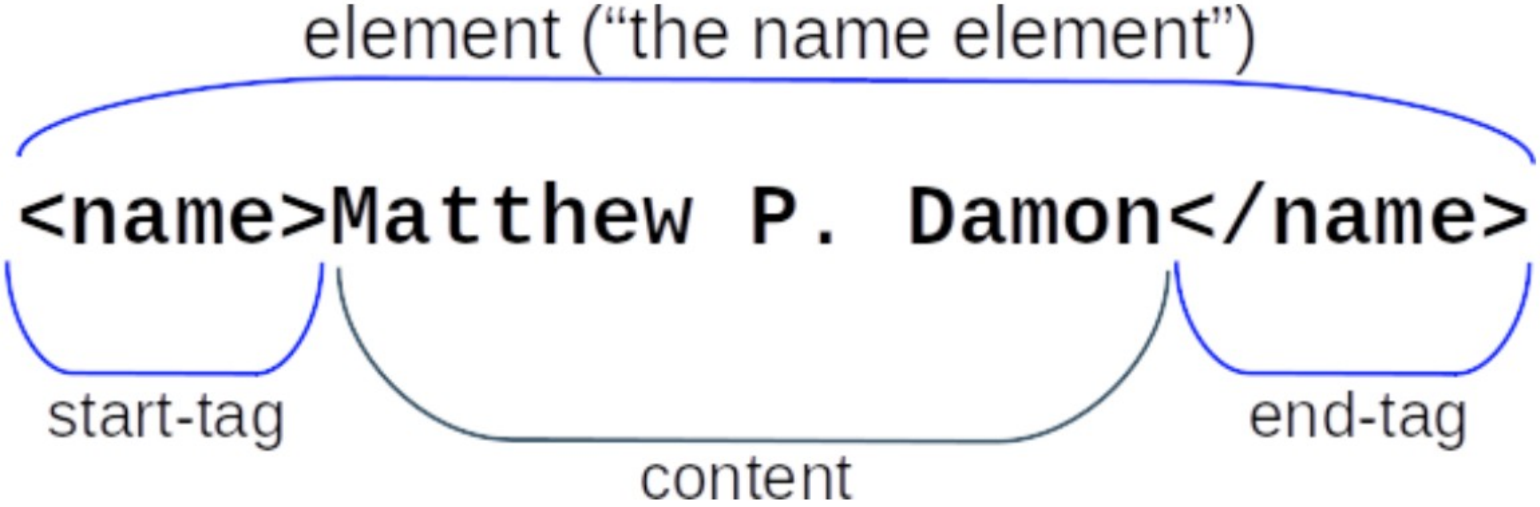
- end-tags by </ … >

```
</heading>
```

- special case: short-hand for an element with no content

```
<anchor/> = <anchor></anchor>
```

Example element



element (“the name element”)

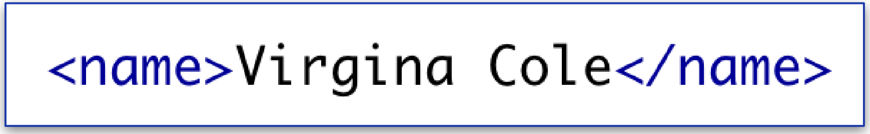
```
<name>Matthew P. Damon</name>
```

start-tag

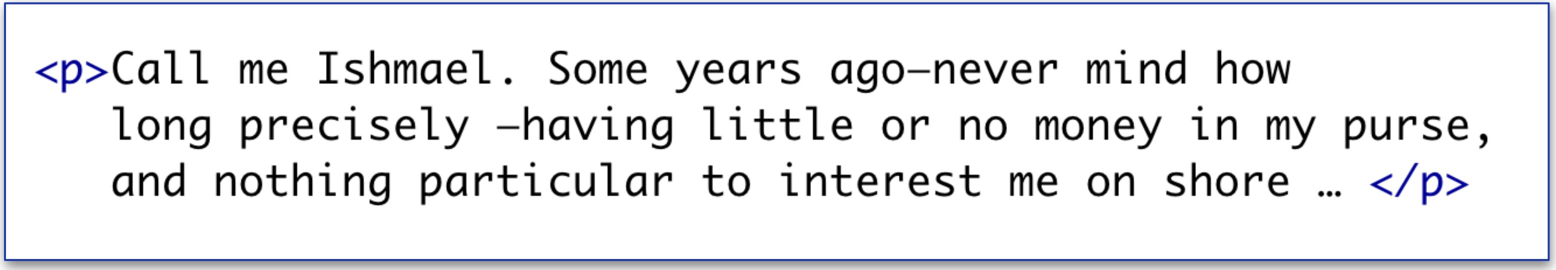
content

end-tag

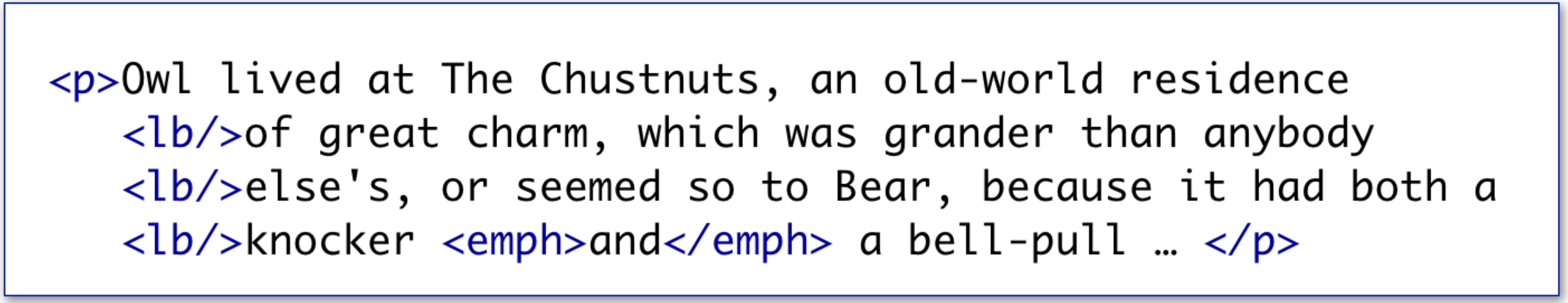
Example elements



```
<name>Virgina Cole</name>
```

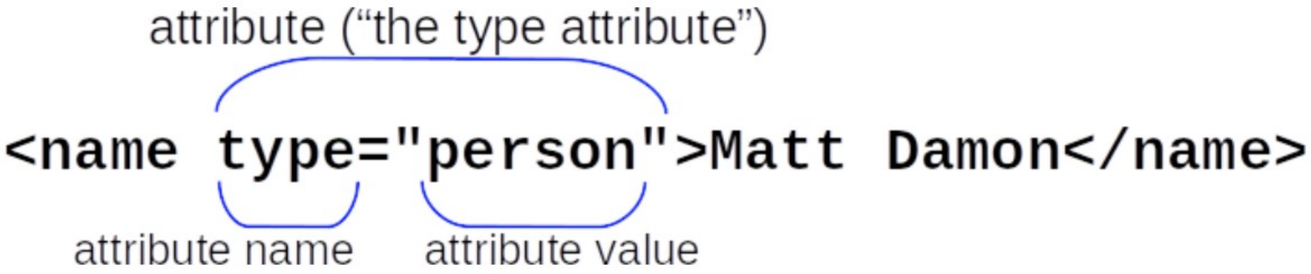


```
<p>Call me Ishmael. Some years ago–never mind how
   long precisely –having little or no money in my purse,
   and nothing particular to interest me on shore … </p>
```



```
<p>Owl lived at The Chustnuts, an old-world residence
   <lb/>of great charm, which was grander than anybody
   <lb/>else's, or seemed so to Bear, because it had both a
   <lb/>knocker <emph>and</emph> a bell-pull … </p>
```

XML attributes



attribute (“the type attribute”)

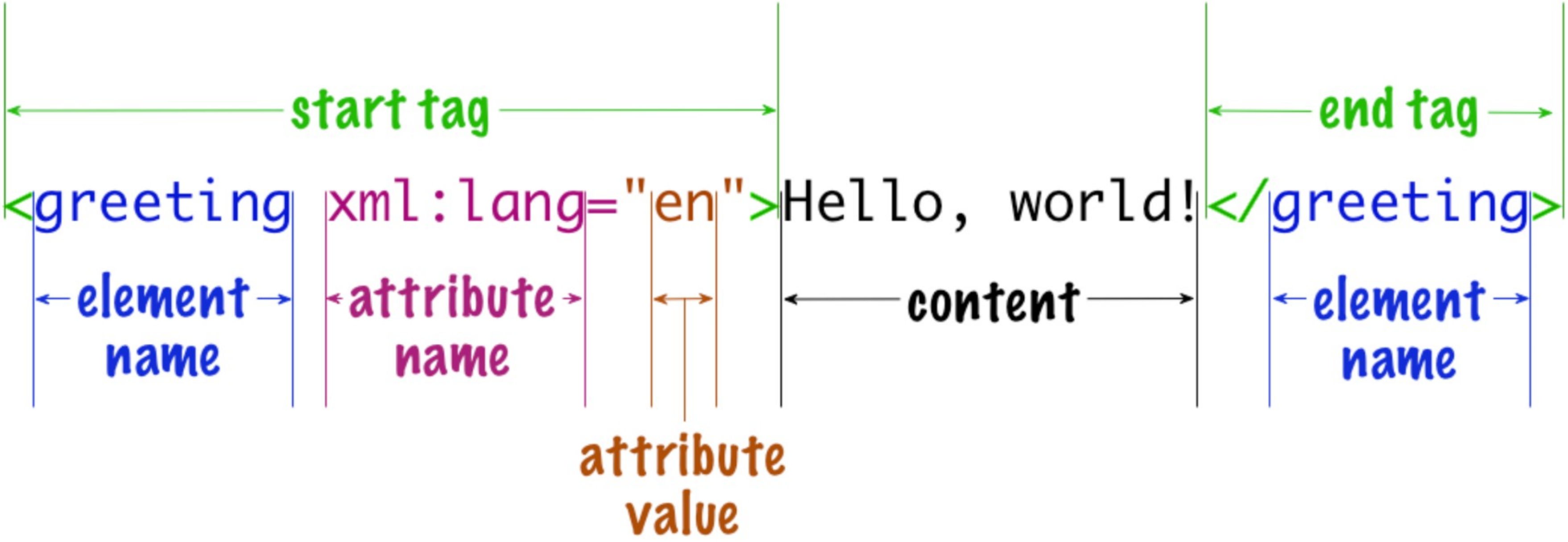
```
<name type="person">Matt Damon</name>
```

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

Anatomy of an element



AI in Oxygen XML

Oxygen XML’s AI Positron has multiple features:

- Generation of code, alt-text, overviews, translations, and more based on the elements used within the document.
- Reviewing code syntax, structure; and provide code annotations
- A chat-bot to mimic human discussion and provide output
- Conversion to other markup languages
- Rewriting code to optimize marketing and search engine outputs

AI Impacts and Workarounds

Some of the many, harmful impacts of AI:

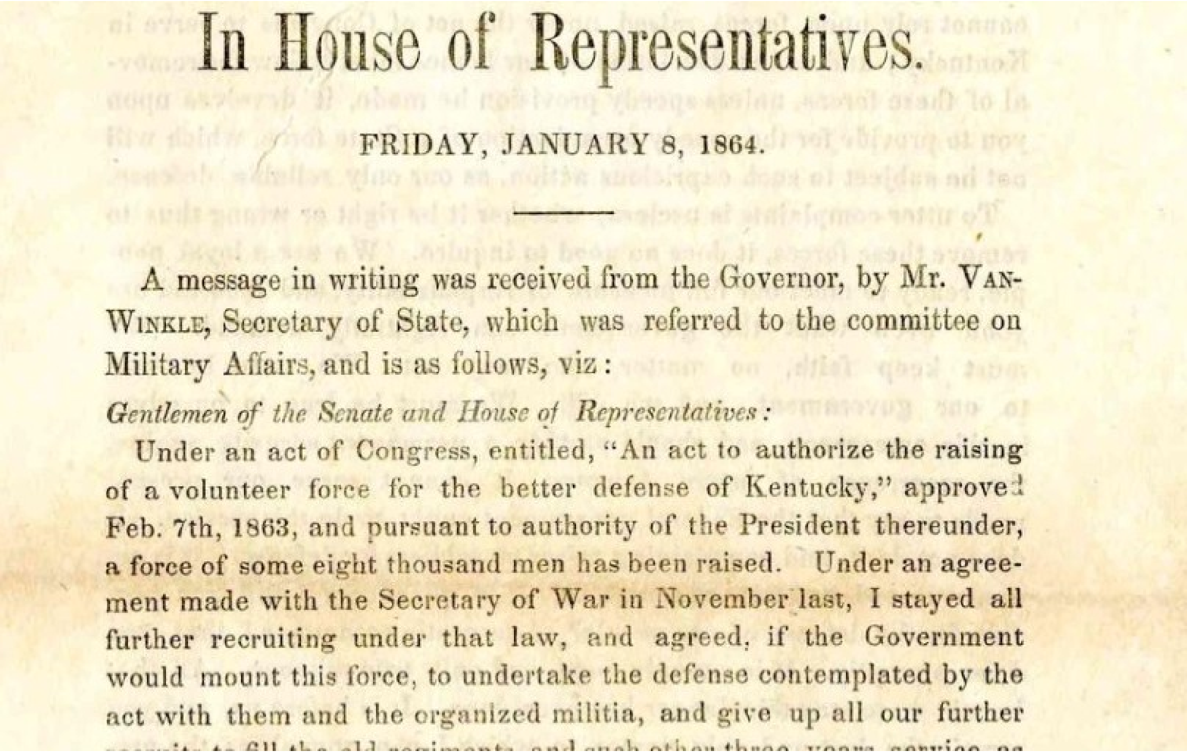
- Environmental damage: high energy, land, and water use
- Critical thinking and psychological damage on users
- Data surveillance of users
- User liability of violating copyright laws
- Issue of companies collecting personal data for profit
- Biased and hallucinated outputs
- Many more

Workarounds for AI in Oxygen XML:

- Note: workarounds reduce, but do not eliminate, harmful impacts of AI. It may be simpler to avoid AI use.
- Pre-plan the number of prompts and design prompts for unbiased output
- Check AI output with human

Introduction to TEI

Sample text fragment



In House of Representatives.

FRIDAY, JANUARY 8, 1864.

A message in writing was received from the Governor, by Mr. VanWinkle, Secretary of State, which was referred to the committee on Military Affairs, and is as follows, viz:

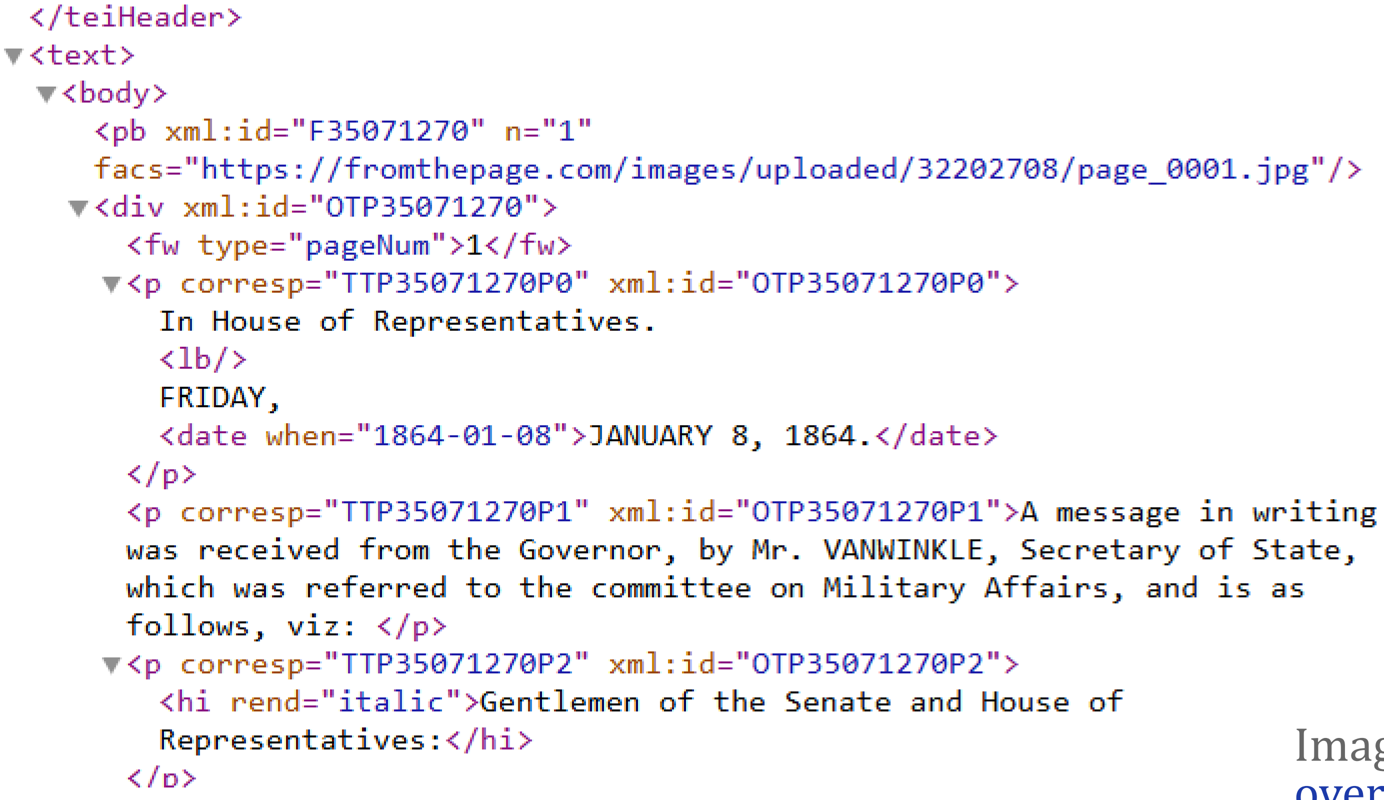
Gentlemen of the Senate and House of Representatives:

Under an act of Congress, entitled, "An act to authorize the raising of a volunteer force for the better defense of Kentucky," approved Feb. 7th, 1863, and pursuant to authority of the President thereunder, a force of some eight thousand men has been raised. Under an agreement made with the Secretary of War in November last, I stayed all further recruiting under that law, and agreed, if the Government would mount this force, to undertake the defense contemplated by the act with them and the organized militia, and give up all our further

Image credit: Civil War Governors of Kentucky

TEI introduction slides from this presentation by the WWP

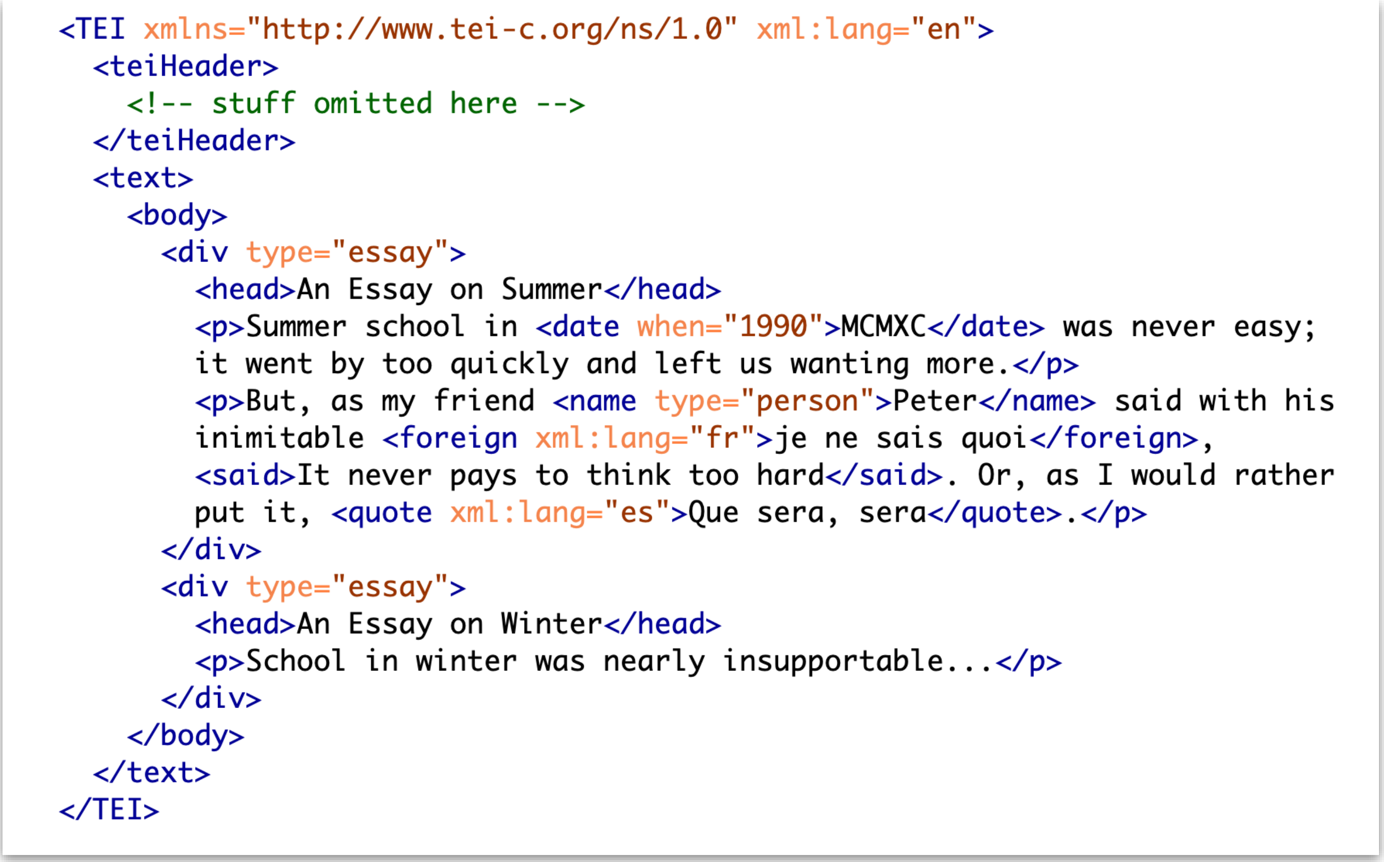
Basic encoding: TEI header



```
 </teiHeader>
<text>
  <body>
    <pb xml:id="F35071270" n="1"
    facs="https://fromthepage.com/images/uploaded/32202708/page_0001.jpg"/>
   <div xml:id="OTP35071270">
      <fw type="pageNum">1</fw>
     <p corresp="TTP35071270P0" xml:id="OTP35071270P0">
        In House of Representatives.
        <lb/>
        FRIDAY,
        <date when="1864-01-08">JANUARY 8, 1864.</date>
      </p>
      <p corresp="TTP35071270P1" xml:id="OTP35071270P1">A message in writing
      was received from the Governor, by Mr. VANWINKLE, Secretary of State,
      which was referred to the committee on Military Affairs, and is as
      follows, viz: </p>
     <p corresp="TTP35071270P2" xml:id="OTP35071270P2">
        <hi rend="italic">Gentlemen of the Senate and House of
        Representatives:</hi>
      </p>
```

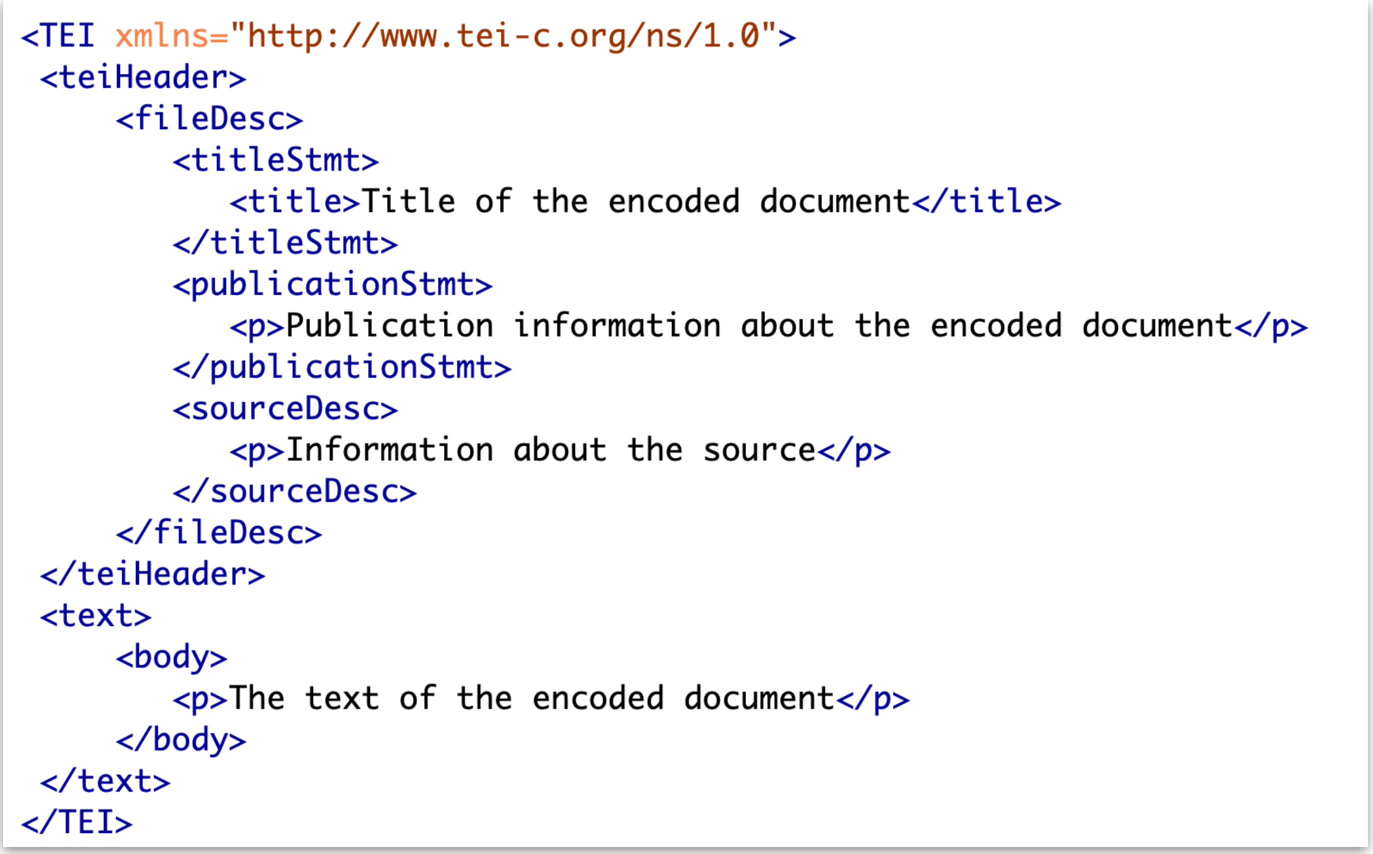
Image credit: Civil War Governors of Kentucky

The bigger picture



```
<TEI xmlns="http://www.tei-c.org/ns/1.0" xml:lang="en">
  <teiHeader>
    <!-- stuff omitted here -->
  </teiHeader>
  <text>
    <body>
      <div type="essay">
        <head>An Essay on Summer</head>
        <p>Summer school in <date when="1990">MCMXC</date> was never easy;
        it went by too quickly and left us wanting more.</p>
        <p>But, as my friend <name type="person">Peter</name> said with his
        inimitable <foreign xml:lang="fr">je ne sais quoi</foreign>,
        <said>It never pays to think too hard</said>. Or, as I would rather
        put it, <quote xml:lang="es">Que sera, sera</quote>.</p>
      </div>
      <div type="essay">
        <head>An Essay on Winter</head>
        <p>School in winter was nearly insupportable...</p>
      </div>
    </body>
  </text>
</TEI>
```

<teiHeader>: metadata



```
<TEI xmlns="http://www.tei-c.org/ns/1.0">
 <teiHeader>
     <fileDesc>
        <titleStmt>
           <title>Title of the encoded document</title>
        </titleStmt>
        <publicationStmt>
           <p>Publication information about the encoded document</p>
        </publicationStmt>
        <sourceDesc>
           <p>Information about the source</p>
        </sourceDesc>
     </fileDesc>
 </teiHeader>
 <text>
     <body>
        <p>The text of the encoded document</p>
     </body>
 </text>
</TEI>
```

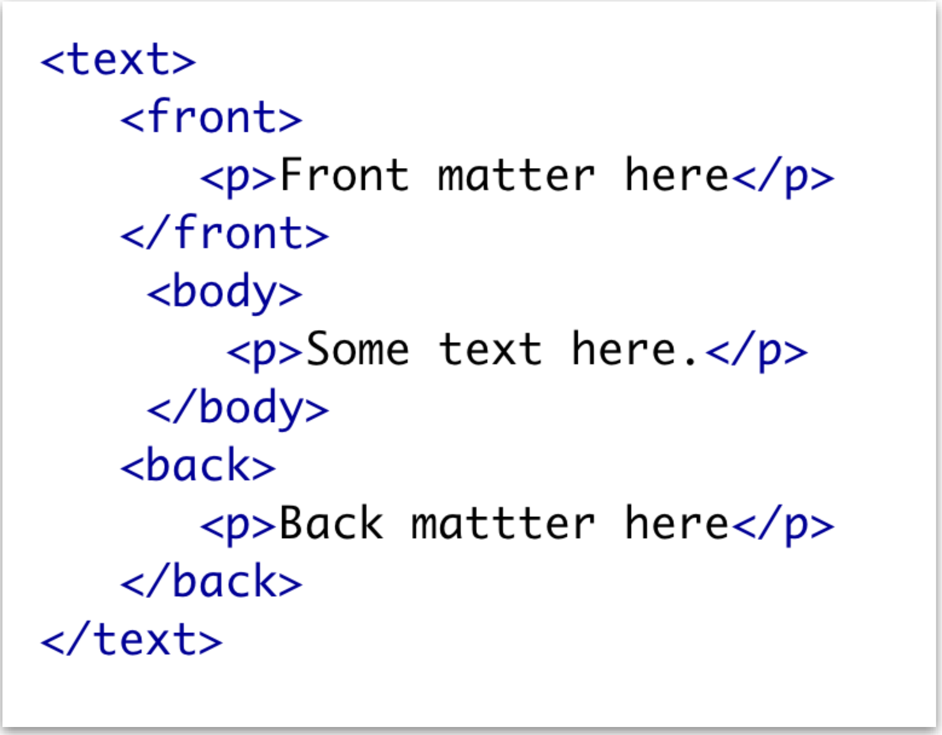
<text> and its contents

<text> contains the transcription of the source text

<front> contains front matter (prefaces, dedications, etc.)

<body> contains the body of the text

<back> contains back matter (indexes, afterwards, appendices, etc.)



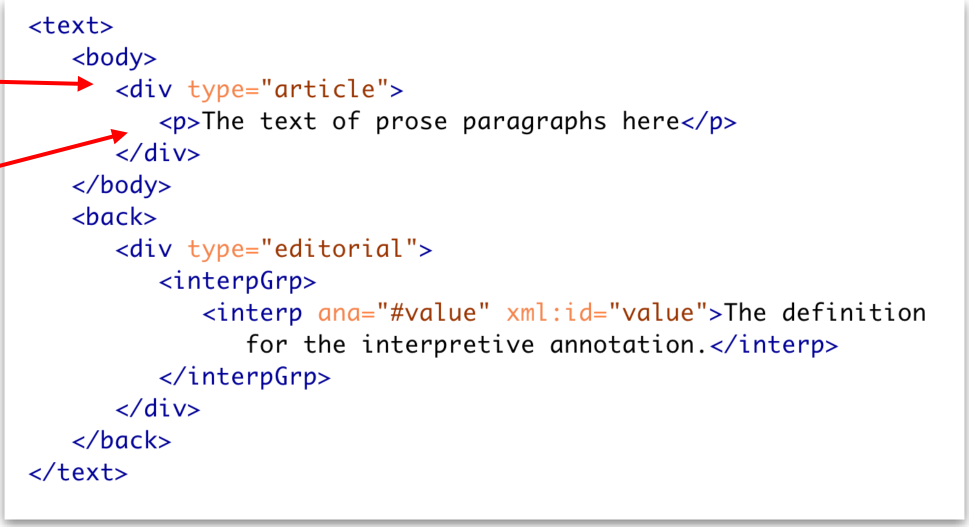
```
<text>
   <front>
      <p>Front matter here</p>
   </front>
    <body>
       <p>Some text here.</p>
    </body>
   <back>
      <p>Back matter here</p>
   </back>
</text>
```

<div> and <p>

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

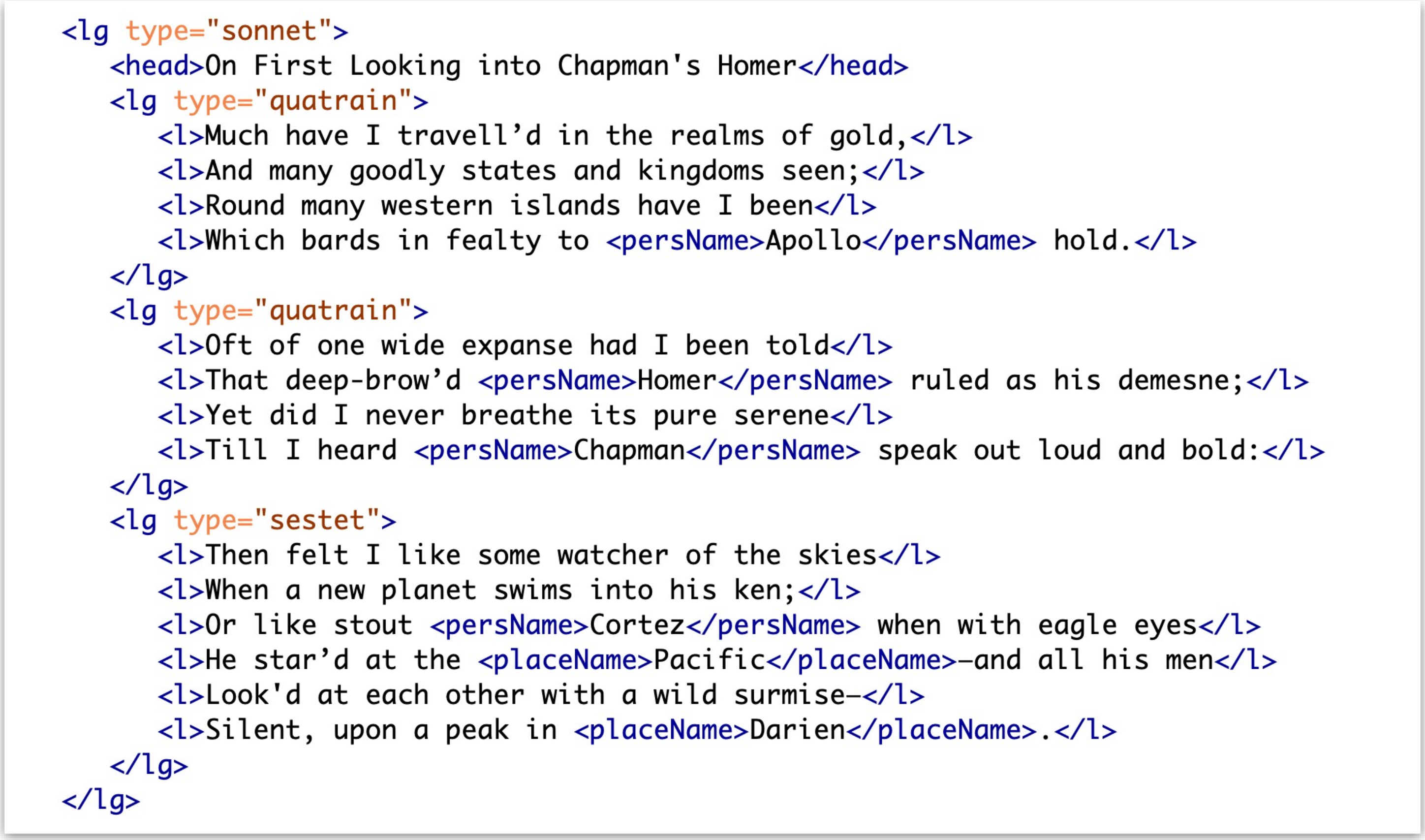
<p> marks paragraphs of prose

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



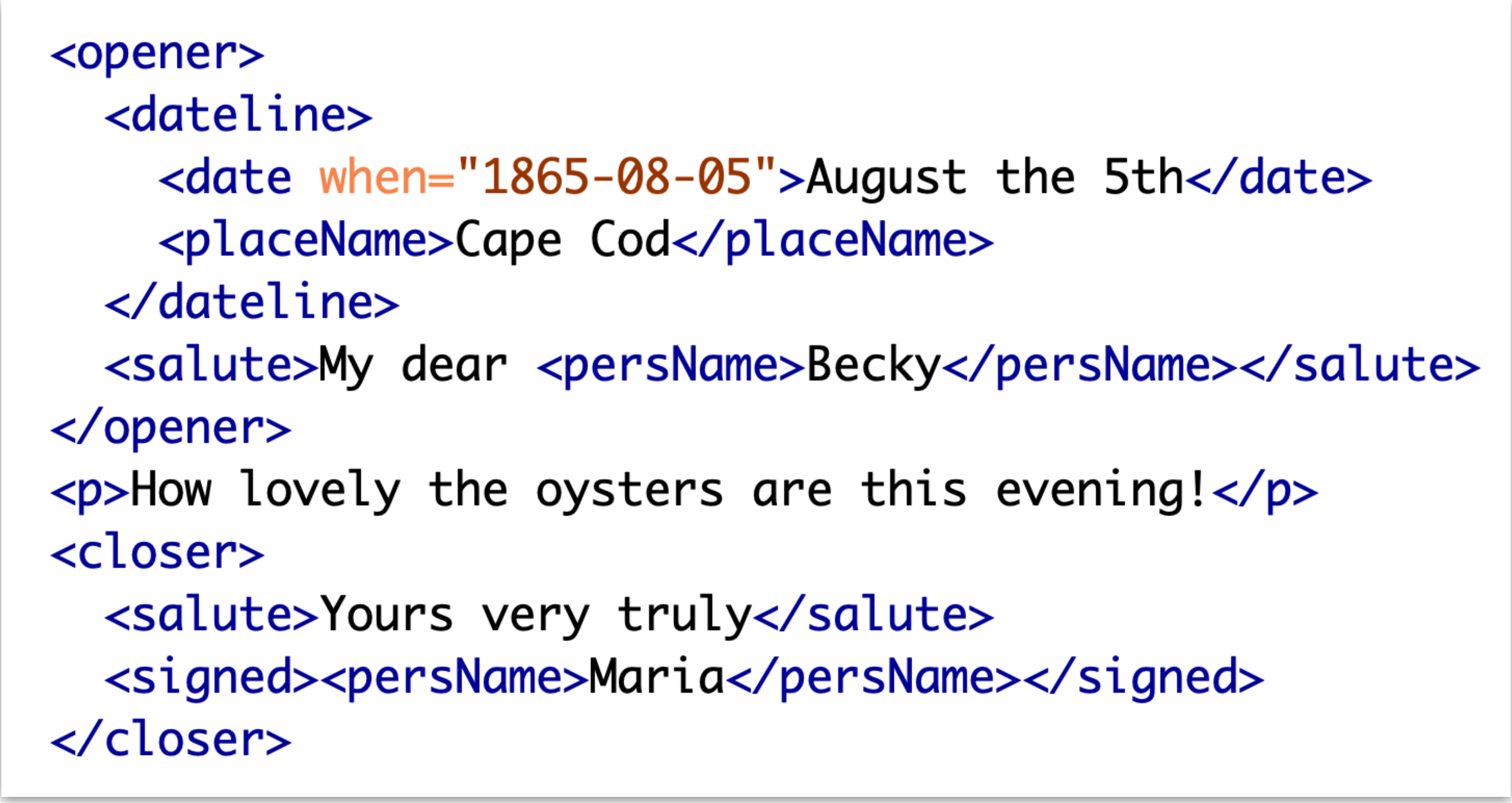
```
<text>
   <body>
      <div type="article">
         <p>The text of prose paragraphs here</p>
      </div>
   </body>
   <back>
      <div type="editorial">
         <interpGrp>
            <interp ana="#value" xml:id="value">The definition
               for the interpretive annotation.</interp>
         </interpGrp>
      </div>
   </back>
</text>
```

Data structures: Encoding poetry



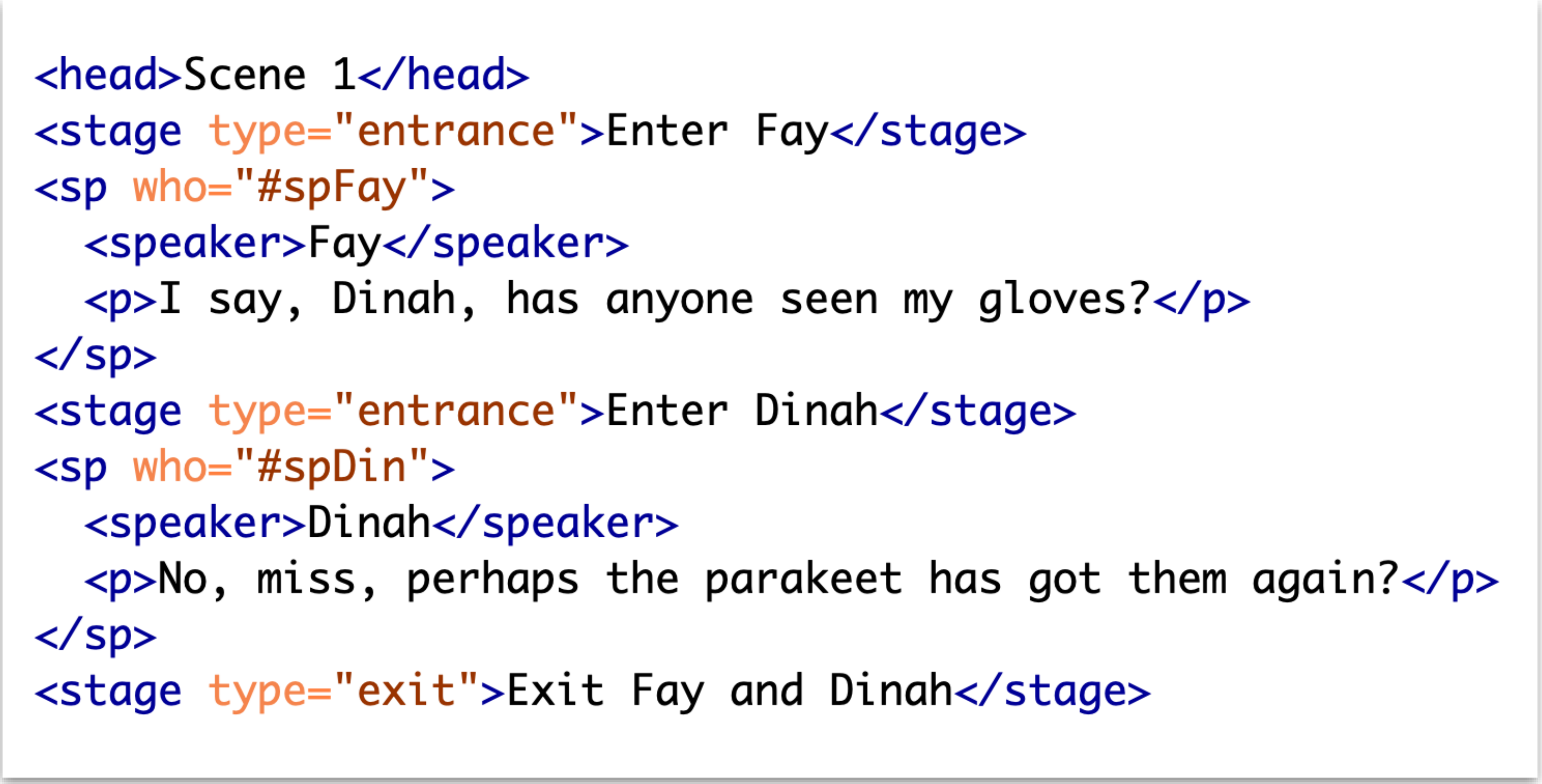
```
<lg type="sonnet">
   <head>On First Looking into Chapman's Homer</head>
   <lg type="quatrain">
      <l>Much have I travell’d in the realms of gold,</l>
      <l>And many goodly states and kingdoms seen;</l>
      <l>Round many western islands have I been</l>
      <l>Which bards in fealty to <persName>Apollo</persName> hold.</l>
   </lg>
   <lg type="quatrain">
      <l>Oft of one wide expanse had I been told</l>
      <l>That deep-brow’d <persName>Homer</persName> ruled as his demesne;</l>
      <l>Yet did I never breathe its pure serene</l>
      <l>Till I heard <persName>Chapman</persName> speak out loud and bold:</l>
   </lg>
   <lg type="sestet">
      <l>Then felt I like some watcher of the skies</l>
      <l>When a new planet swims into his ken;</l>
      <l>Or like stout <persName>Cortez</persName> when with eagle eyes</l>
      <l>He star’d at the <placeName>Pacific</placeName>–and all his men</l>
      <l>Look'd at each other with a wild surmise–</l>
      <l>Silent, upon a peak in <placeName>Darien</placeName>.</l>
   </lg>
</lg>
```

Encoding letters



```
<opener>
  <dateline>
    <date when="1865-08-05">August the 5th</date>
    <placeName>Cape Cod</placeName>
  </dateline>
  <salute>My dear <persName>Becky</persName></salute>
</opener>
<p>How lovely the oysters are this evening!</p>
<closer>
  <salute>Yours very truly</salute>
  <signed><persName>Maria</persName></signed>
</closer>
```

Encoding drama

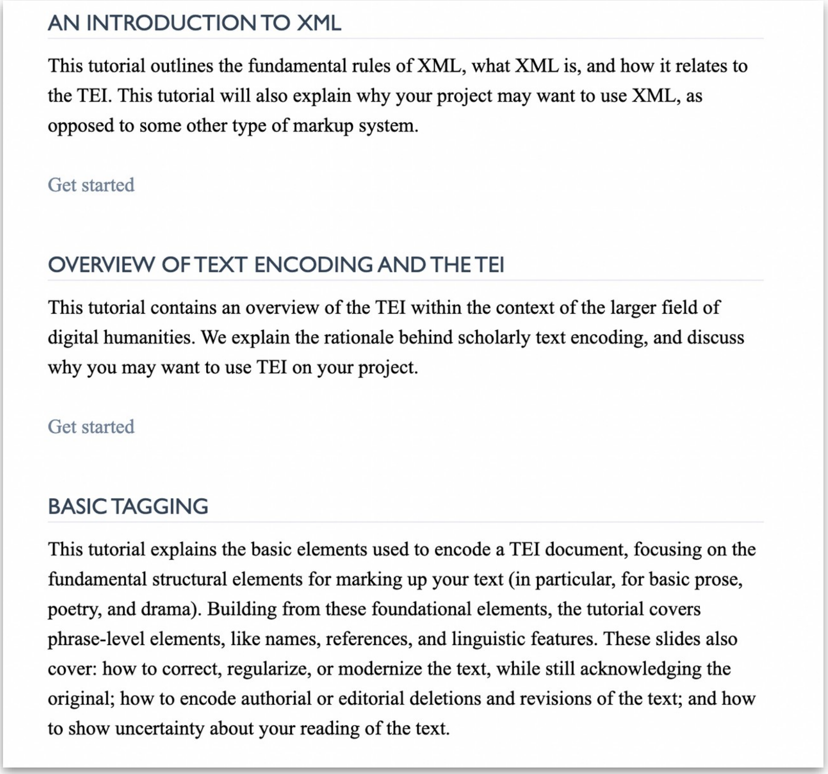


```
<head>Scene 1</head>
<stage type="entrance">Enter Fay</stage>
<sp who="#spFay">
  <speaker>Fay</speaker>
  <p>I say, Dinah, has anyone seen my gloves?</p>
</sp>
<stage type="entrance">Enter Dinah</stage>
<sp who="#spDin">
  <speaker>Dinah</speaker>
  <p>No, miss, perhaps the parakeet has got them again?</p>
</sp>
<stage type="exit">Exit Fay and Dinah</stage>
```

Getting started with encoding

WWP Tutorials

The WWP provides a set of tutorials that cover the concepts we have introduced today in more detail. The TEI Primer should have all the information you need to get started with encoding. You can also find more information on the WWP's resources page.



AN INTRODUCTION TO XML

This tutorial outlines the fundamental rules of XML, what XML is, and how it relates to the TEI. This tutorial will also explain why your project may want to use XML, as opposed to some other type of markup system.

Get started

OVERVIEW OF TEXT ENCODING AND THE TEI

This tutorial contains an overview of the TEI within the context of the larger field of digital humanities. We explain the rationale behind scholarly text encoding, and discuss why you may want to use TEI on your project.

Get started

BASIC TAGGING

This tutorial explains the basic elements used to encode a TEI document, focusing on the fundamental structural elements for marking up your text (in particular, for basic prose, poetry, and drama). Building from these foundational elements, the tutorial covers phrase-level elements, like names, references, and linguistic features. These slides also cover: how to correct, regularize, or modernize the text, while still acknowledging the original; how to encode authorial or editorial deletions and revisions of the text; and how to show uncertainty about your reading of the text.

Editing TEI documents

XML is expressed in **plain text**, which means that you can use any text editor (including Notepad and TextEdit) to write and edit TEI files. However, it is usually easier to work with an **XML-aware editor**, such as Oxygen XML Editor.

Oxygen offers free 30-day trials, or you can contact Sarah Connell for a license for longer-term use provided through the NU Library.

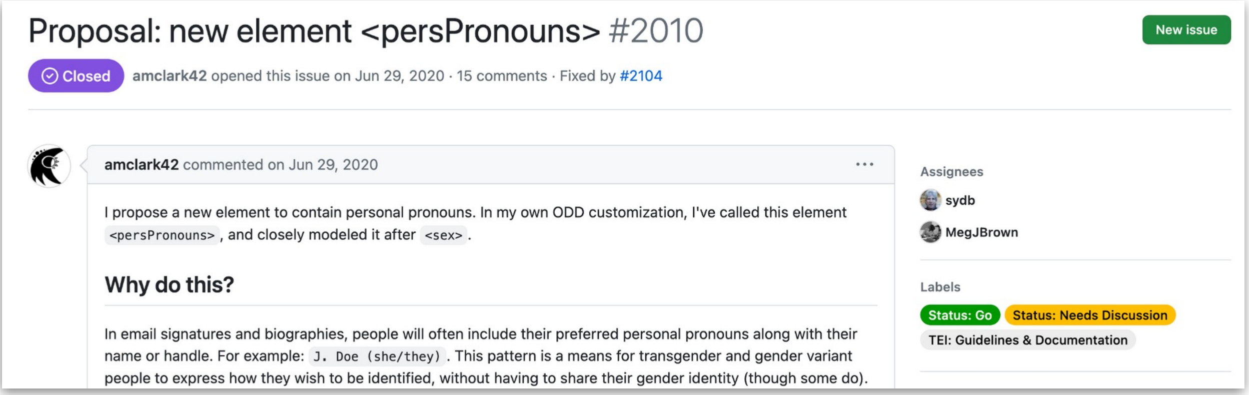
Basic commands in Oxygen

- **To insert a new element:** type a **<** (less-than sign) and choose the element you want from the dropdown list.
- **To insert a new attribute:** with your cursor inside of the element's start tag, just after the name of the element but before the closing **>** character, hit the **spacebar** and choose the attribute you want from the dropdown list.
- If you want to **surround existing text** with element start and end tags: select that text and type **`control-E or command-E`** and pick the element you want from the dropdown list.
- If you want the **text to wrap**: hit **`control/command-shift-Y.`**
- **To add a comment:** type **`<!`** (less-than and then exclamation point) and select the "XML Comment" option from the dropdown.

The TEI community

The TEI is an international standard, developed and contributed to by the TEI community. The TEI consortium publishes the *TEI Guidelines.*

For an example of how the TEI community works, check out this [GitHub issue].



Proposal: new element <persPronouns> #2010

amclark42 opened this issue on Jun 29, 2020 · 15 comments · Fixed by #2104

amclark42 commented on Jun 29, 2020

I propose a new element to contain personal pronouns. In my own ODD customization, I've called this element `<persPronouns>`, and closely modeled it after `<sex>`.

Why do this?

In email signatures and biographies, people will often include their preferred personal pronouns along with their name or handle. For example: `J. Doe (she/they)`. This pattern is a means for transgender and gender variant people to express how they wish to be identified, without having to share their gender identity (though some do).

Assignees

sydb

MegJBrown

Labels

Status: Go

Status: Needs Discussion

TEI: Guidelines & Documentation

Contents of the *TEI Guidelines*

From the main page of the *TEI Guidelines*, you can access:

- Chapters describing different "modules", such as Verse; Names, Dates, People, and Places; and The TEI Header
  - Modules are the major organizing unit for both the *Guidelines* and the TEI elements themselves
- Appendix C: Elements
- Many other resources that you likely won't need for this class

Reading an element entry

The things you will find most useful are:

- The element definition
- What the element is **contained by** and **can contain**
- Any special attributes that belong to the element
- All other attributes that the element can have
- Examples!

Check out the <persName> element entry for an example.

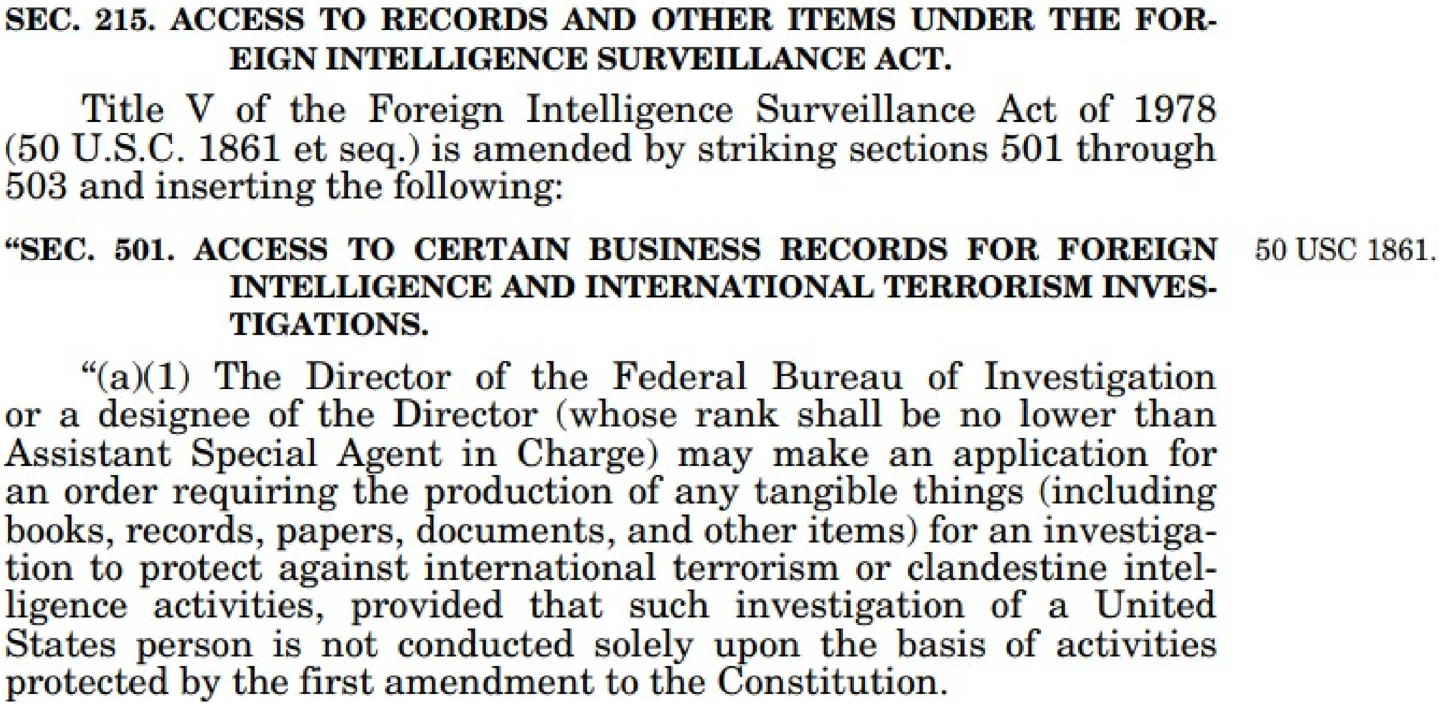
Thought experiment: tagging

Modeling primary sources

Thinking about what we have just learned about **selection, description,** and **formalism**, decide which aspects of our sample text you would want to tag, and how you would describe these.

Sketch out your tagging system. Consider: how would you label the significant aspects of the text? Where do these features start and end? How do you know this? What additional information might you want to model? What **categories** of information are significant to you?

Sample text: US Patriot Act



SEC. 215. ACCESS TO RECORDS AND OTHER ITEMS UNDER THE FOREIGN INTELLIGENCE SURVEILLANCE ACT.

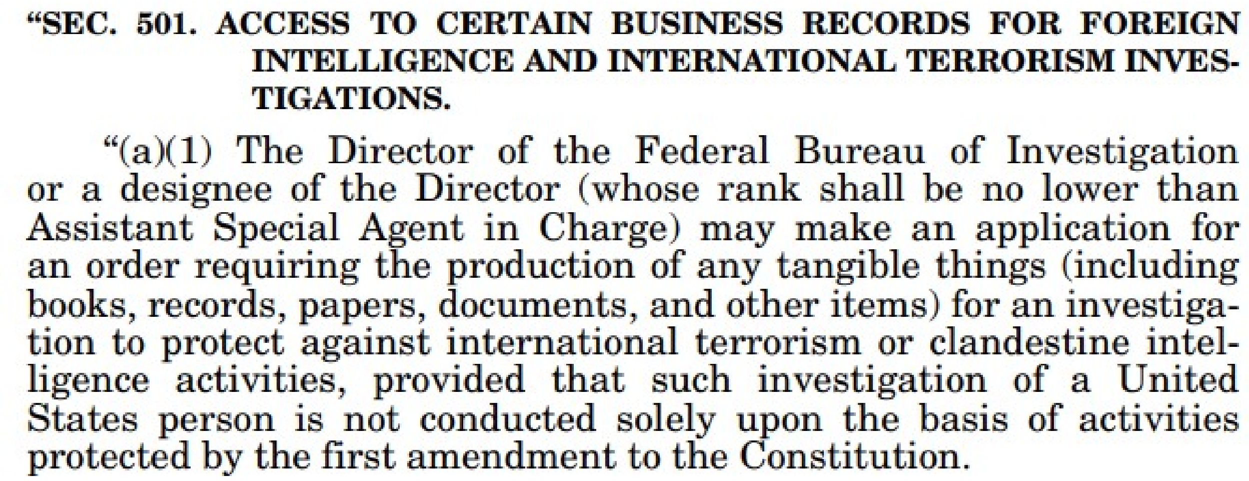
Title V of the Foreign Intelligence Surveillance Act of 1978 (50 U.S.C. 1861 et seq.) is amended by striking sections 501 through 503 and inserting the following:

“SEC. 501. ACCESS TO CERTAIN BUSINESS RECORDS FOR FOREIGN INTELLIGENCE AND INTERNATIONAL TERRORISM INVESTIGATIONS. 50 USC 1861.

“(a)(1) The Director of the Federal Bureau of Investigation or a designee of the Director (whose rank shall be no lower than Assistant Special Agent in Charge) may make an application for an order requiring the production of any tangible things (including books, records, papers, documents, and other items) for an investigation to protect against international terrorism or clandestine intelligence activities, provided that such investigation of a United States person is not conducted solely upon the basis of activities protected by the first amendment to the Constitution.

Image credit: US Congressional record

What do we encode?



“SEC. 501. ACCESS TO CERTAIN BUSINESS RECORDS FOR FOREIGN INTELLIGENCE AND INTERNATIONAL TERRORISM INVESTIGATIONS.

“(a)(1) The Director of the Federal Bureau of Investigation or a designee of the Director (whose rank shall be no lower than Assistant Special Agent in Charge) may make an application for an order requiring the production of any tangible things (including books, records, papers, documents, and other items) for an investigation to protect against international terrorism or clandestine intelligence activities, provided that such investigation of a United States person is not conducted solely upon the basis of activities protected by the first amendment to the Constitution.

- Fill in the blank: section heading
- What other information would you tag?

<tei header>

<text>

<body>

<div n= “501” type= “________”>

<head> ______________________________</head>

Discussion, questions

- What kinds of metadata are important to tag?
- How did you label the significant aspects of the text?
- How did you determine where these features start and end?
- What additional information might you want to model?
- What **categories** of information are significant to your encoding system?
- What questions do you have about encoding?

Thank you!

—**Developed by** Sarah Connell, Claire Lavarreda, Ayah Aboelela, Avery Blankenship, Dipa Desai and Juniper Johnson

- For more information on DITI, please see: https://bit.ly/diti-about
- Schedule an appointment with us! https://bit.ly/diti-meeting
- If you have any questions, contact us at: nulab.info@gmail.com
- We'd love your feedback! Please fill out a short survey here: https://bit.ly/diti-feedback
- Find all of the course materials at: https://bit.ly/sp26-parr-hist1357-multi