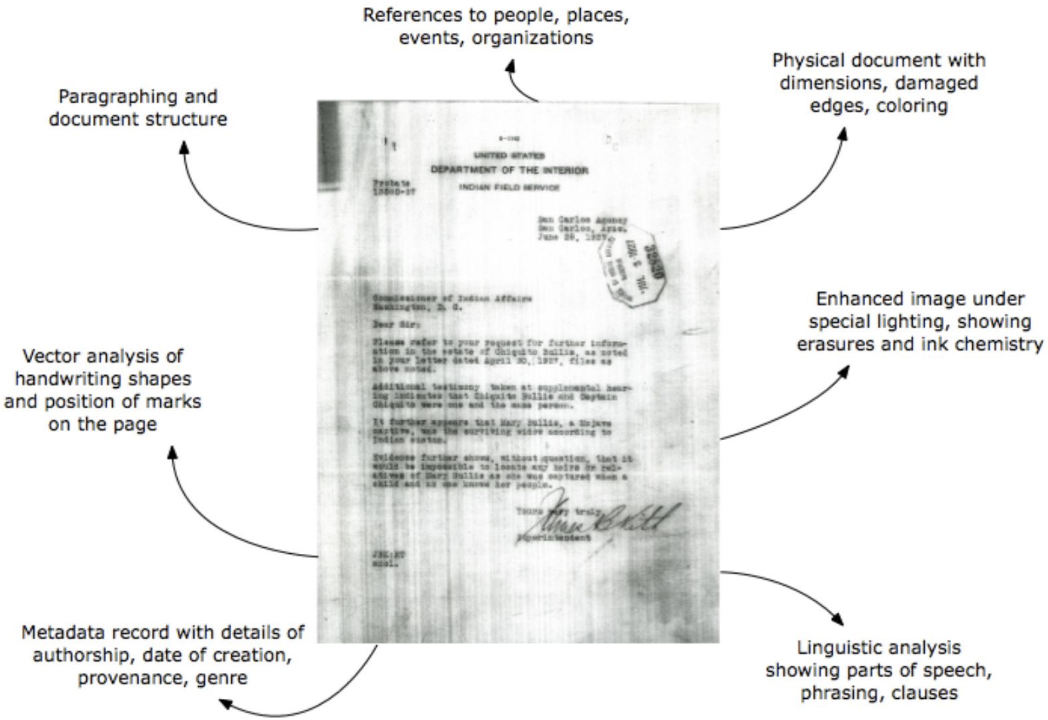
Introduction to Text Encoding

Sarah Connell and Claire Lavarreda
Digital Integration Teaching Initiative

Black Digital Humanities
Nicole Aljoe
https://bit.ly/fa24-aljoe-tei

What is text encoding?

Representing research objects


References to people, places, events, organizations
Physical document with dimensions, damaged edges, coloring
Paragraphing and document structure
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
13800-27

San Carlos Agency
San Carlos, Ariz.
June 20, 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

JBL:ET
encl.

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

Background on the TEI

The TEI is:

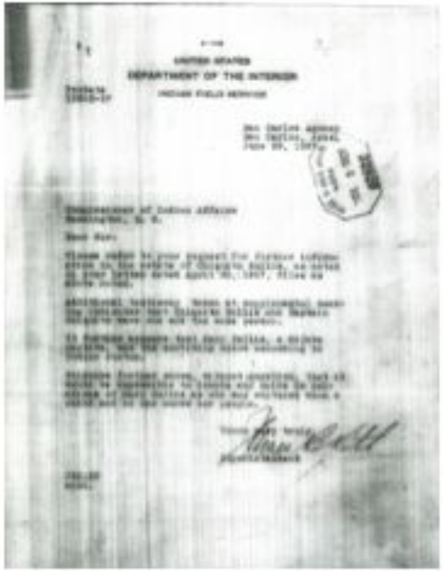
- A markup language (a text encoding language)
- Developed by an international consortium; free and open-source
- Both a community standard and a community research effort

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

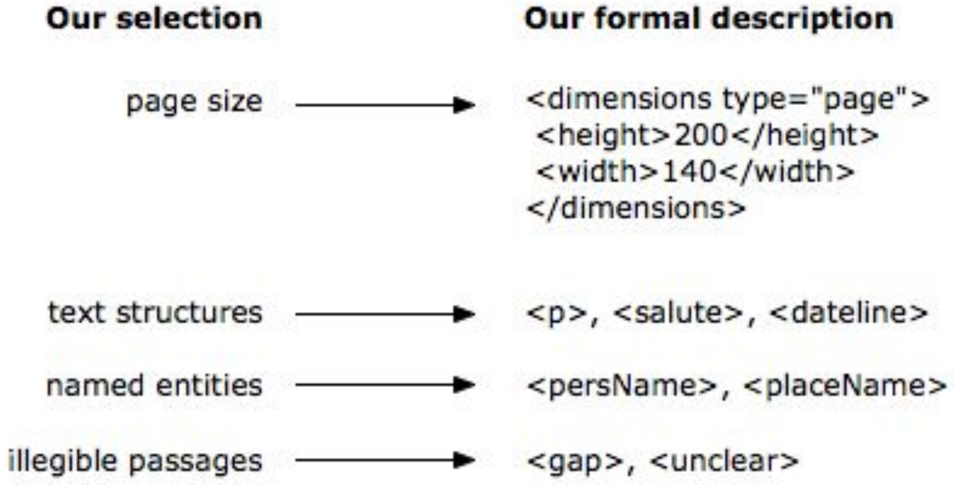
Raw stuff



UNITED STATES
DEPARTMENT OF THE INTERIOR
INDIAN FIELD SERVICE

Commissioner of Indian Affairs
Washington, D. C.

Dear Sir:



Our selection → **Our formal description**

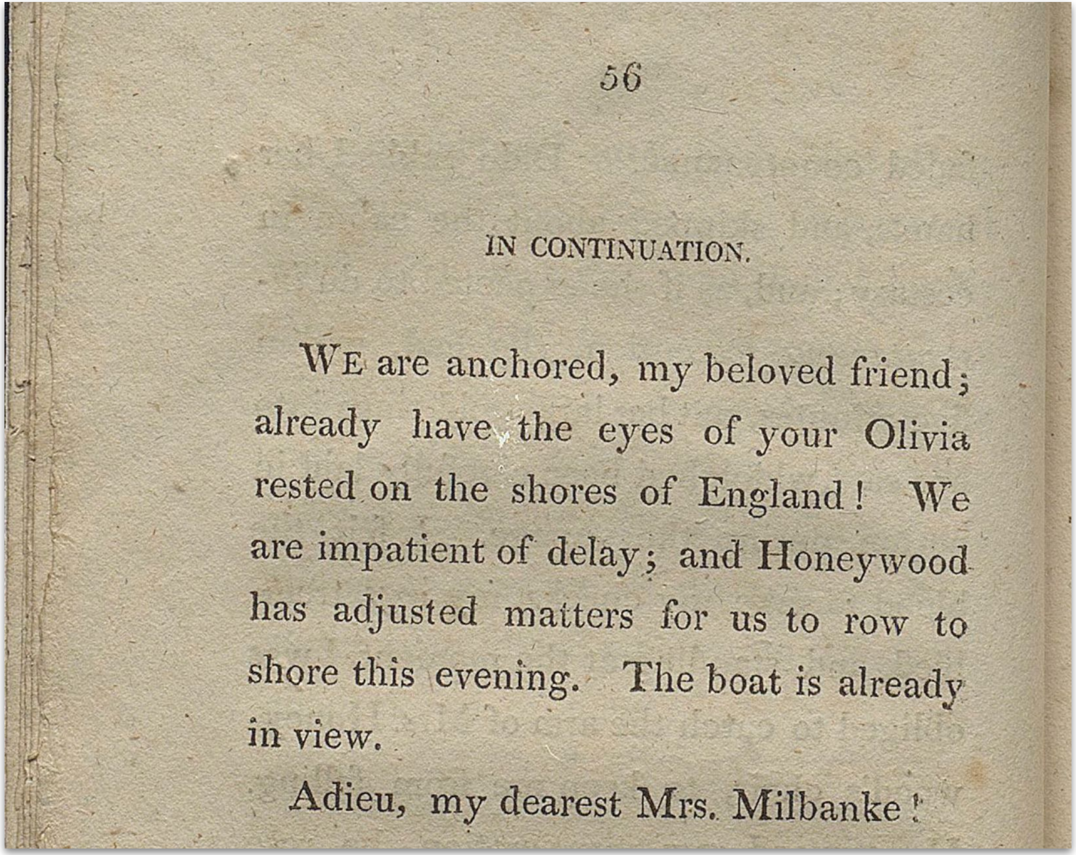
page size → <dimensions type="page">
 <height>200</height>
 <width>140</width>
</dimensions>

text structures → <p>, <salute>, <dateline>

named entities → <persName>, <placeName>

illegible passages → <gap>, <unclear>

Sample text: *The Woman of Colour*



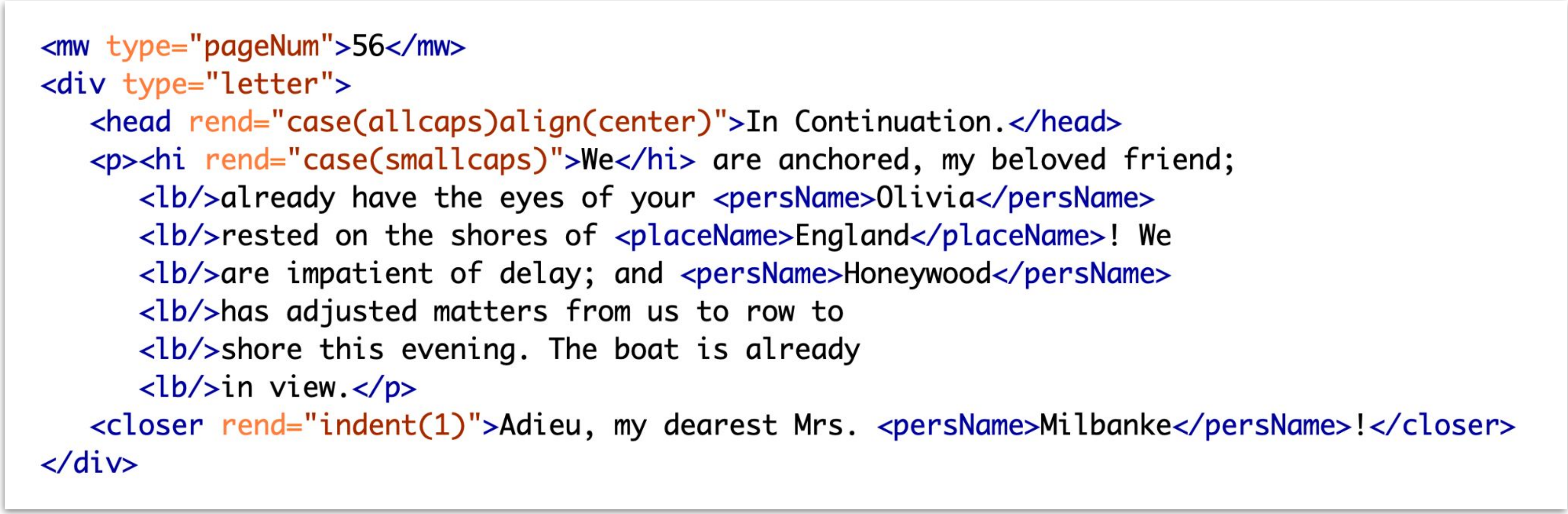
IN CONTINUATION.

WE are anchored, my beloved friend; already have the eyes of your Olivia rested on the shores of England! We are impatient of delay; and Honeywood has adjusted matters for us to row to shore this evening. The boat is already in view.

Adieu, my dearest Mrs. Milbanke!

—Unknown, *The Woman of Colour*, 1808
Image credit: The British Library

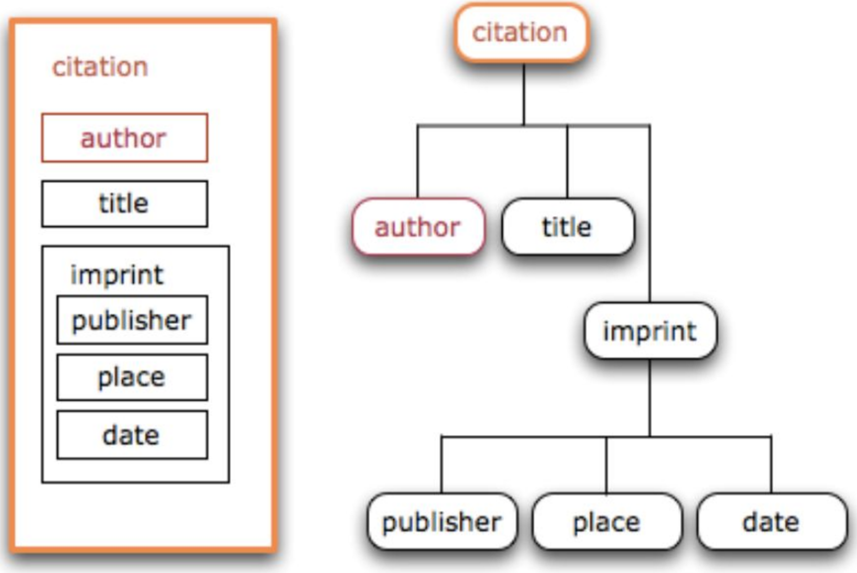
Sample encoding

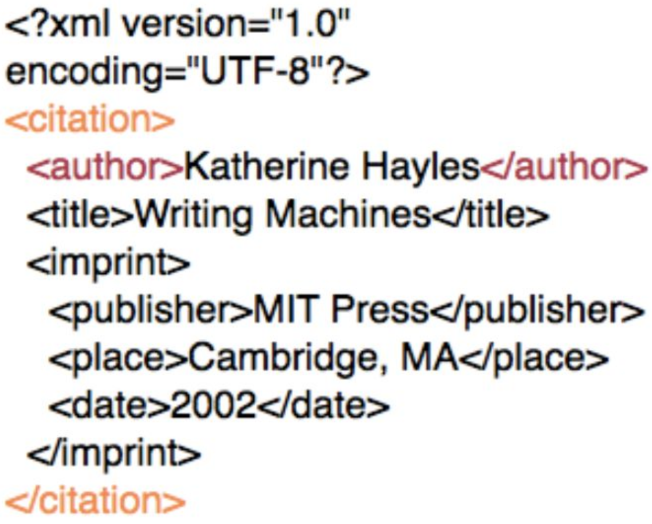


```
<mw type="pageNum">56</mw>
<div type="letter">
   <head rend="case(allcaps)align(center)">In Continuation.</head>
   <p><hi rend="case(smallcaps)">We</hi> are anchored, my beloved friend;
      <lb/>already have the eyes of your <persName>Olivia</persName>
      <lb/>rested on the shores of <placeName>England</placeName>! We
      <lb/>are impatient of delay; and <persName>Honeywood</persName>
      <lb/>has adjusted matters from us to row to
      <lb/>shore this evening. The boat is already
      <lb/>in view.</p>
   <closer rend="indent(1)">Adieu, my dearest Mrs. <persName>Milbanke</persName>!</closer>
</div>
```

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

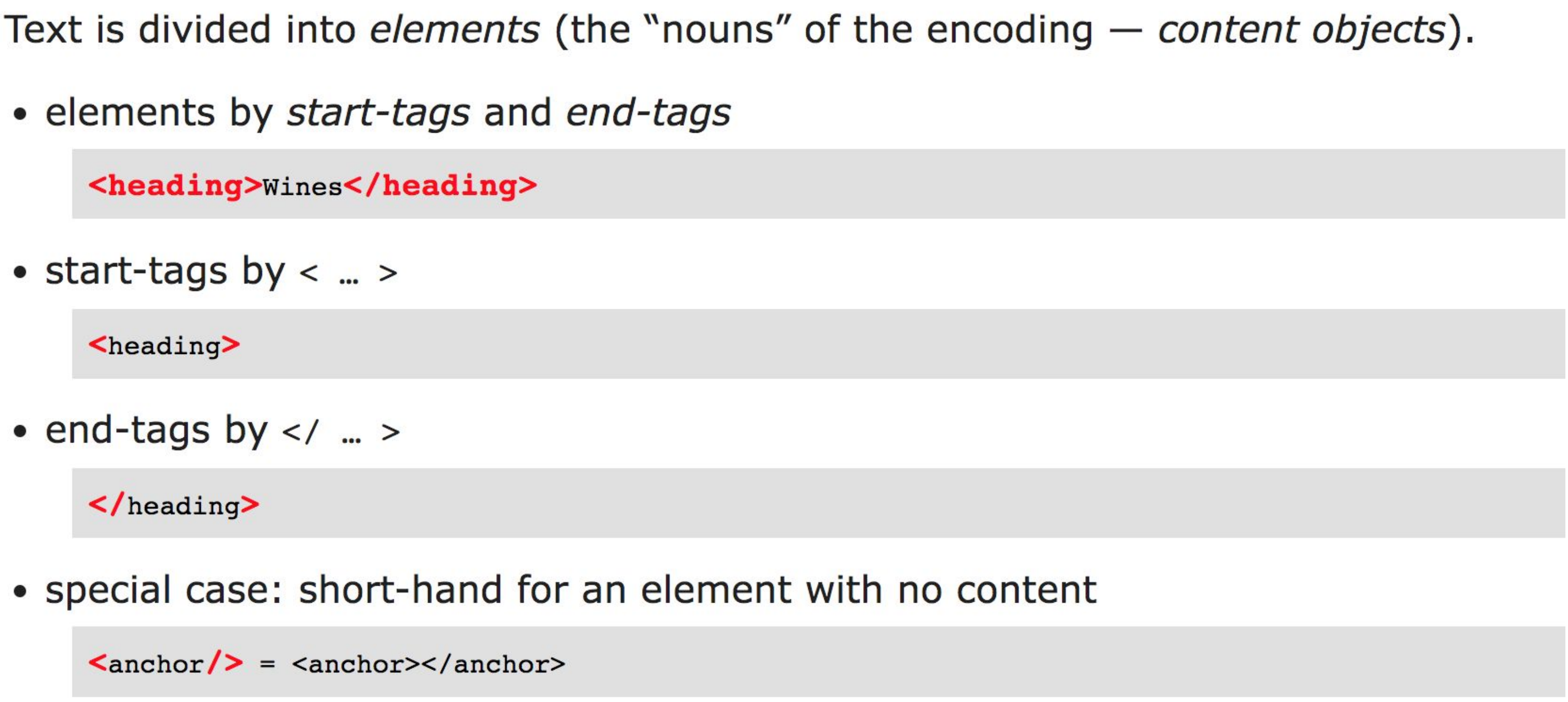
Language:
vocabulary and grammar

```
<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>
```

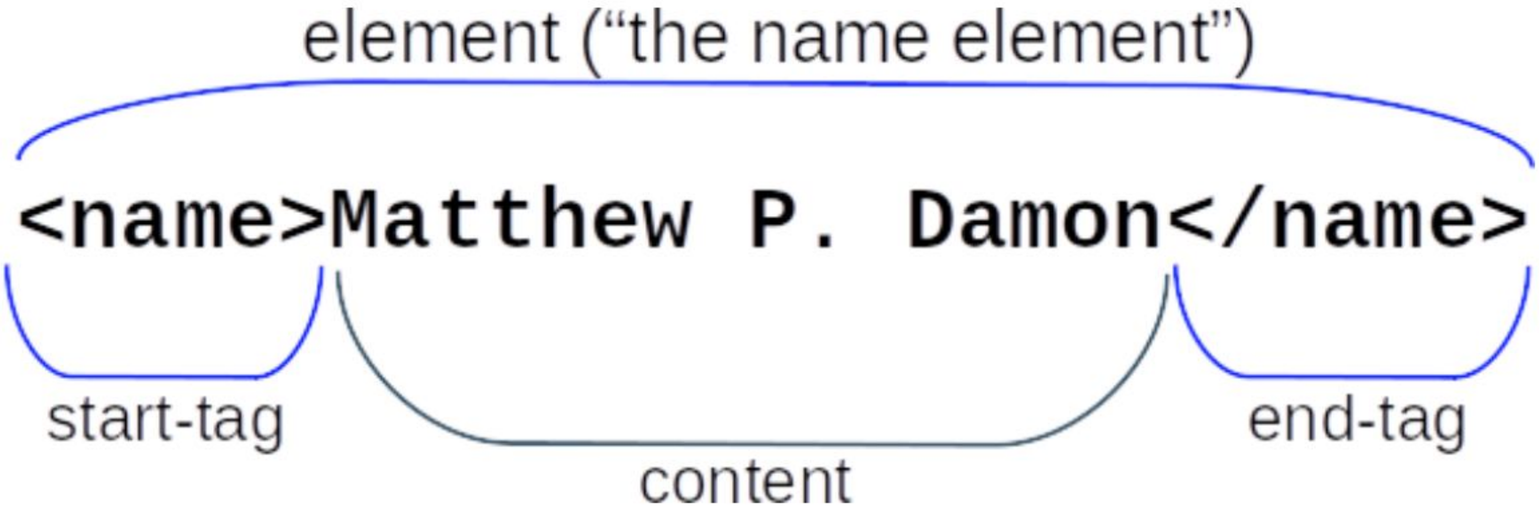
- end-tags by </ … >

```
</heading>
```

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

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

Example element



element (“the name element”)

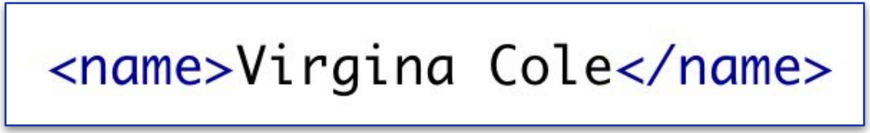
```
<name>Matthew P. Damon</name>
```

start-tag

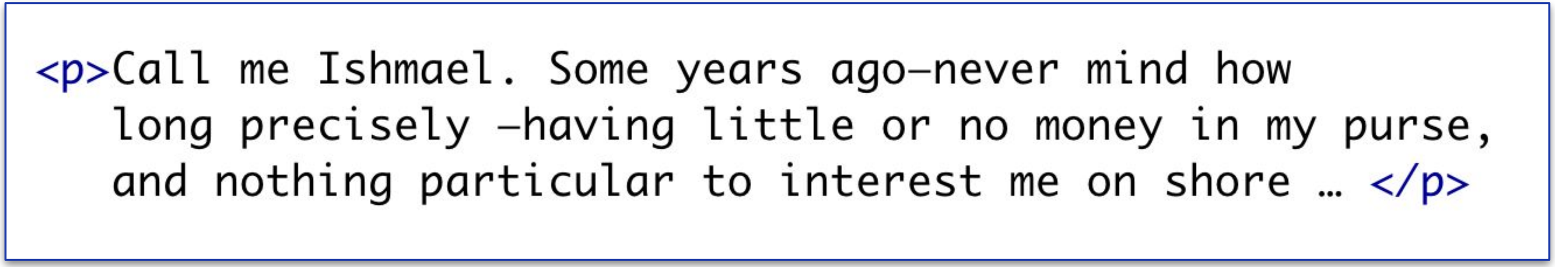
content

end-tag

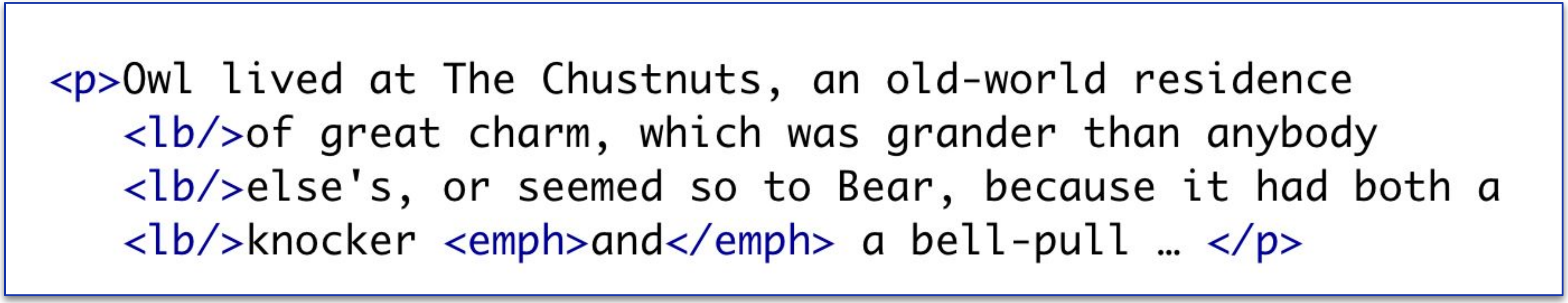
Example elements



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

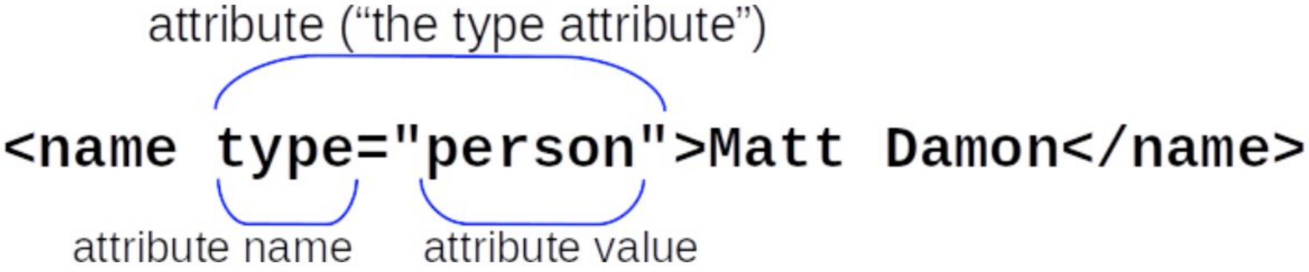


```
<p>Call me Ishmael. Some years ago–never mind how
   long precisely –having little or no money in my purse,
   and nothing particular to interest me on shore … </p>
```



```
<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”)

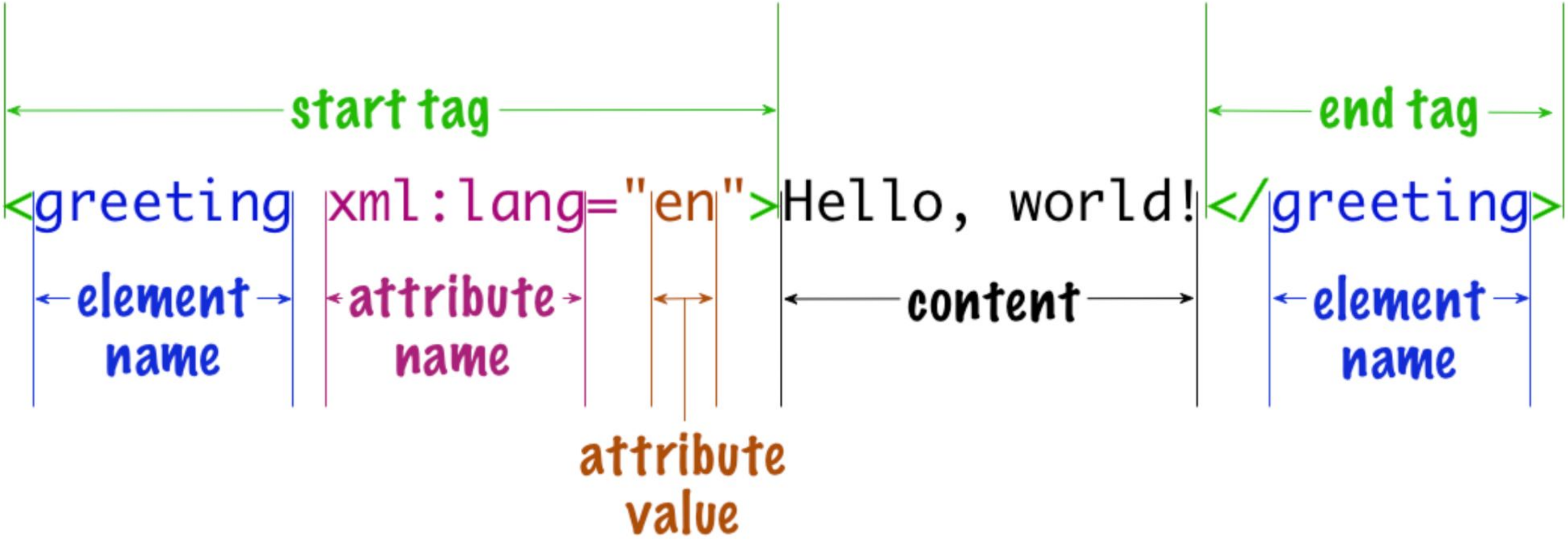
```
<name type="person">Matt Damon</name>
```

attribute name

attribute value

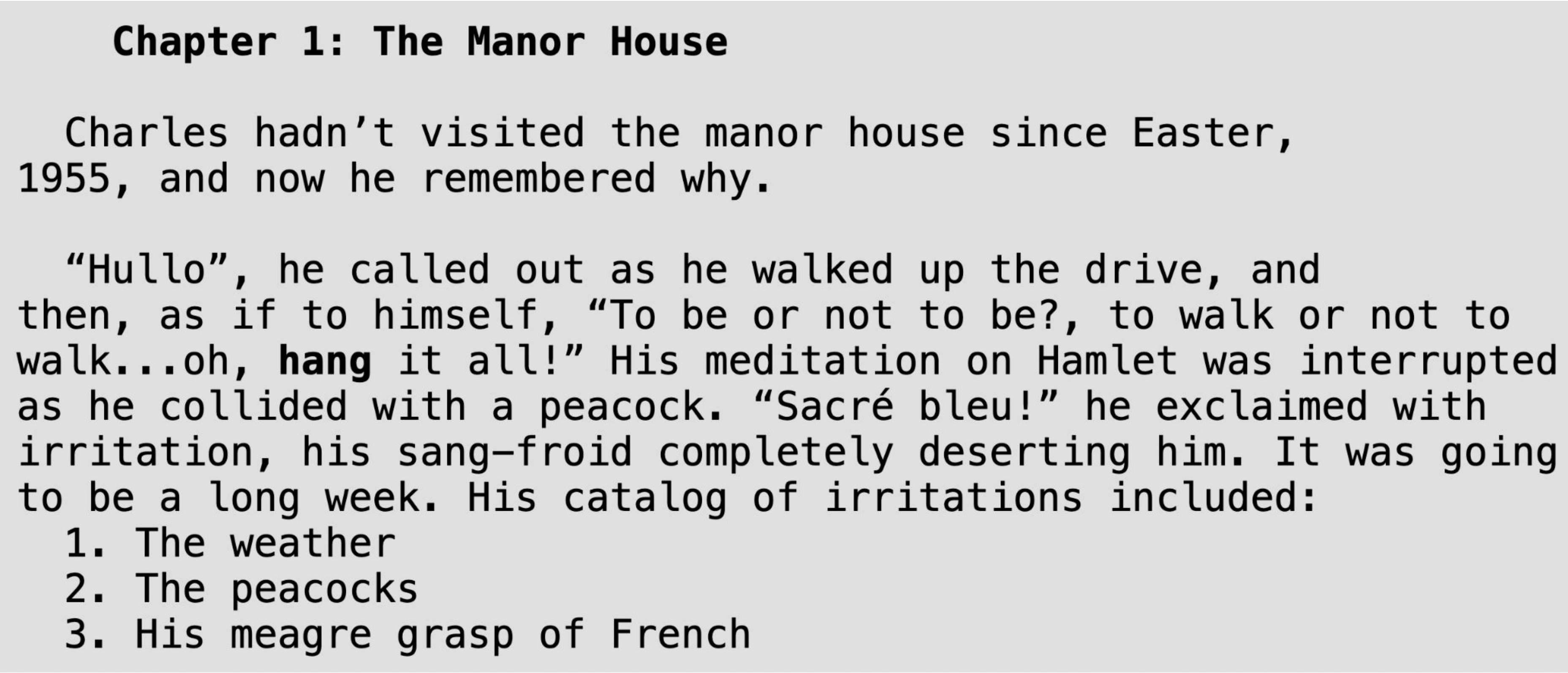
- any number of attributes can be specified on a given start- (or empty-) tag
- but **only one** with a given name!
- order does not matter
- whitespace can be adjusted to make it look good to humans

Anatomy of an element



Introduction to TEI

Sample text fragment



Chapter 1: The Manor House

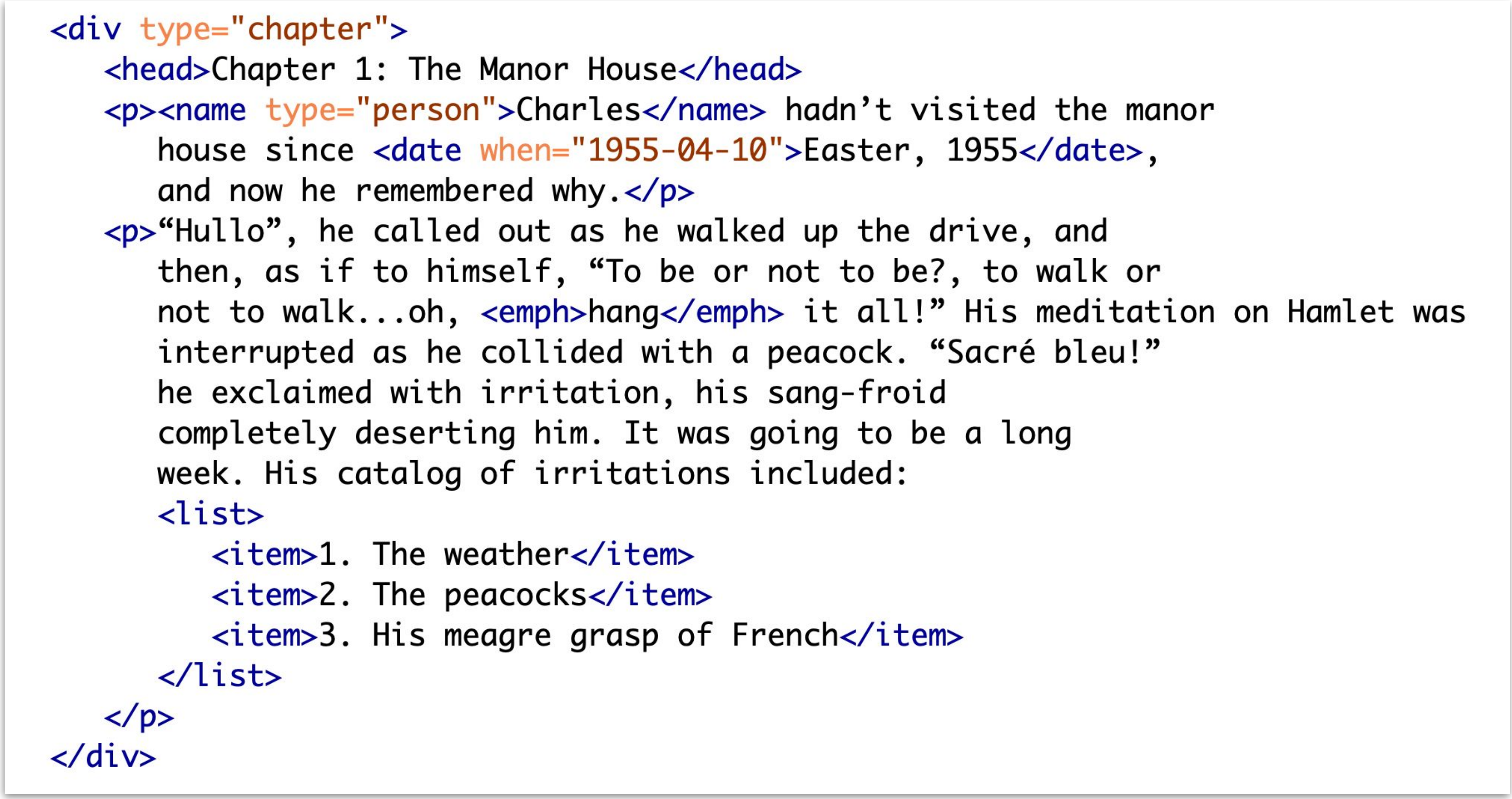
Charles hadn’t visited the manor house since Easter, 1955, and now he remembered why.

“Hullo”, he called out as he walked up the drive, and then, as if to himself, “To be or not to be?, to walk or not to walk...oh, **hang** it all!” His meditation on Hamlet was interrupted as he collided with a peacock. “Sacré bleu!” he exclaimed with irritation, his sang–froid completely deserting him. It was going to be a long week. His catalog of irritations included:

1. The weather
2. The peacocks
3. His meagre grasp of French

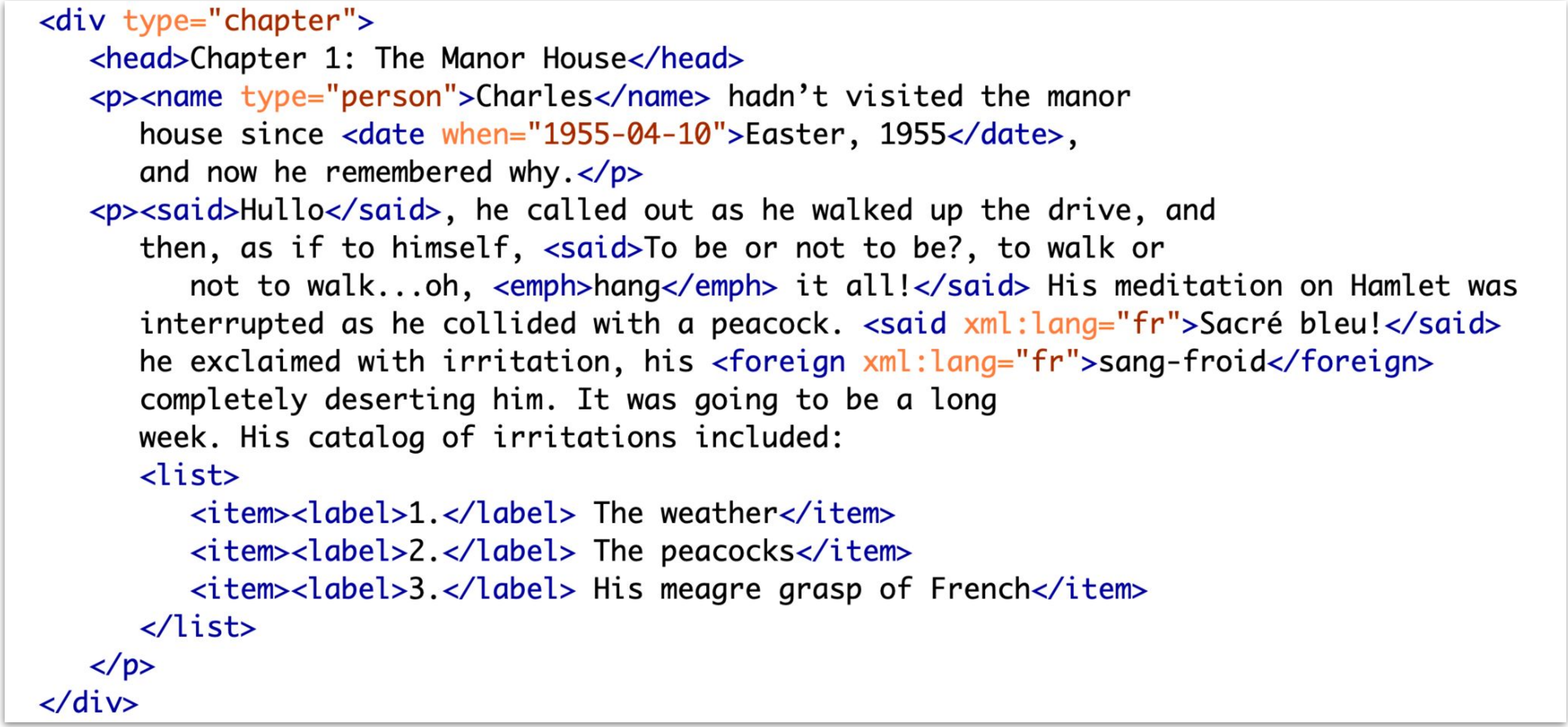
TEI introduction slides from this presentation by the WWP

Basic prose tagging



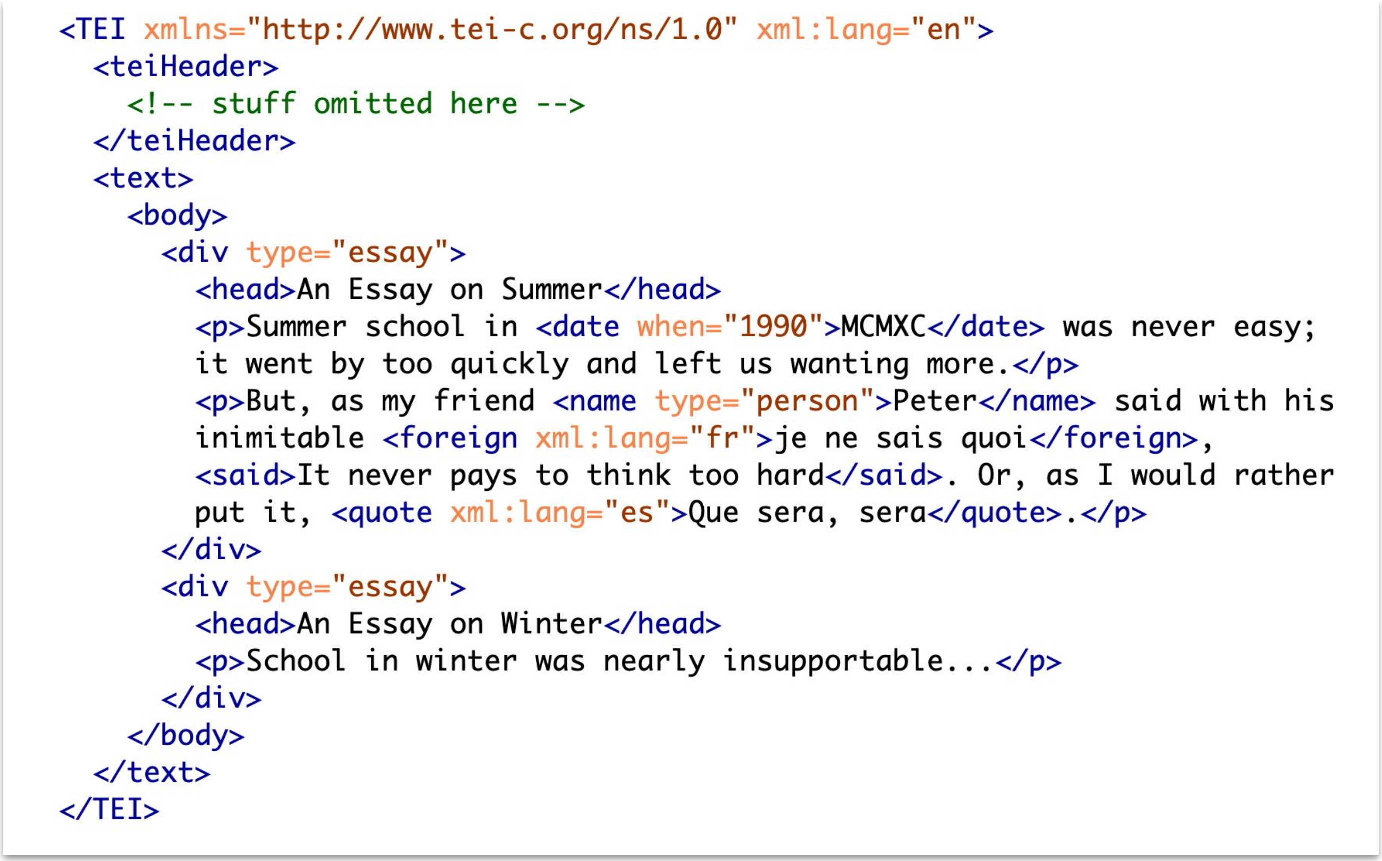
```
<div type="chapter">
   <head>Chapter 1: The Manor House</head>
   <p><name type="person">Charles</name> hadn’t visited the manor
      house since <date when="1955-04-10">Easter, 1955</date>,
      and now he remembered why.</p>
   <p>“Hullo”, he called out as he walked up the drive, and
      then, as if to himself, “To be or not to be?, to walk or
      not to walk...oh, <emph>hang</emph> it all!” His meditation on Hamlet was
      interrupted as he collided with a peacock. “Sacré bleu!”
      he exclaimed with irritation, his sang-froid
      completely deserting him. It was going to be a long
      week. His catalog of irritations included:
      <list>
         <item>1. The weather</item>
         <item>2. The peacocks</item>
         <item>3. His meagre grasp of French</item>
      </list>
   </p>
</div>
```

More detailed encoding



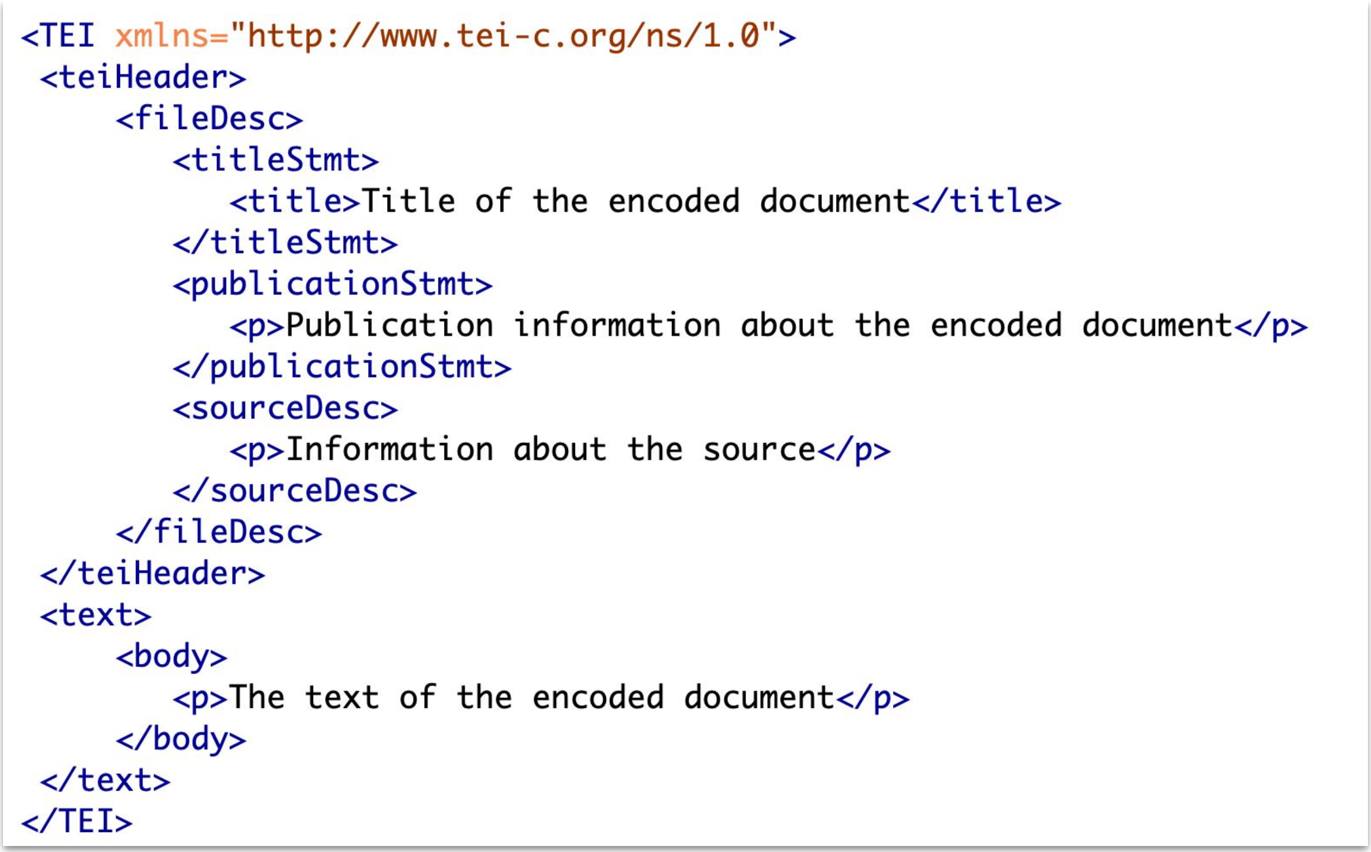
```
<div type="chapter">
   <head>Chapter 1: The Manor House</head>
   <p><name type="person">Charles</name> hadn’t visited the manor
      house since <date when="1955-04-10">Easter, 1955</date>,
      and now he remembered why.</p>
   <p><said>Hullo</said>, he called out as he walked up the drive, and
      then, as if to himself, <said>To be or not to be?, to walk or
         not to walk...oh, <emph>hang</emph> it all!</said> His meditation on Hamlet was
      interrupted as he collided with a peacock. <said xml:lang="fr">Sacré bleu!</said>
      he exclaimed with irritation, his <foreign xml:lang="fr">sang-froid</foreign>
      completely deserting him. It was going to be a long
      week. His catalog of irritations included:
      <list>
         <item><label>1.</label> The weather</item>
         <item><label>2.</label> The peacocks</item>
         <item><label>3.</label> His meagre grasp of French</item>
      </list>
   </p>
</div>
```

The bigger picture



```xml
<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>
```

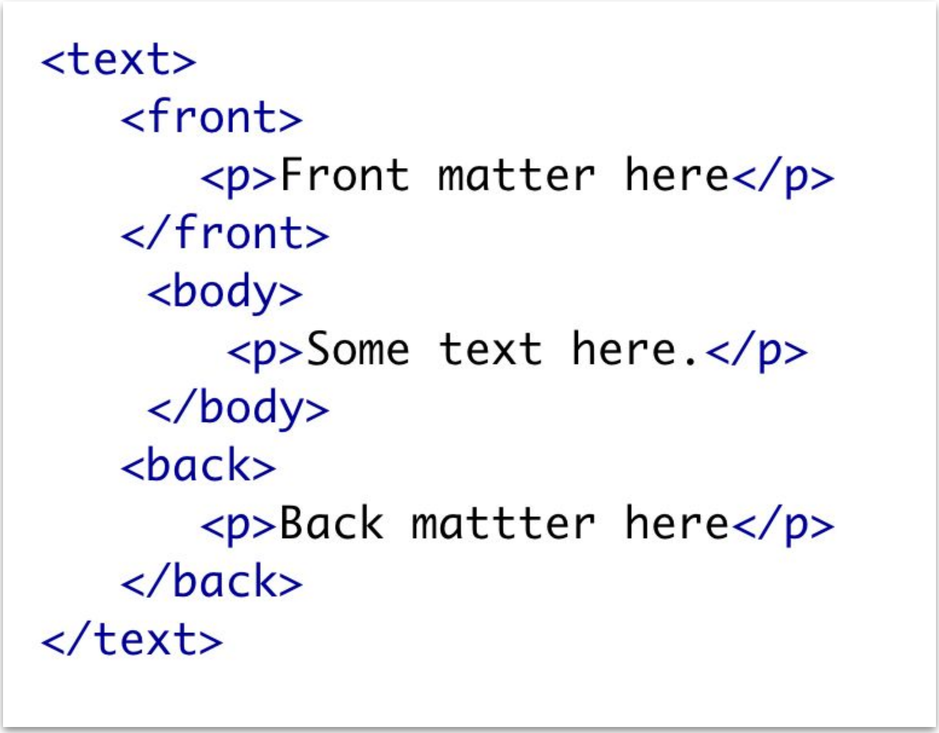
<text> and its contents

<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.)



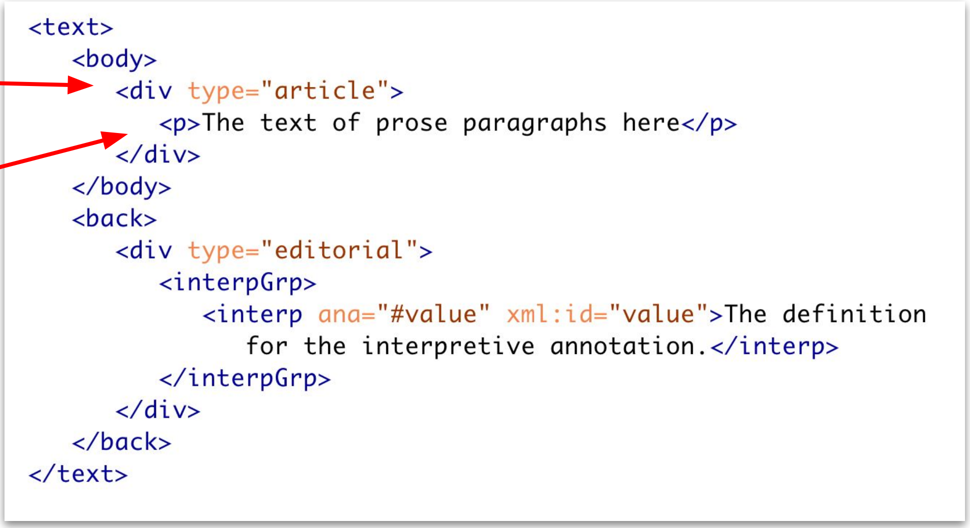
```
<text>
   <front>
      <p>Front matter here</p>
   </front>
    <body>
       <p>Some text here.</p>
    </body>
   <back>
      <p>Back matter here</p>
   </back>
</text>
```

<div> and <p>

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

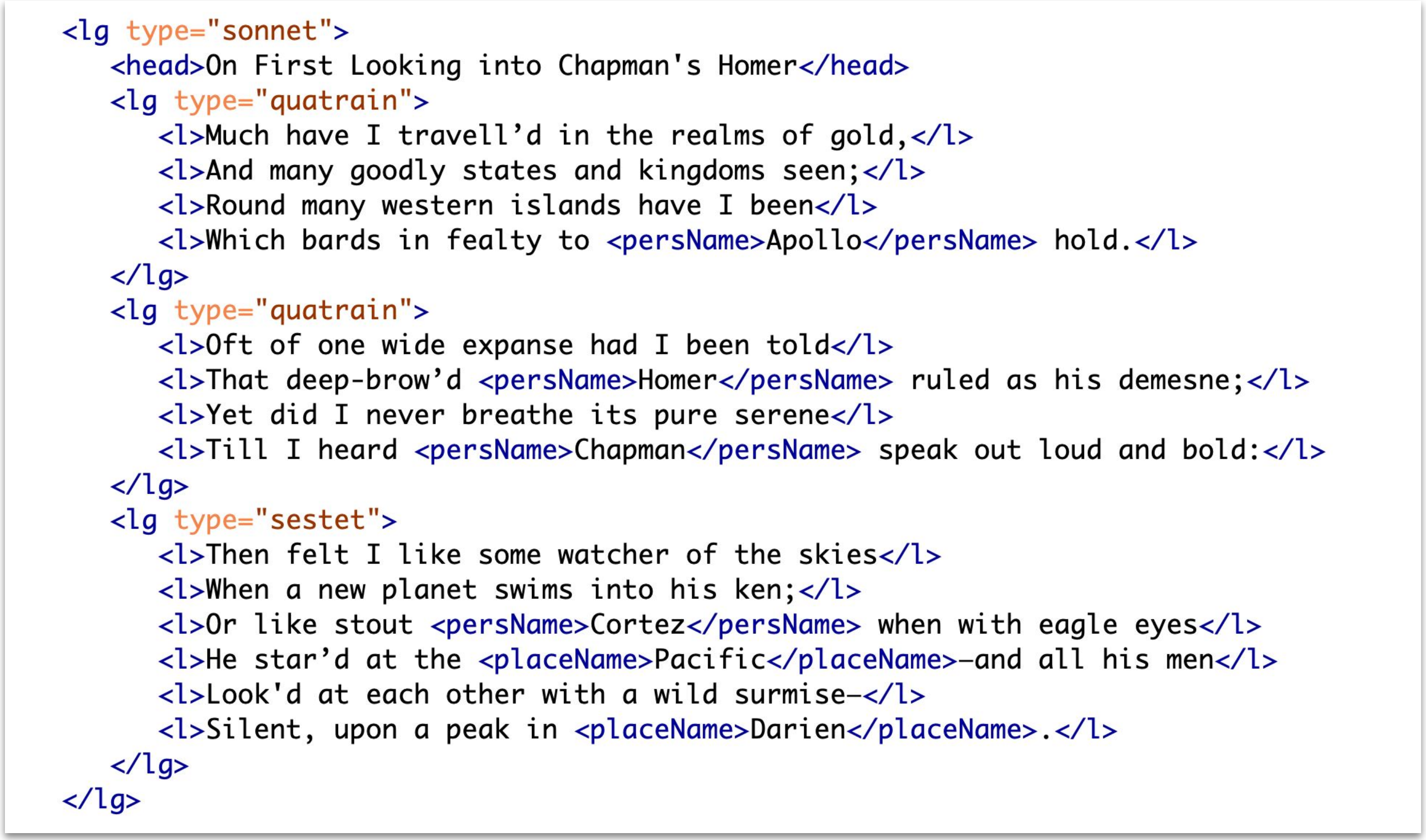
<p> marks paragraphs of prose

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



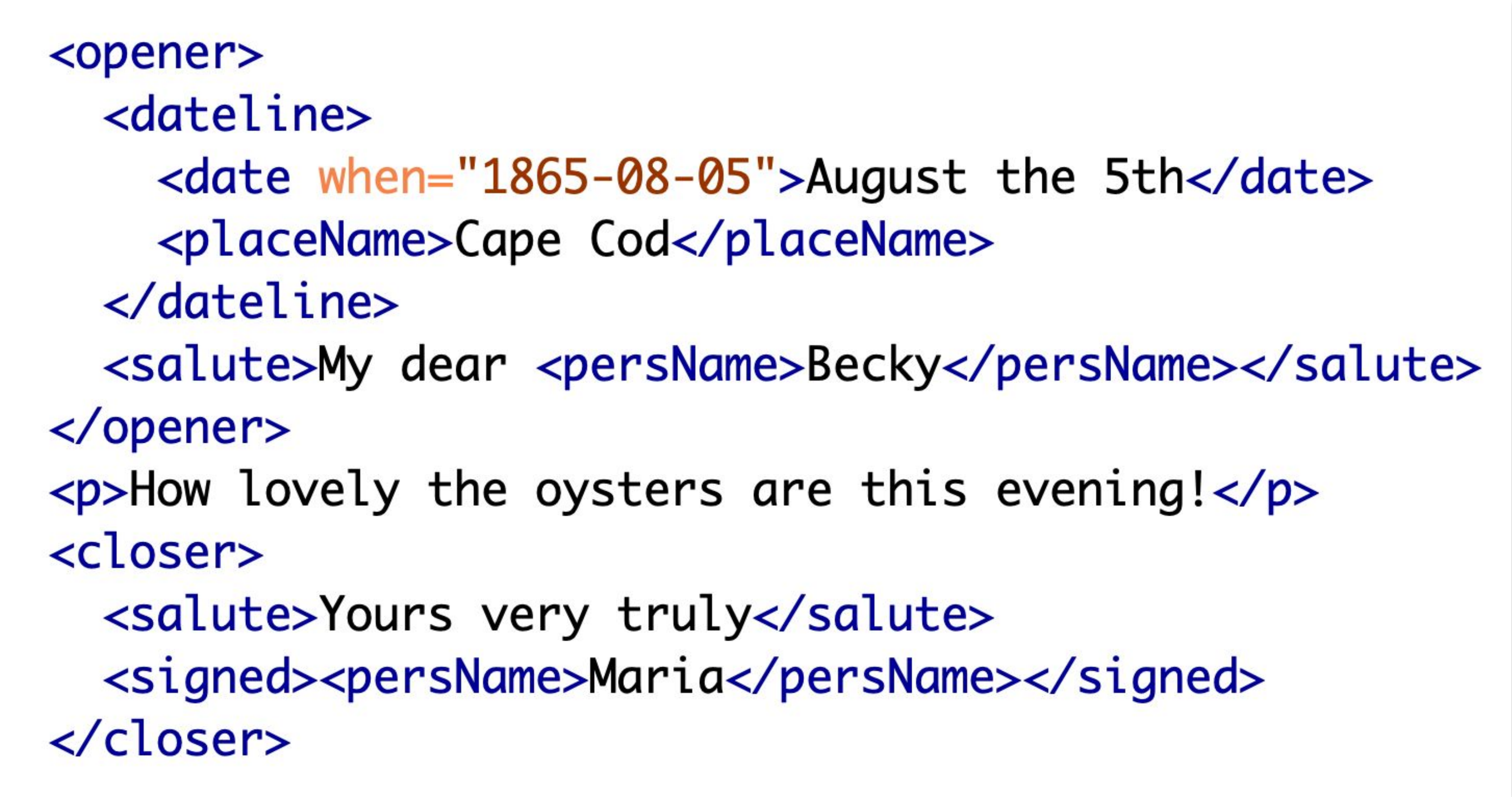
```
<text>
   <body>
      <div type="article">
         <p>The text of prose paragraphs here</p>
      </div>
   </body>
   <back>
      <div type="editorial">
         <interpGrp>
            <interp ana="#value" xml:id="value">The definition
               for the interpretive annotation.</interp>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

Encoding poetry



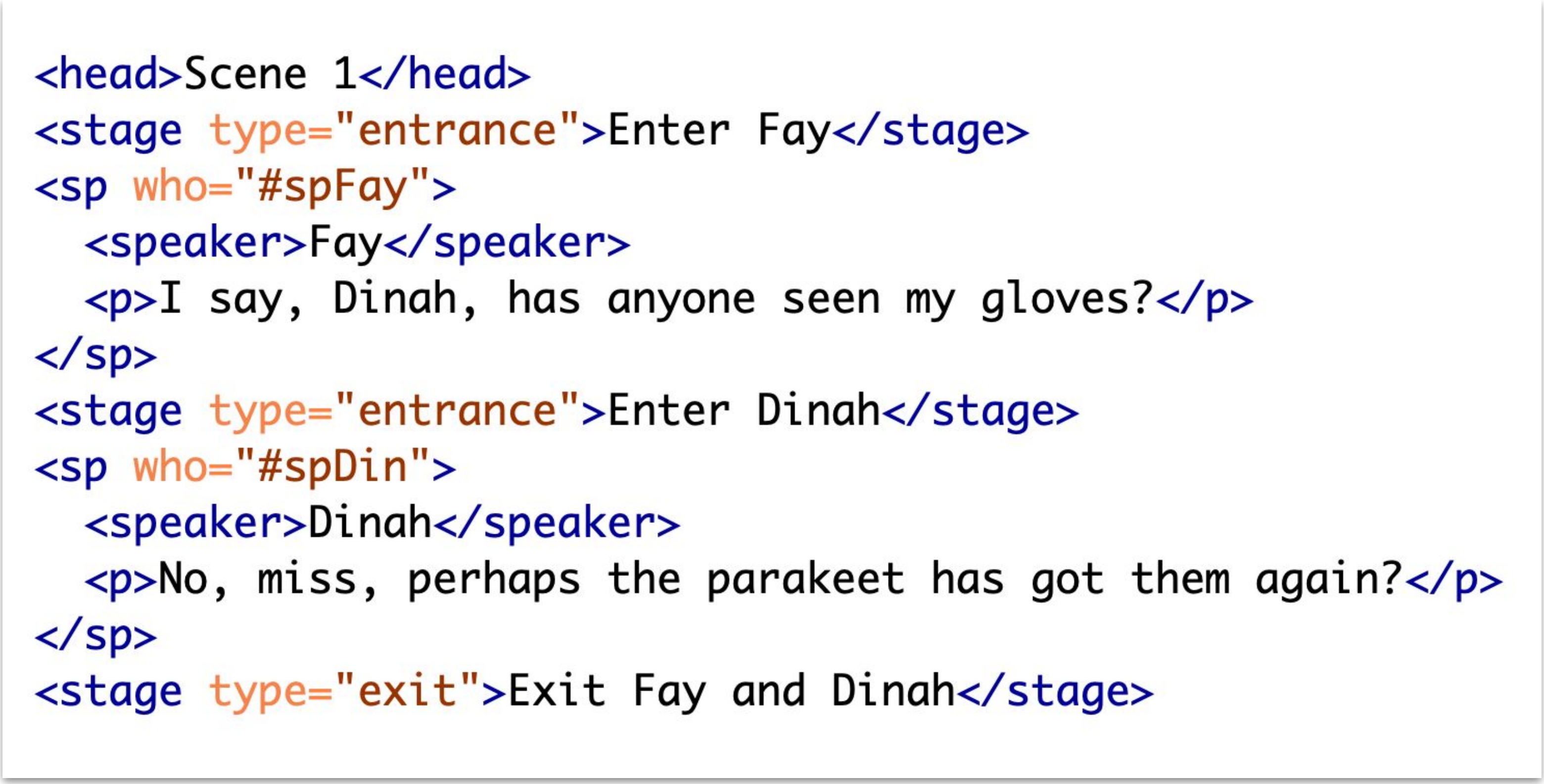
```
<lg type="sonnet">
   <head>On First Looking into Chapman's Homer</head>
   <lg type="quatrain">
      <l>Much have I travell’d in the realms of gold,</l>
      <l>And many goodly states and kingdoms seen;</l>
      <l>Round many western islands have I been</l>
      <l>Which bards in fealty to <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

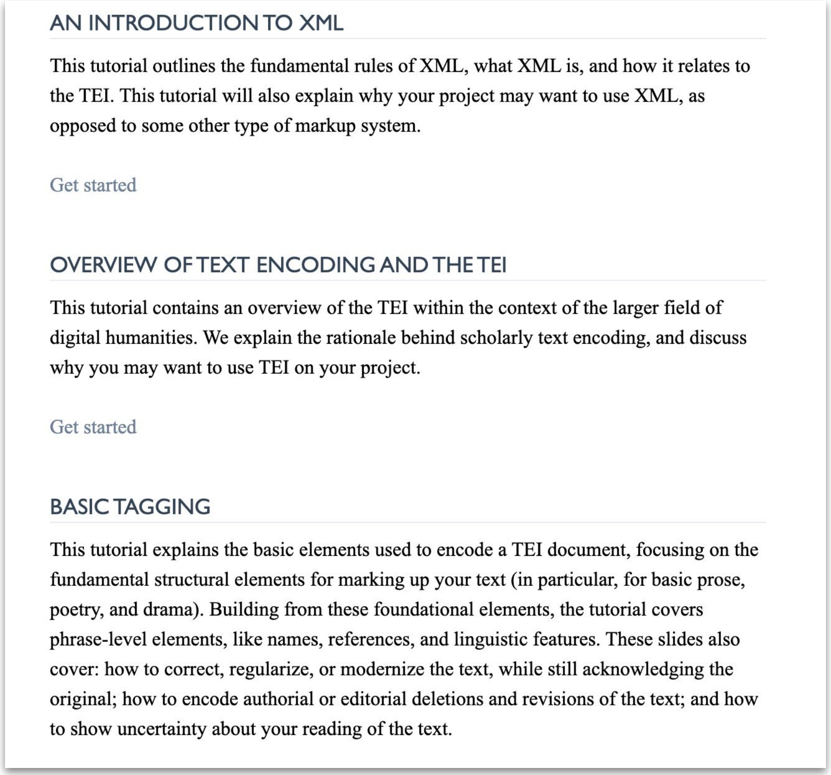


```
<head>Scene 1</head>
<stage type="entrance">Enter Fay</stage>
<sp who="#spFay">
  <speaker>Fay</speaker>
  <p>I say, Dinah, has anyone seen my gloves?</p>
</sp>
<stage type="entrance">Enter Dinah</stage>
<sp who="#spDin">
  <speaker>Dinah</speaker>
  <p>No, miss, perhaps the parakeet has got them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

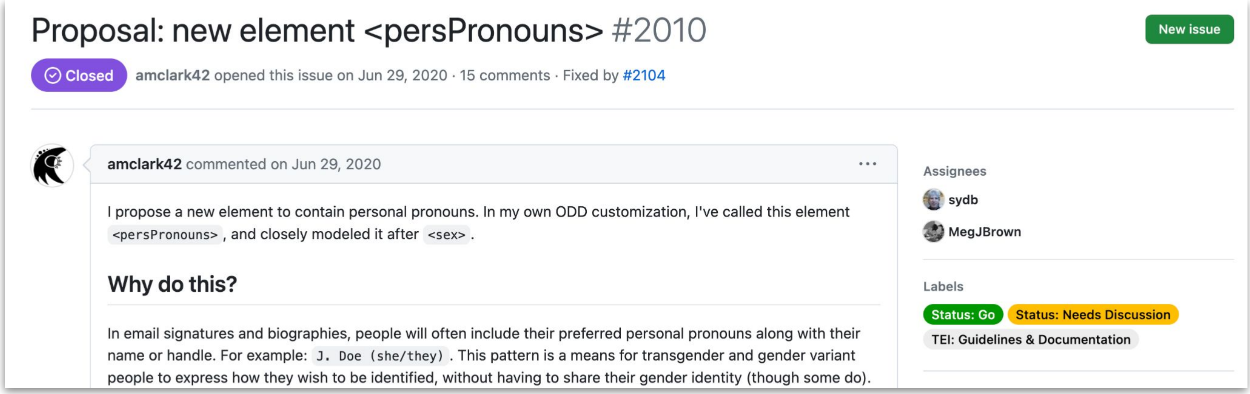
If you want the **text to wrap**: hit **`control/command-shift-Y.`**

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.

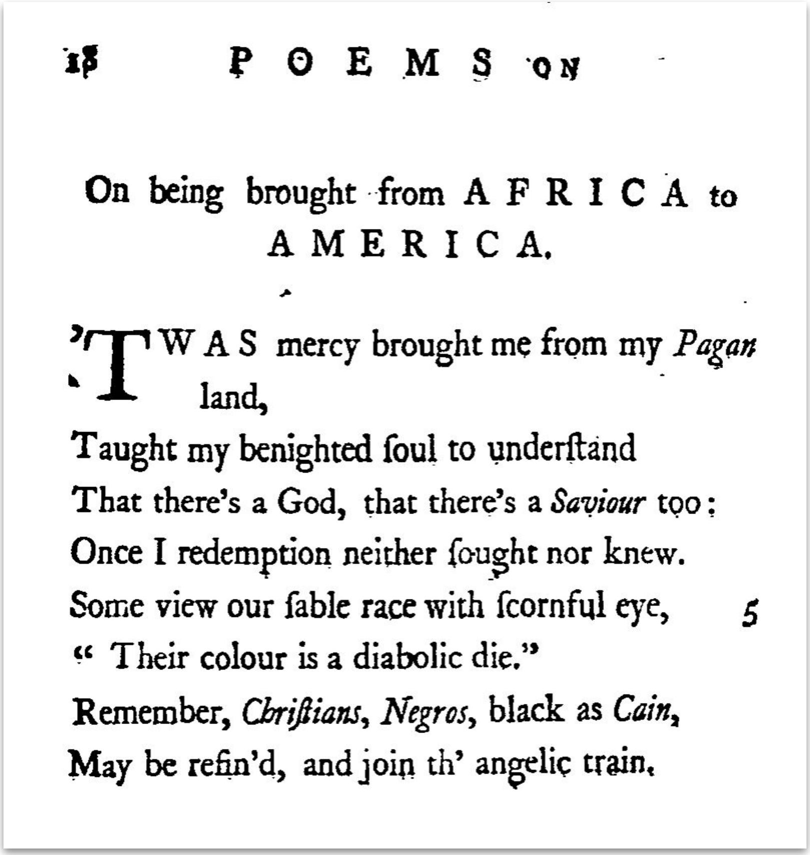
Thought experiment: tagging

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 on the handout. 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



18 POEMS ON

On being brought from AFRICA to AMERICA.

'TWAS mercy brought me from my *Pagan* land,
Taught my benighted ſoul to underſtand
That there's a God, that there's a *Saviour* too:
Once I redemption neither ſought nor knew.
Some view our ſable race with ſcornful eye, 5
"Their colour is a diabolic die."
Remember, *Chriſtians*, *Negros*, black as *Cain*,
May be refin'd, and join th' angelic train.

—Phillis Wheatley Peters, *Poems on Various Subjects, Religious and Moral*, 1773
Image credit: Eighteenth-century Collections Online

Discussion, questions

- 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, and Avery Blankenship

- 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
- Slides and supporting materials are available at: https://bit.ly/fa24-aljoe-tei