

INTRODUCTION TO TEI

Dot Porter, University of Pennsylvania

PhillyDH@Penn

June 2013

What we'll cover today

- Concept of TEI
- What can you do with TEI?
- The TEI Guidelines
- Customizing TEI
- Learning more

What is “TEI”



- Text Encoding Initiative
- De facto “standard” for humanities text encoding
- Described in the TEI Guidelines
 - Encoding scheme
 - Formal documentation
- Expressed using XML
- *Modular and customizable*

Also: TEI *Consortium*

Text Encoding Initiative

- Poughkeepsie Conference, Vassar College, 1987
- Major early influences
 - Digital libraries and text collections
 - Language corpora
 - Scholarly datasets
- Institutional members and individual subscribers
- Board and Council
- SIGS and Workgroups
- TEI P5 published in 2007
- Annual Conference (October 2-5, Rome)



Is TEI a Standard?

- Not a “standard”, but standardized guidelines
- Standards-based functionality
 - Datatyping
 - Controlled vocabularies
- Control within projects and disciplines (e.g., EpiDoc)
- Separation of content and display

“TEI Guidelines”

- The encoding scheme
 - 500+ tags (human-readable)
 - 21 modules (4 required)
 - Model classes and attribute classes

- Documentation
 - Describes the encoding scheme

TEI is expressed in XML

- Extensible Markup Language

Advantages of XML

- Simple and flexible
- Enables (communities of) users to create their own tag sets
- Widely used – support and tools
- ISO standard
- Software independent
- Facilitates the movement of data
- Separates content from display

TEI is *Modular* and *Customizable*

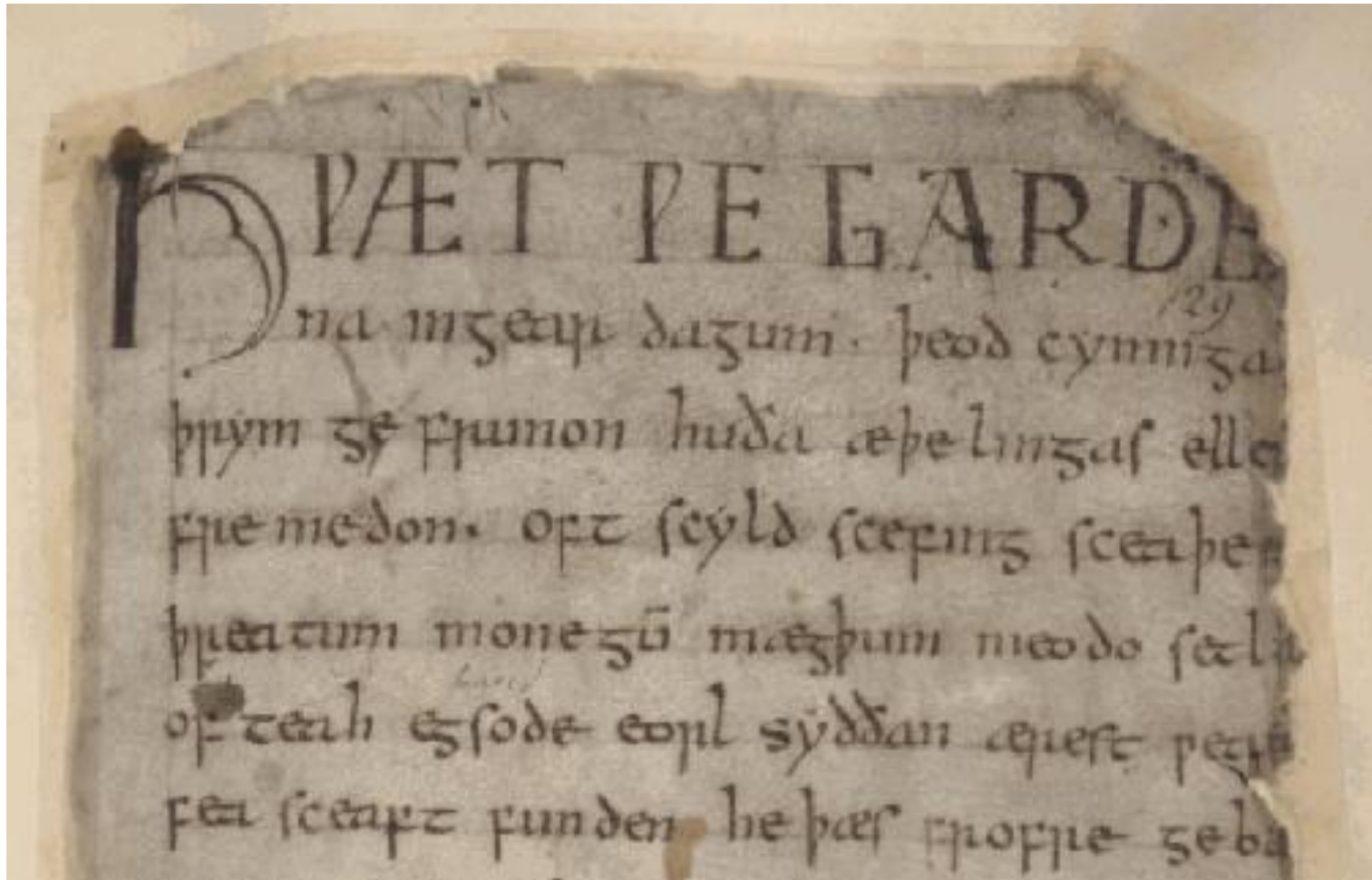
- 500+ tags divided into 21 modules
- Only four modules required:
 - tei
 - header
 - core
 - textstructure
- Customize: Use only the modules and tags you need (e.g. TEI Lite, EpiDoc)

Customize TEI using *ROMA* and *ODD*

Next: Text analysis and markup

What is a text?

Damaged letters
Special characters
Folio lines
Abbreviations
(etc.)



What is a text?

Poetic lines and half-lines
Marginal notes
Etc.

Hwæt wē Gār-Dena in geār-dagum
þēod-cyninga þrym gefrūnon,
hū ðā æþelingas ellen fremedon.

Oft Scyld Scēfing sceapena brēatum,
5 monegum mægþum meodo-setla oftēah;
egsode Eorl[e], syððan ārest wearð
fēasceaft funden; hē þæs frōfre gebād:
wēox under wolcnum, weorð-myndum þāh,
oðþæt him æghwylc þāra ymb-sittendra
10 ofer hron-rāde hýran scolde,

All of these and more

- Only that which is explicit can be reliably processed by a computer

February 8, 1976



`<date>February 8, 1976</date>`



`<date when="1976-02-08">February 8, 1976</date>`



Markup or Encoding

Markup!

- Name and characterize parts of a text in a formalized way

Markup!

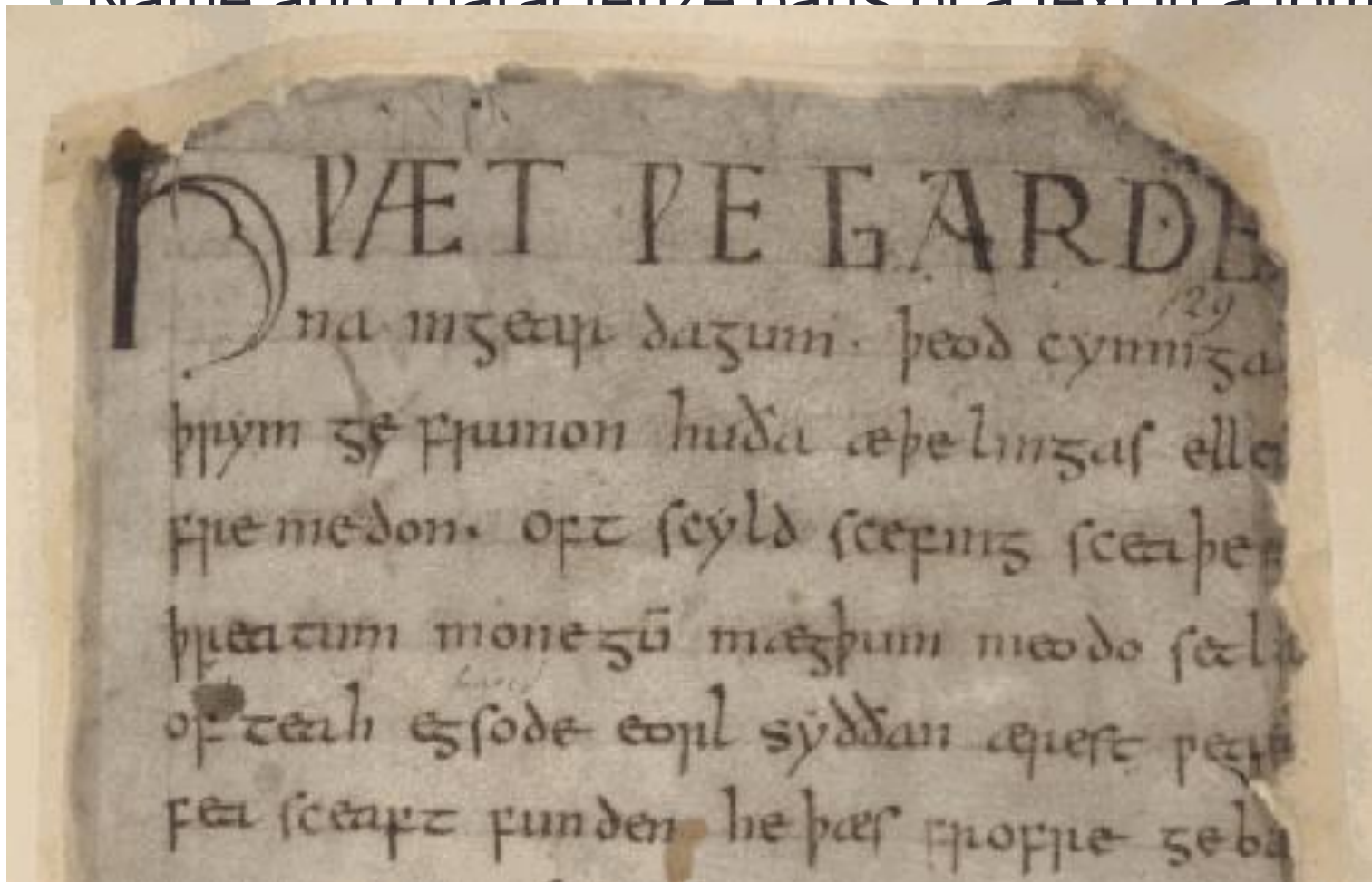
Damaged letters: <damage><unclear></unclear><damage>
or <damage><supplied></supplied></damage>

Special characters: <g></g>

Folio lines: <lb/> or <line></line>

Abbreviations: <choice><abbr></abbr><expan></expan></choice>
(etc.)

- Name and characterize parts of a text in a formalized way



Markup!

Poetic lines and half-lines: <lg>

<| n="1">...<caesura/>...</|>

<| n="2">...<caesura/>...</|>

</lg>

Marginal notes: <add place="margin"></add>

Etc.

- Name and characterize parts of a text in a formalized way



Hwæt wē Gār-Dena in geār-dagum

þēod-cyninga þrym gefrūnon,

hū ðā æþelingas ellen fremedon.

Oft Scyld Scēfing sceapena þrēatum,

5 monegum mægþum meodo-setla oftēah;

egsode Eorl[e], syððan ærest wearð

fēasceaft funden; hē þæs frōfre gebād:

wēox under wolcnum, weorð-myndum þāh,

oðþæt him æghwylc þāra ymb-sittendra

10 ofer hron-rāde hýran scolde,

Markup!



normalized way

Hwæt wē Gār-Dena in geār-dagum
þēod-cyninga þrym gefrūnon,
hū ðā æþelingas ellen fremedon.

Oft Scyld Scēfing sceapena þrēatum,
5 monegum mægþum meodo-setla oftēah;
egsode Eorl[e], syððan ærest wearð
fēasceaft funden; hē þæs frōfre gebād:
wēox under wolcnum, weorð-myndum þāh,
oðþæt him æghwylc þāra ymb-sittendra
10 ofer hron-rāde hýran scolde,

Name what things *are* rather than what they *look like*

What does TEI make explicit?

- ▢ structural divisions within a text

 - ★ title-page, chapter, scene, stanza, line, paragraph, etc.

- ▢ typographical elements

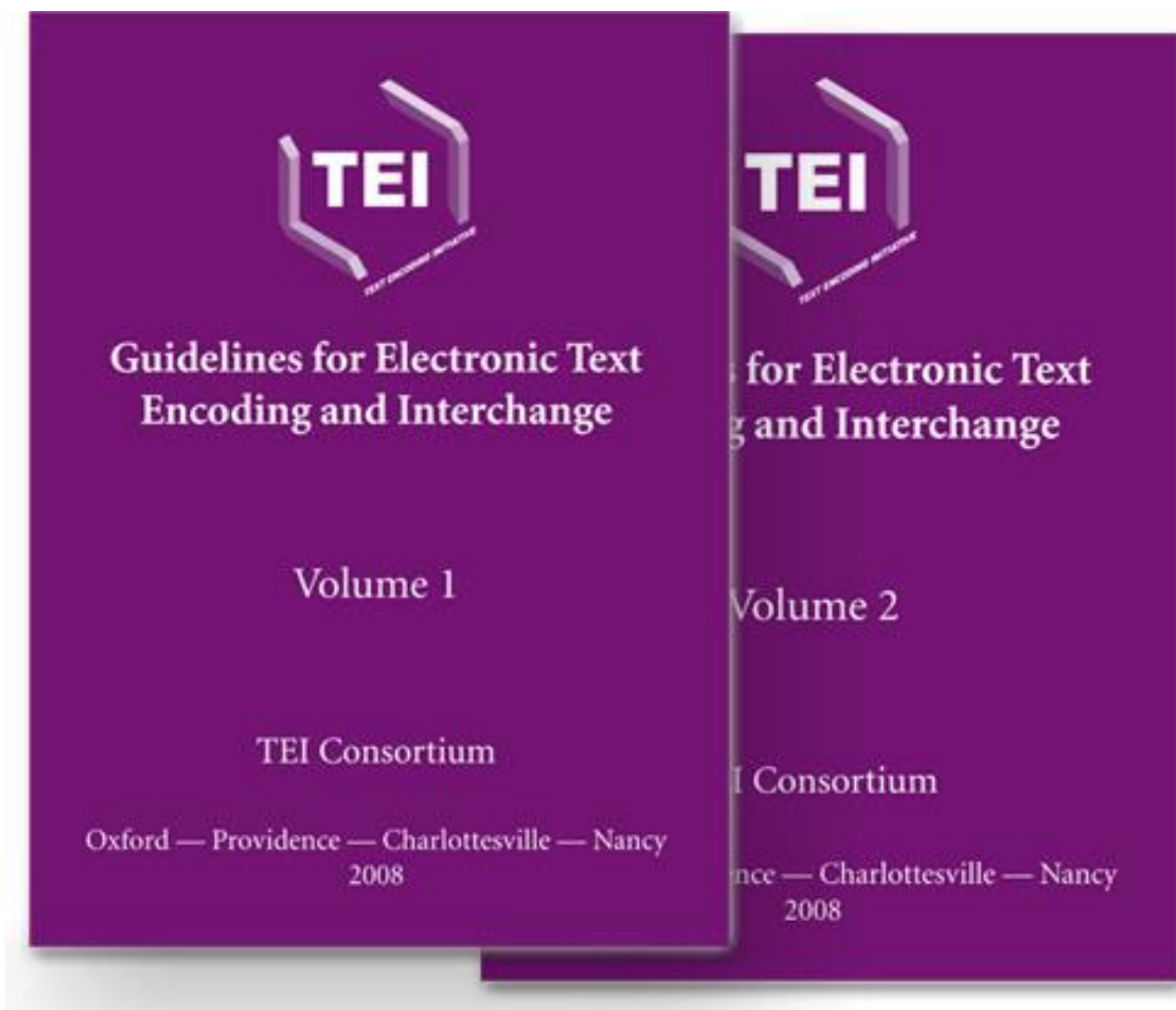
 - ★ changes in typeface, special characters, etc.

- ▢ other features

 - ★ people, places, events

 - ★ grammatical structures, location of illustrations, variant forms, etc.

How does TEI make it explicit?



How does TEI make it explicit?

- The encoding scheme
 - 500+ tags (human-readable)
 - 21 modules (4 required)
 - Model classes and attribute classes

- Documentation
 - Describes the encoding scheme

Original appearance/display: *Moby Dick*

HTML: `<i>Moby Dick</i>`

TEI XML: `<title type="main" level="m">Moby Dick</title>`

What else can markup do?

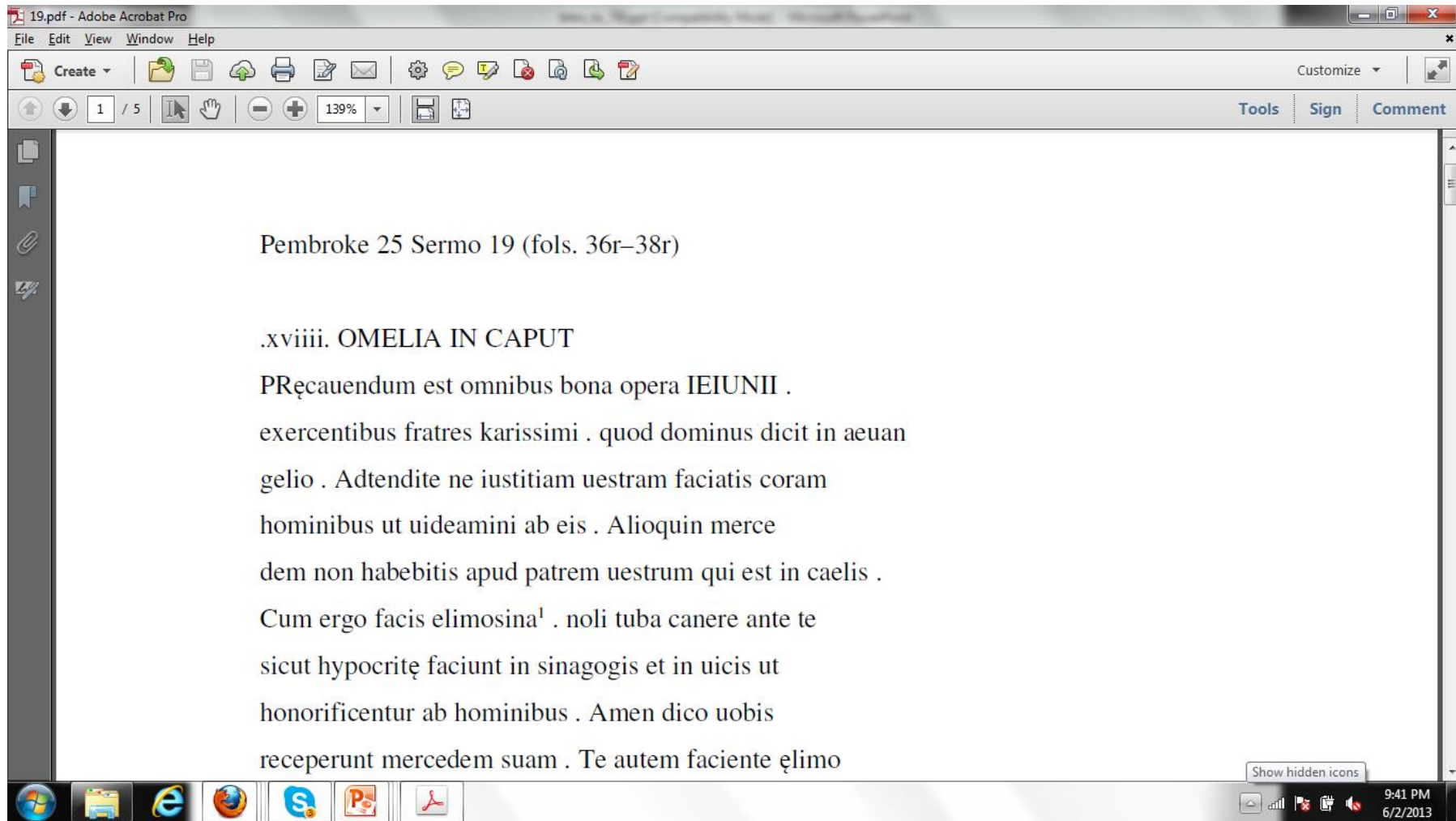
- Add value by supplying multiple annotations to the same text
- Facilitate re-use of the same material
 - in different formats
 - in different contexts
 - by different users

Hwæt wē Gār-Dena in geār-dagum
þēod-cýninga þrym gefrūnon,
hū ðā æþelingas ellen fremedon.
Oft Scyld Scēfing sceapenā prēatum,
5 monegum mægþum meodo-setla oftēah;
egsode Eorl[e], - syððan ærest wearð
feascraft funden; hē þæs frōfre gebād:
wēox under wolcnum, weorð-myndum þāh,
oðþæt him æghwylc þāra ymb-sittendra
10 ofer hron-rāde hyran scolde, (of him)

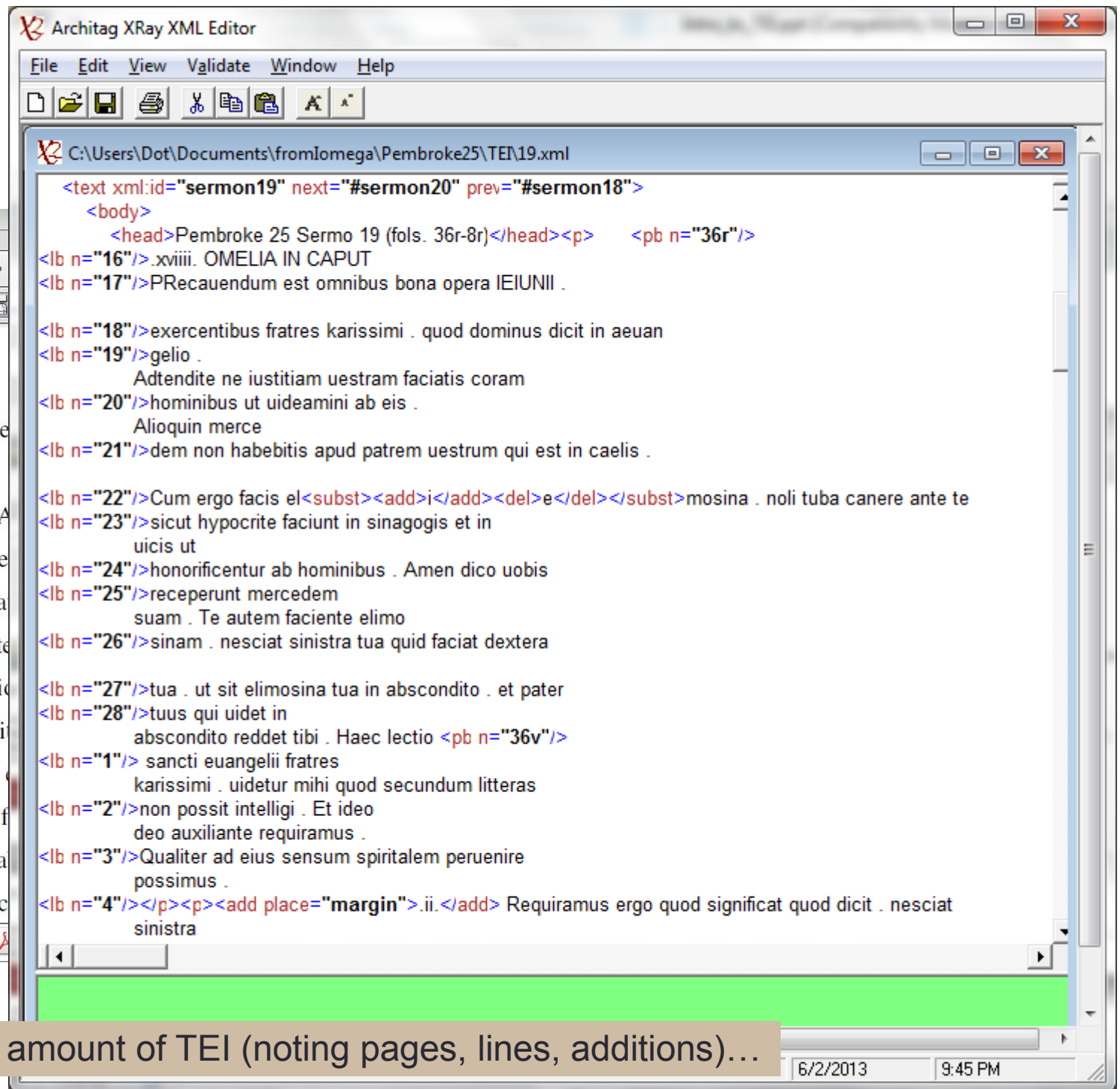
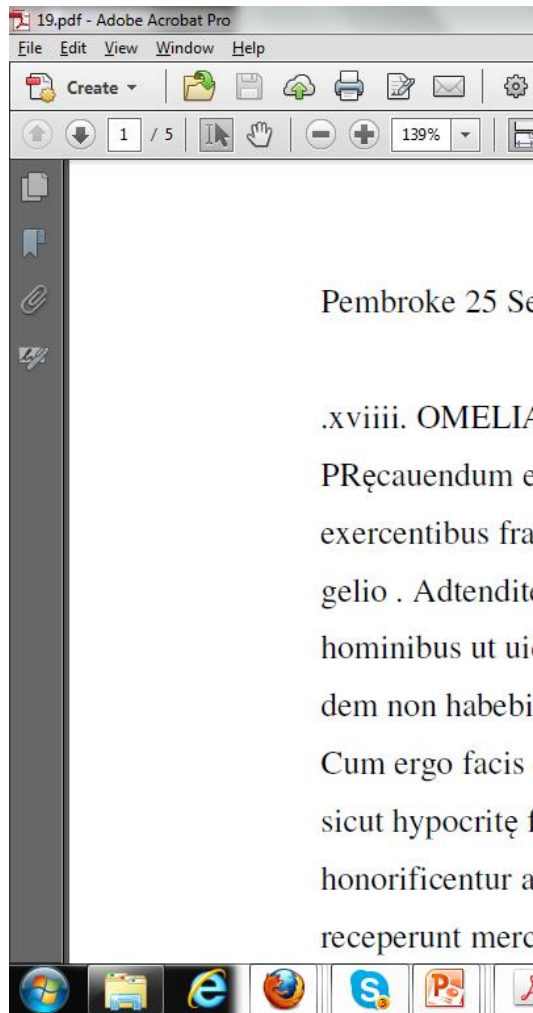


Why?

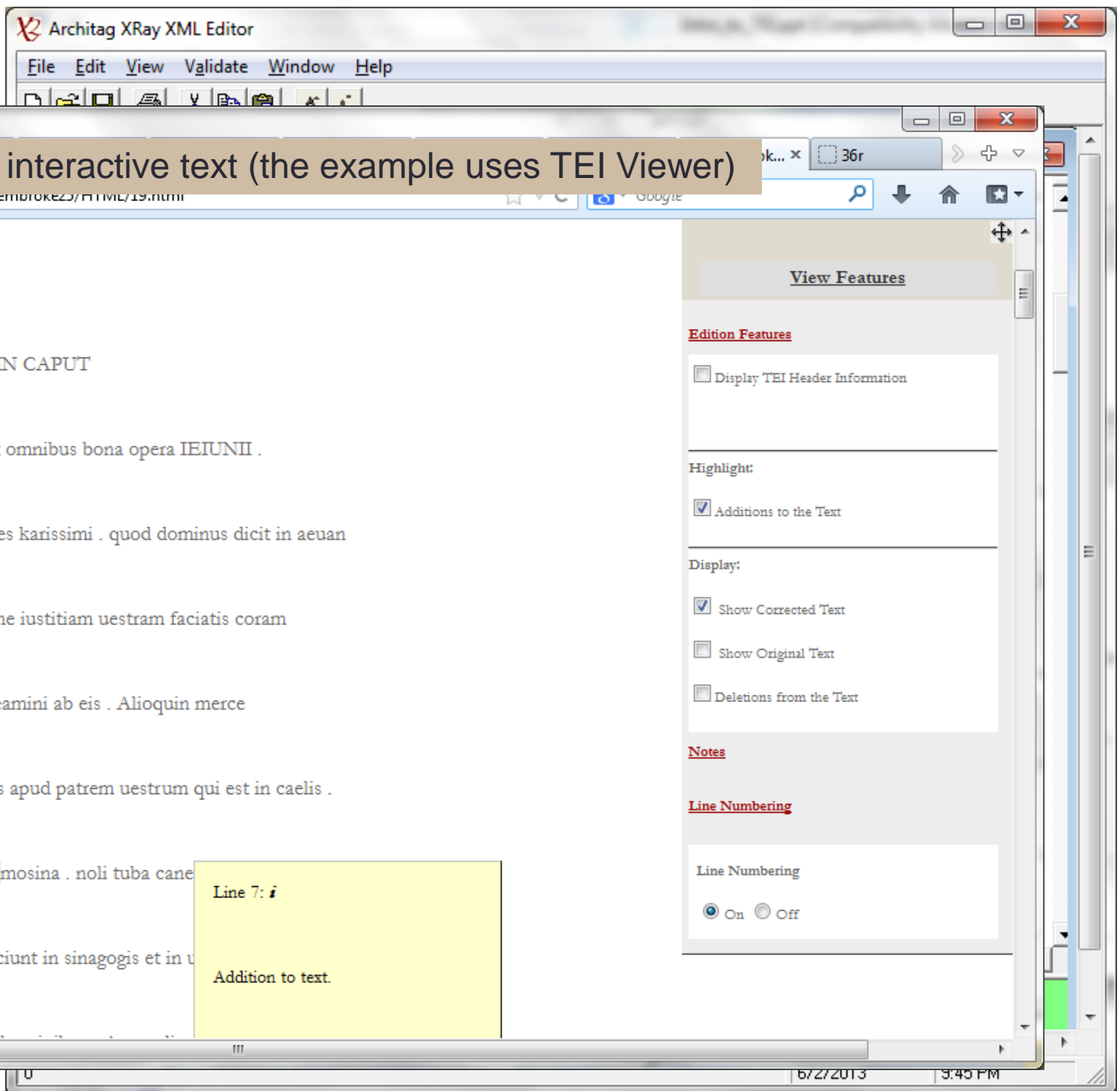
Why use TEI when you can just make a PDF?



Why?



With just a small amount of TEI (noting pages, lines, additions)...



You can create an interactive text (the example uses TEI Viewer)

[36r]

16 .xviii. OMELIA IN CAPUT

17 PRecauendum est omnibus bona opera IEIUNII .

18 exercentibus fratres karissimi . quod dominus dicit in aeuan

19 gelio . Adtendite ne iustitiam uestram faciatis coram

20 hominibus ut uideamini ab eis . Alioquin merce

21 dem non habebitis apud patrem uestrum qui est in caelis .

22 Cum ergo facis elimosina . noli tuba cane

23 sicut hypocrite faciunt in sinagogis et in u

Line 7: i

Addition to text.

View Features

Edition Features

☐ Display TEI Header Information

Highlight:

☒ Additions to the Text

Display:

☒ Show Corrected Text

☐ Show Original Text

☐ Deletions from the Text

Notes

Line Numbering

Line Numbering

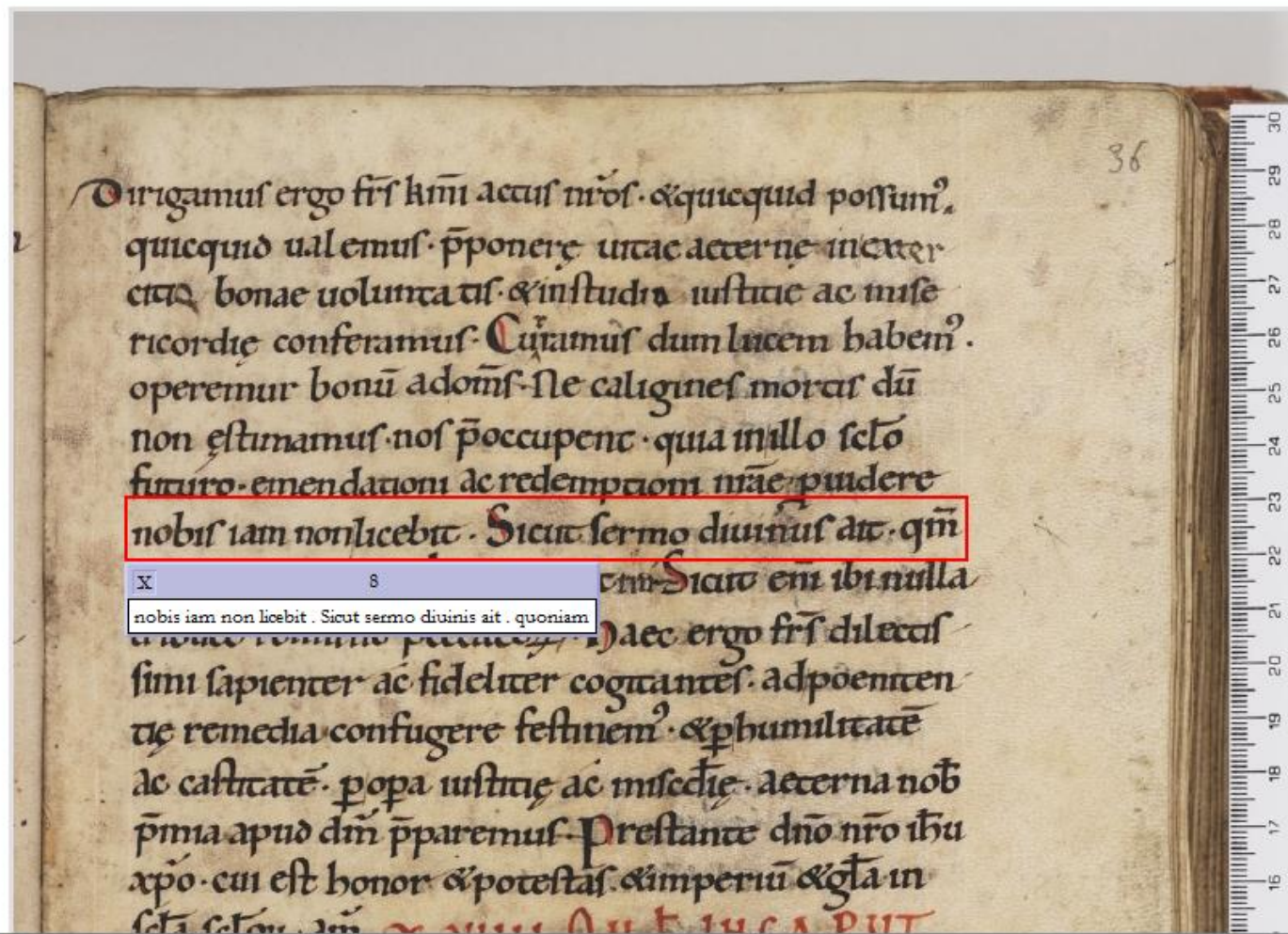
☒ On ☐ Off

You can build on it to link text to image (the examples uses UVIC Image Markup Tool)

file:///C:/Users/Dot/Documents/fromIomega/Pembroke25/IMT/36r/36r.htm

Google

36r



Annotations

Line-by-line transcriptions

- 1
- 2
- 3
- 4
- 5
- 6
- 7
- 8
- 9
- 10
- 11
- 12
- 13
- 14
- 15
- 16
- 17
- 18
- 19
- 20
- 21
- 22
- 23
- 24
- 25
- 26
- 27
- 28

You can use it to generate Linked Data (the example is the Canonical Text Service)

CTS • Homer Multitext

The Homer Multitext project: Canonical Text service

CTS: [home](#) | [credits](#)

Homer, *Iliad*: 1.10

A

(= urn:cts:greekLit:tlg0012.tlg001.msA:1.10)

1

1 Μῆνιν ᾄειδε θεὰ Πηληϊάδεω Ἀχιλῆος

2 οὐλομένην· ἣ μυρ' Ἀχαιοῖς ἄλγε' ἔθηκεν·

3 πολλὰς δ' ἰφθίμους ψυχὰς Ἄϊδι προΐαψεν

4 ἡρώων· αὐτοὺς δὲ ἐλώρια τεῦχε κύνεσσιν

5 οἰωνοῖσ' τε πᾶσι· Λιὸς δ' ἐτελείετο βουλή·

<http://www.tei-c.org/Activities/Projects/>

Introduction to XML

Tags

<tag> content </tag>

<author> Charles Darwin </author>

<date> 15 April, 1860 </date>

<underlined> especially </underlined>

Tags

•

<pb/>
<opener>
 <placeName>Sudbrook Park</placeName>
 <placeName>Richmond</placeName>
 <date>Saturday <supplied>30 June 1860</supplied></date>
 <salute>Dear Sir</salute>
</opener>
<p>I write a line to thank you much for the kind manner with which you have received my rather unreasonable request. — If you find any pollen-masses removed, will you watch a group of the Bee-orchis for 1/4 or 1/2 an hour<add>, and see what sort of insect visits them</add>. I have received account of pollen-masses of this plant having been seen on proboscis of a day-moth; but I cannot<pb/> help feeling a little sceptical about the identification. If you send any <add>other</add> orchids perhaps you would kindly enclose a Bee-orchis, (especially if you find one or more with pollen-masses removed) for this summer I have as yet searched in vain for specimen near my home; <choice>
 <reg>and</reg>
 <orig>&</orig>
</choice> I want to have pollen-masses for standard of comparison with those observed on the probosces of moths. —</p>
<p>The Spiranthes would be <hi>especially</hi> valuable to me <choice>
 <reg>and</reg>
 <orig>&</orig>
</choice> the Epipactus.</p>
<p>I shall return home on next Thursday <add>(5<hi>th</hi>)</add> (to Down, Bromley, Kent) <choice>
 <reg>and</reg>
 <orig>&</orig>
</choice> on Monday 9<hi>th</hi> or 10<hi>th</hi> I shall go to

Attributes

<tag attribute="value"> content </tag>

<author ref="#CD1"> Charles Darwin </author>

<date when="1860-04-15"> 15 April, 1860 </date>

<person gender="m" id="CD1"> Darwin, Charles </person>

Attributes

- ```
<pb n="1"/>
<opener>
 <placeName>Sudbrook Park</placeName>
 <placeName>Richmond</placeName>
 <date when="1860-06-30">Saturday <supplied>30 June 1860</supplied></date>
 <salute>Dear Sir</salute>
</opener>
<p>I write a line to thank you much for the kind manner with which you have received my
rather unreasonable request. — If you find any pollen-masses removed, will you
watch a group of the Bee-orchis for 1/4 or 1/2 an hour<add>, and see what sort of insect
visits them</add>. I have received account of pollen-masses of this plant having been seen on
proboscis of a day-moth; but I cannot<pb n="2"/> help feeling a little sceptical about
the identification. If you send any <add>other</add> orchids perhaps you would kindly enclose a
Bee-orchis, (especially if you find one or more with pollen-masses removed) for this
summer I have as yet searched in vain for specimen near my home; <choice>
 <reg>and</reg>
 <orig>&</orig>
</choice> I want to have pollen-masses for standard of comparison with those observed on
the probosces of moths. —</p>
<p>The Spiranthes would be <hi rend="underline">especially</hi> valuable to me <choice>
 <reg>and</reg>
 <orig>&</orig>
</choice> the Epipactus.</p>
<p>I shall return home on next Thursday <add>(5<hi rend="sup">th</hi>)</add> (to
Down, Bromley, Kent) <choice>
 <reg>and</reg>
 <orig>&</orig>
</choice> on Monday 9<hi rend="sup">th</hi> or 10<hi rend="sup">th</hi> I shall go to
```

# Advantages of XML

- Simple and flexible
- Enables (communities of) users to create their own tag sets
- Widely used – support and tools
- ISO standard
- Software independent
- Facilitates the movement of data
- Separates content from display

# Widely used

- Text Encoding Initiative (**TEI**)
- Math Markup Language (**MathML**)
- **MusicXML**
- Encoded Archival Description (**EAD**)
- Keyhole Markup Language (**KML**)
- Metadata Object Description Schema (**MODS**)
- Metadata Encoding and Transmission Standard (**METS**)

*And pretty much everything on the  
World Wide Web*



# XML is defined by *schemas*

- Rules for a specified set of XML tags and attributes

`<name when="1976-02-08">Dot Porter</name>`

Which attributes may appear on which tags. Example: no attribute "when" on tag "name"

# XML is defined by *schemas*

- Rules for a specified set of XML tags and attributes

 <name when="1977-12-08">Dot Porter</name>

Which attributes may appear on which tags. Example: no attribute "when" on tag "name"

# XML is defined by *schemas*

- Rules for a specified set of XML tags and attributes

<name when="1977-12-08">Dot Porter</name>



Once upon <c><w>a</w></c> dream

Which tags may nest within which other tags.  
Example: “c” tag (character) must nest within “w” tag (word)

# XML is defined by *schemas*

- Rules for a specified set of XML tags and attributes

<name when="1977-12-08">Dot Porter</name>

Once upon <c><w>w</w></c> dream

Which tags may nest within which other tags.  
Example: "c" tag (character) must nest within "w"  
tag (word)

# XML is defined by *schemas*

- Rules for a specified set of XML tags and attributes

<name when="1977-12-08">Dot Porter</name>

Once upon <c><w>w</w></c> dream

<sourceDesc><p>Venetus A</p></sourceDesc>  
<sourceDesc><p>Venetus B</p></sourceDesc>

Which tags may repeat, and what order they may appear in. Example: sourceDesc may not appear more than once

# XML is defined by *schemas*

- Rules for a specified set of XML tags and attributes

<name when="1977-12-08">Dot Porter</name>

Once upon <c><w></w></c> dream

<sourceDesc><p>Veritas A</p></sourceDesc>  
<sourceDesc><p>Veritas B</p></sourceDesc>

Which tags may repeat, and what order they may appear in. Example: sourceDesc may not appear more than once

# XML is defined by *schemas*

- Rules for a specified set of XML tags and attributes

<name when="1977-12-08">Dot Porter</name>

Once upon <c><w>w</w></c> dream

<sourceDesc><p>Version A</p></sourceDesc>  
<sourceDesc><p>Version B</p></sourceDesc>

What content may appear in which tags. Example: limiting the content of <sex> to a controlled list

<sex>YES</sex>

# XML is defined by *schemas*

- Rules for a specified set of XML tags and attributes

<name when="1977-12-08">Dot Porter</name>

Once upon <c><w> w</w></c> dream

<sourceDesc><p>Veritas A</p></sourceDesc>  
<sourceDesc><p>Veritas B</p></sourceDesc>

<sex>S</sex>



# XML is defined by *schemas*

- Rules for a specified set of XML tags and attributes

<name when="1977-12-08">Dot Porter</name>

Once upon <c><w> w</w></c> dream

<sourceDesc><p>Veritas A</p></sourceDesc>  
<sourceDesc><p>Veritas B</p></sourceDesc>

<sex>S</sex>

<abbr type="contraction">ds</abbr>

Which values are allowed on which attributes. Example: limiting the values of types of abbreviations to a controlled list.

# XML is defined by *schemas*

- Rules for a specified set of XML tags and attributes

<name when="1977-12-08">Dot Porter</name>

Once upon <c><w> w</w></c> dream

<sourceDesc><p>Veritas A</p></sourceDesc>  
<sourceDesc><p>Veritas B</p></sourceDesc>

<sex>S</sex>

<abbr type="connection">ds</abbr>

# More about schemas

- Different schema formats available
  - TEI recommends RelaxNG schema format
  - Schematron for tighter restrictions
  - Roma will output other formats
- Schemas enable *datatyping*
- Users can combine schemas using *namespaces* (e.g., embed TEI in a METS file)

XML does not require a schema!

# Basic XML: *well-formed* vs. *valid*

<XML>Here is my <term>XML</XML>

Every opening tag must have a corresponding closing tag

# Basic XML: *well-formed* vs. *valid*

<XML>Here is my  <term>XML</XML>

# Basic XML: *well-formed* vs. *valid*

<XML>Here is my  <term>XML</XML>

<XML>Here is my <term>XML</XML></term>

A tag that opens after another one must close before the first one closes (aka, correct nesting)

# Basic XML: *well-formed* vs. *valid*

<XML>Here is my  <term>XML</XML>

<XML>Here is my <term>XML</XML>  </term>

# Basic XML: *well-formed* vs. *valid*

<XML>Here is my  <term>XML</XML>

<XML>Here is my <term>XML</XML>  </term>

<XML type=wellFormed>Here is my <term>XML</term></XML>


Attribute values must be enclosed in double quotes



# Basic XML: *well-formed* vs. *valid*

<XML>Here is my  <term>XML</XML>

<XML>Here is my <term>XML</XML>  </term>

 <XML type=wellFormed>Here is my <term>XML</term></XML>

# Basic XML: *well-formed* vs. *valid*


<XML>Here is my <term>XML</XML>

A large red 'X' mark is placed over the closing tag </XML> in the XML snippet, indicating it is not well-formed because the root element is not properly closed.

<XML>Here is my <term>XML</XML></term>

A large red 'X' mark is placed over the closing tag </term> in the XML snippet, indicating it is not well-formed because the root element <XML> is not properly closed.

<XML type=wellFormed>Here is my <term>XML</term></XML>

A large red 'X' mark is placed over the attribute value wellFormed in the XML snippet, indicating it is not well-formed because attribute values must be quoted.

<XML type="wellFormed" type="valid">Here is my <term>XML</term></XML>

Attributes may not repeat in a tag

# Basic XML: *well-formed* vs. *valid*


<XML>Here is my <term>XML</XML>

A large red 'X' mark is placed over the closing tag </XML> in the XML snippet, indicating it is not well-formed because the opening <term> tag is not closed.

<XML>Here is my <term>XML</XML></term>

A large red 'X' mark is placed over the closing tag </term> in the XML snippet, indicating it is not well-formed because the opening <XML> tag is not closed.

<XML type=wellFormed>Here is my <term>XML</term></XML>

A large red 'X' mark is placed over the attribute wellFormed in the XML snippet, indicating it is not well-formed because the attribute value is not quoted.

<XML type="wellFormed" type="valid">Here is my <term>XML</term></XML>

A large red 'X' mark is placed over the second type attribute in the XML snippet, indicating it is not well-formed because there are two type attributes, which is not allowed in XML.

# Basic XML: *well-formed* vs. *valid*


<XML>Here is my <term>XML</XML>



<XML>Here is my <term>XML</XML></term>



<XML type=wellFormed>Here is my <term>XML</term></XML>



<XML type="wellFormed" type="valid">Here is my <term>XML</term></XML>



<XML type="wellFormed valid">Here is my <term>XML</term></XML>

This is *well-formed* XML. It may also be *valid* if it conforms to (i.e., follows all the rules of) a schema



# Relationships in XML

<parent><child>...</child></parent>

<sibling1>...</sibling1>

<sibling2>...</sibling2>

<ancestor><parent><child>...</child></parent></  
ancestor>

# Using the TEI Guidelines

<http://www.tei-c.org/Guidelines/P5/>

At this point in the workshop we go to the Guidelines. We look at how the chapters are organized, and how the element and attribute indices work. The key to successfully using TEI is understanding how to use the Guidelines.

# Customizing the TEI

# “TEI Guidelines”

- The encoding scheme
  - 500+ tags (human-readable)
  - 21 modules (4 required)
  - Model classes and attribute classes

- Documentation
  - Describes the encoding scheme



# “TEI Guidelines”

- The encoding scheme
  - 500+ tags (human-readable)
  - 21 modules (4 required)
  - Model classes and attribute classes

- Documentation
  - Describes the encoding scheme

```
graph TD; A[TEI Guidelines] --> B[ROMA]; C[Documentation] --> B;
```

ROMA

The modules and documentation are placed in the Roma system

# “TEI Guidelines”

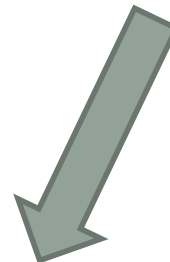
- The encoding scheme
  - 500+ tags (human-readable)
  - 21 modules (4 required)
  - Model classes and attribute classes

- Documentation
  - Describes the encoding scheme

Roma outputs a customized schema and documentation, depending on project needs

ROMA

Customized  
Schema +  
Documentation



# “TEI Guidelines”

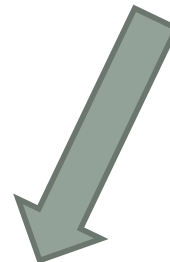
- The encoding scheme
  - 500+ tags (human-readable)
  - 21 modules (4 required)
  - Model classes and attribute classes

- Documentation
  - Describes the encoding scheme

Roma also outputs  
an ODD file...

ROMA

One Document  
Does it all  
(ODD)



# “TEI Guidelines”

- The encoding scheme
  - 500+ tags (human-readable)
  - 21 modules (4 required)
  - Model classes and attribute classes

- Documentation
  - Describes the encoding scheme

One Document  
Does it all  
(ODD)

ROMA

Customized  
Schema +  
Documentation

Which can be fed back into Roma to output the customized schema again.

# Learning more about the TEI

- TEI By Example: <http://www.teibyexample.org/>
- TEI@Oxford Teaching Pages:  
<http://tei.oucs.ox.ac.uk/Oxford>
- Teach Yourself TEI (at TEI website; out of date?):  
<http://www.tei-c.org/Support/Learn/tutorials.xml>
- Workshops:
  - DHSI (Victoria, BC): <http://www.dhsi.org/>
  - DigitalHumanities@Oxford Summer School:  
<http://digital.humanities.ox.ac.uk/dhoxss/>
  - DHWI (University of Maryland): <http://mith.umd.edu/dhwi/> [check for future workshops]
  - WWP Seminars on Scholarly Text Encoding (Brown U.):  
<http://www.wwp.brown.edu/outreach/seminars/>

# Questions?

- Dot Porter
- [dorp@upenn.edu](mailto:dorp@upenn.edu) or [dot.porter@gmail.com](mailto:dot.porter@gmail.com)
- [dotporterdigital.org](http://dotporterdigital.org)