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
- 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.)
Poetic lines and half-lines
Marginal notes
Etc.

What is a text?

Hwæt wē Gār-Dena in geār-dagum
ūod-cyninga ārym gefrūnon,
Hu ðā æþelingas ellen fremedon.
Oft Scyld Scēfing sceapēna þrēatum,
monegum mægbum meodo-setla ofþæah;
egsode Eorl[e], syðan ærest wearð
fēasceafte funden; hē þæs frōfre gebād:
wēox under wolcnûm, weord-þymendum þāh,
oðþæt him æghwylc þāra ym-b-sittendra
ofre hron-rāde Ḣyran 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
• 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.)
Poetic lines and half-lines: <lg>
<l n="1">…<caesura/>…</l>
<l n="2">…<caesura/>…</l>
</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 þyrm geфрūnōn,
hū ēā æþelingas ellīn fremedon.

Oft Scyld Scēfing sceæþena þrēatum,
5 moneγum meægbum meodo-setla ofteah;
egsode Eorl[e], syððan ærest wear∂
fēasceaf funden; hē ðæs frōfre gebād:
wēox under wolcnnum, weor∂-myndum þāh,
oðæt him æghwylc þāra ymb-sittendra
10 ofer hron-rāde hīrān 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
Why?

Why use TEI when you can just make a PDF?

Pembroke 25 Sermo 19 (fols. 36r–38r)

.xviii. OMELIA IN CAPUT
PRẹcautendum est omnibus bona opera IEIUNII.
exercentibus fratres karissimi, quod dominus dicit in æuan
gelio. Adtendite ne iustitiam uestram faciatis coram
hominibus ut uideamini ab eis. Aliquuin merce
dem non habebitis apud patrem uestrum qui est in caelis.
Cum ergo facis elimosina', noli tuba canere ante te
sicut hypocrite faciunt in sinagogis et in uicis ut
honorificentur ab hominibus. Amen dico ubis
receptor mortem suam. Te autem faciente elimo
Why?

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

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

16. xviii. OMELLA IN CAPUT

17. PREcaudum est omnibus bona opera IEIUNII.

18. exercentibus fratres kanissimi. quod dominus dicit in aevan

19. gelio. Adtendite ne iustitiam uestram facatis coram

20. hominibus ut uideamini ab eis. Alloquin merce

21. dcem non habebit apud patrem uestrum qui est in caulis.

22. Cum ergo facias elmosina. noli tuba canare

23. sicut hypocrite faciant in sinagogis et in uici.

Line 7: i

Addition to text.
You can build on it to link text to image (the examples uses UVIC Image Markup Tool).
You can use it to generate Linked Data (the example is the Canonical Text Service)

http://www.tei-c.org/Activities/Projects/
Introduction to XML
<tag> content </tag>

<author> Charles Darwin </author>

<date> 15 April, 1860 </date>

<underlined> especially </underlined>
Dear Sir,

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, and see what sort of insect visits them. I have received account of pollen-masses of this plant having been seen on proboscis of a day-moth; but I cannot help feeling a little sceptical about the identification. If you send any other 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; The Spiranthes would be especially valuable to me. I want to have pollen-masses for standard of comparison with those observed on the probosces of moths. —

The Spiranthes would be especially valuable to me. The Epipactus.

I shall return home on next Thursday (5th) (to Down, Bromley, Kent) on Monday 9th or 10th 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>
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, and see what sort of insect visits them. I have received account of pollen-masses of this plant having been seen on proboscis of a day-moth; but I cannot help feeling a little sceptical about the identification. If you send any other 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; I want to have pollen-masses for standard of comparison with those observed on the probosces of moths. —

The Spiranthes would be especially valuable to me especially the Epipactus. I shall return home on next Thursday (5th) to Down, Bromley, Kent) on Monday 9th or 10th 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

```xml
<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="1962-02-08">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="1976-02-08">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="1976-02-08">Dot Porter</name>
```

Once upon `<c>`<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

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

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

```xml
<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

```xml
<name when="1982-02-08">Dot Porter</name>

Once upon <c><w>a</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="1976-02-08">Dot Porter</name>
```

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

```
<sourceDesc><p>Venetus A</p></sourceDesc>
<sourceDesc><p>Venetus 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

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

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

<sourceDesc><p>Venetus A</p></sourceDesc>
<sourceDesc><p>Venetus B</p></sourceDesc>
<sex>YES</sex>
```
XML is defined by **schemas**

- Rules for a specified set of XML tags and attributes

```xml
<name when="1961-02-08">Dot Porter</name>

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

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

<sex>YES</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="1976-02-08">Dot Porter</name>

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

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

<sex>YES</sex>

<abbr type="contraction">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
<XML>
Here is my <term>XML</XML>
</XML>
```

Every opening tag must have a corresponding closing tag
Basic XML: well-formed vs. valid

<XML>Here is my <term>XML</term></XML>
Basic XML: **well-formed** vs. **valid**

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

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

```xml
<XML>
  Here is my <term>XML</XML></term>
</XML>
```
Basic XML: *well-formed* vs. *valid*

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

<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</term></XML>

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

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

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

```xml
<XML type="wellFormed">Here is my <term>XML</term></XML>
```
Basic XML: *well-formed* vs. *valid*

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

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

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

<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>`

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

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

- `<XML type="wellFormed" type="valid">Here is my <term>XML</term></XML>`
Basic XML: well-formed vs. valid

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

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

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

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…

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/](http://www.teibyexample.org/)
- **TEI@Oxford Teaching Pages:** [http://tei.oucs.ox.ac.uk/Oxford](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](http://www.tei-c.org/Support/Learn/tutorials.xml)
- **Workshops:**
  - **DigitalHumanities@Oxford Summer School:** [http://digital.humanities.ox.ac.uk/dhoxss/](http://digital.humanities.ox.ac.uk/dhoxss/)
  - **DHWI** (University of Maryland): [http://mith.umd.edu/dhwi/](http://mith.umd.edu/dhwi/) [check for future workshops]
Questions?

- Dot Porter
- dorp@upenn.edu or dot.porter@gmail.com
- dotporterdigital.org