Introduction to XML/TEI

Creating Digital Editions:
An Introduction to the Text Encoding Initiative
January 17, 2014

Why encode in TEI?

You need...

- "a simple data format that balances the needs of people to read/write data with the needs of machines to read/write data."

- a format for interchanging bits of structured data

- a standard way of annotating these documents.

- *...to make explicit certain features of text in order to aid the processing of that text by computer programs.*

What features?

Almost anything! Have to make choices based on your project's goals and resources:

- Making the information conveyed by a page layout readable to a machine? (capturing paragraphs as <p>, tagging decorative initial capitals)

- Adding scholarly interpretation that makes the text richer? (adding notes, or glosses, adding explanation for why/how various phenomena occur, associating a pronoun with the name of the person being referenced)

- Standardizing bits of the text so they can be processed more easily? (expanding abbreviations, associating dates and place names with standard formats)

To what end?

The goal is always to make the text readable to a machine—so what you choose to encode depends on what you want an application to be able to read.

Think about what aspects of the text are not susceptible to machine processing in their current format, and what the benefits would be of addressing that (dates, names, places)

Can be tricky: if you're encoding for unknown future uses.

Let's look at how some real life websites make use of the various kinds of encoding options we talked about.

Encoding text structure

- Enables sensible rendering of the text online
- Enables searching restricted to specific regions of the text

```
<DIV1 TYPE="dedication">
    <PB REF="2"/>
    <HEAD>TO <HI>ERASMUS EARLE,</HI>
        <ABBR>Esq</ABBR></HEAD>
    <OPENER><SALUTE>SIR,</SALUTE></OPENER>
    <P><SEG REND="decorInit">T</SEG>HE <HI>Weak</HI> naturally fly to the
        <HI>Brave</HI> for Protec|tion; and, methinks, a <HI>Play</HI> in which
    a <HI>Soldier</HI> carries off the <HI>Prize,</HI> has a sort of Right to a
        <HI>Patron</HI> out of the <HI>Camp.</HI> Your condescending to serve,
    for some little Time, as Subaltern Officer, while You are in the actual
    Possession of so Noble a Paternal Estate, is equall
    by All who have the Happiness to know You. This is
    an Honour to the <HI>Bri|tish</HI> Service; And how
    desir'd, that so Bright an Exam|ple, may be Imitate
    the best Families, and For|tunes.</P>
    <P>We have scarce an Instance in <HI>An|tient,</HI> or
        of any Hero who has <HI>Commanded</HI> WELL, with|o
        what it was to OBEY. The Greatest among the <HI>Gre
        <HI>Roman</HI> Generals, made their First Campa
        Voluntiers; and the Greatest <HI>Modern Hero,</HI>
        lamented, serv'd first in the <HI>same Post</HI> Yo
    <P>IF a Fine <HI>Person,</HI> a most Grace|ful, and Eas
    joyn'd to a great Knowledge of the <HT>World </HT>
```

Page [unnumbered]

THE ARTIFICE. A COMEDY.

As It Is ACTED At the THEATRE-ROYAL in Drury-Lane. By His *MAJESTY*'s Company of COMEDIANS.

By Mrs. *CENT-LIVRE*.

It is a Kind of *Slander*, to trust to RUMOUR.

Ben. Jonson's Cataline.

LONDON: Printed for T. PAYNE, near *Stationers-Hall*, M.DCC.XXIII. [Price 1 *s.* 6 *d.*]

Page [unnumbered]

TO *ERASMUS EARLE*, Esq

SIR,

THE *Weak* naturally fly to the *Brave* for Protec|tion; and, methinks, a *Play* in which a *Soldier* carries off the *Prize*, has a sort of Right to a *Patron* out of the *Camp*. Your condescending to serve, for some little Time, as Subaltern Officer, while You are in the actual Possession of so Noble a Paternal Estate, is equally Admir'd, and Applauded, by All who have the Happiness to know You. This is

Encoding text structure

- Enables sensible rendering of the text online
- Enables searching restricted to specific regions of the text

Normalizing non-standard forms

- Enables thorough searching and sorting
- Allows simultaneous fidelity to the source and ease of processing
- Enables creation of visualizations like maps and timelines

```
</choice> Primate <lb/>of all <choice>
    <abbr>Engl:</abbr>
    <expan>England</expan>
</choice> &amp; Metropolitan &amp; to our right
    <abbr>Councell<hi rend="superscript">r</hi><
    <expan>Councellor</expan>
</choice>
<lb/>George <choice>
    <abbr>L<hi rend="superscript">d</hi></abbr>
    <expan>Lord</expan>
</choice> Ieffreys <choice>
    <abbr>L<hi rend="superscript">d</hi>.</abbr>
    <expan>Lord</expan>
</choice>
<choice>
    <abbr>Ch<hi rend="superscript">r</hi>.</abbr>
    <expan>Chancellor</expan>
</choice> of <choice>
    <abbr>Engl:</abbr>
    <expan>England</expan>
```

1. Iames the 2nd by ye Grace of God of England, Scotl: France & Ireland King, Defender of ye faith &c. To ye most Reverend Father in God our right trusty & right well-beloved Councellr Wm Ld Archbishop of Cant: Primate of all Engl: & Metropolitan & to our right Trusty & right well-beloved Councellr George Ld Ieffreys Ld. Chr. of Engl: & to our right Trusty & right well-beloved Cosin & Councellr Laurence E. of Rochester, Ld high Treasurer of Engl: & to our right Trusty & right well-beloved Cosin & Councellr Robt E. of Sunderland, Prsident of our Councel & our principall Secretary of State, Nathaniel Ld. BP of Durham, & to ye right Reverend Father in God our Trusty & well-beloved Councellr. Sr Edw: Herbert Kt. Cheif Iustice of the Pleas before us to be holden assign'd Greeting.

1. Iames the 2nd by the Grace of God of England, Scotland France & Ireland King, Defender of the faith &c. To the most Reverend Father in God our right trusty & right well-beloved Councellor William Lord Archbishop of Canterbury Primate of all England & Metropolitan & to our right Trusty & right well-beloved Councellor George Lord Ieffreys Lord Chancellor of England & to our right Trusty & right well-beloved Cosin & Councellor Laurence Earl of Rochester, Lord high Treasurer of England & to our right Trusty & right well-beloved Cosin & Councellor Robert Earl of Sunderland, President of our Councel & our principall Secretary of State, Nathaniel Lord Bishop of Durham, & to the right Reverend Father in God our Trusty & well-beloved Councellor Sir Edward Herbert Knight Cheif Iustice of the Pleas before us to be holden assign'd Greeting.

http://www.newtonproject.sussex.ac.uk/

Adding interpretation/glossing

- Enables references between text and external factors such as people or known events
- Richly encoded file becomes the basis of your edition

http://www.vangoghletters.org/vg/

XML in brief (1)

Open, non-proprietary standard

Stored in plain text but usually thought of as contrasting with it

Marks beginning and ends of spans of text using tags:
<sentence>This is a sentence.</sentence>

XML in brief (2)

Spans of text must nest properly:

Wrong:
\<sentence\>Overlap is \<emphasis\>not allowed!\</sentence\>\</emphasis\>

Right:
\<sentence\>Overlap is \<emphasis\>not allowed!\</emphasis\>\</sentence\>

Elements (tags), attributes, values, content

<sentence type="declarative">This is a sentence.</sentence>

<sentence type="interrogative">Is this is a sentence?</sentence>

Elements (tags), attributes, values, content

Elements may have one attribute, many attributes, or none, but each attribute on any given element must be unique.

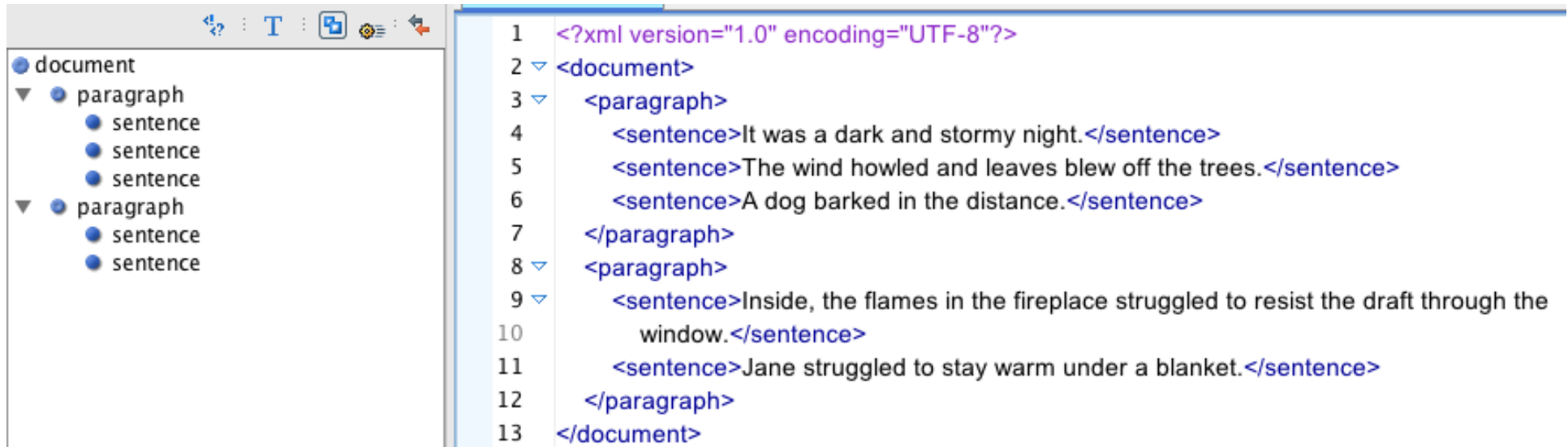Valid: <sentence type="declarative">This is a sentence.</sentence>

Valid: <sentence type="interrogative" xml:lang="en">Is this is a sentence?</sentence>

Valid: <sentence>This is a sentence.</sentence>

Invalid: <sentence type="declarative" type="true">This is a sentence.</sentence>

XML as a tree

- We use family tree terms: parent, child, sibling, ancestor, and descendent.
- *Remember, everything must nest properly!*

```
document
▼ ● paragraph
     ● sentence
     ● sentence
     ● sentence
▼ ● paragraph
     ● sentence
     ● sentence
```

```
1    <?xml version="1.0" encoding="UTF-8"?>
2  ▽ <document>
3  ▽    <paragraph>
4          <sentence>It was a dark and stormy night.</sentence>
5          <sentence>The wind howled and leaves blew off the trees.</sentence>
6          <sentence>A dog barked in the distance.</sentence>
7      </paragraph>
8  ▽    <paragraph>
9  ▽       <sentence>Inside, the flames in the fireplace struggled to resist the draft through the
10              window.</sentence>
11         <sentence>Jane struggled to stay warm under a blanket.</sentence>
12     </paragraph>
13 </document>
```

Structure, not appearance

Most people use XML to describe the structure of a document rather than its appearance. Information about how to render various components of the document is usually stored separately, in a *stylesheet*.

Questions?