

Introduction to XML/TEI

Why encode a text in TEI?

You need...

- a system that allows annotations of the text in order to label the various components
- "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 Protection; 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 equally
    by All who have the Happiness to know You. This is
    an Honour to the <HI>British</HI> Service; And how desir'd, that so Bright an Example, may be Imitate
    the best Families, and Fortunes.</P>
  <P>We have scarce an Instance in <HI>Antient,</HI> or
    of any Hero who has <HI>Commanded</HI> WELL, withlo
    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 Gracelful, and Eas
    ioun'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 Protection; 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

```
<SP>
  <PB N="2" REF="9"/>
  <SPEAKER>Sir John.</SPEAKER>
  <P>What Game are you in Pursuit of?</P>
</SP>
<SP>
  <SPEAKER>Fain.</SPEAKER>
  <P>The <HI>old standing Dish,</HI> a WIFE! Here's no Hopes of
    <HI>War,</HI> no <HI>Rising</HI> in View! and
    <HI>Subaltern</HI>'s Pay will make a Man <HI>rub but slowly thr
    the World.</HI> I have got Leave of my Colonel to be absent
    my Post for Two Months, in which Time, <HI>I intend to make my
    Fortune.</HI> -You know the Widow <HI>Headless</HI>!</P>
</SP>
<SP>
  <SPEAKER>Sir John.</SPEAKER>
  <P>Very well! Ha! -is not that <HI>her Livery?</HI></P>
</SP>
"
```

Eighteenth Century Collections Online
Text Creation Partnership

Home Search Browse Bookbag Help

Basic Search

Basic Search Boolean Proximity Bibliographic History

Search in full text
Find addresses and signatures
Tip drama
illustrations
notes
section and work titles
subjects
verse
ESTC number

Limit to in author

Restrict by date range: 1700 and 1800

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>Councill<hi rend="superscript">r</hi></abbr>
  <expan>Councillor</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. James the 2nd by y^e Grace of God of England, Scotl: France & Ireland King, Defender of y^e faith &c. To y^e most Reverend Father in God our right trusty & right well-beloved Councill^r W^m L^d Archbishop of Cant: Primate of all Engl: & Metropolitan & to our right Trusty & right well-beloved Councill^r George L^d Ieffreys L^d. Ch^r. of Engl: & to our right Trusty & right well-beloved Cosin & Councill^r Laurence E. of Rochester, L^d high Treasurer of Engl: & to our right Trusty & right well-beloved Cosin & Councill^r Rob^t E. of Sunderland, P^rsident of our Council & our principall Secretary of State, Nathaniel L^d. BP of Durham, & to y^e right Reverend Father in God our Trusty & well-beloved Councill^r. S^r Edw: Herbert K^t. Cheif Iustice of the Pleas before us to be holden assign'd Greeting.

1. James 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 Councillor William Lord Archbishop of Canterbury Primate of all England & Metropolitan & to our right Trusty & right well-beloved Councillor George Lord Ieffreys Lord Chancellor of England & to our right Trusty & right well-beloved Cosin & Councillor Laurence Earl of Rochester, Lord high Treasurer of England & to our right Trusty & right well-beloved Cosin & Councillor Robert Earl of Sunderland, President of our Council & our principall Secretary of State, Nathaniel Lord Bishop of Durham, & to the right Reverend Father in God our Trusty & well-beloved Councillor Sir Edward Herbert Knight Cheif Iustice of the Pleas before us to be holden assign'd Greeting.

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

```
<div type="translation">
  <pb f="1r" n="1" xml:id="pb-trans-1r-1" facs="#zone-pb-1r-1"/>
  <ab>My dear Theo,</ab>
  <ab>Many thanks for your letter and 200 francs enclosed.</ab>
  <ab rend="indent">Thanks for letting me know the size of the frame, for
    which I'm thinking of making a little woman spinning, after the large
    study.<anchor n="1" xml:id="note-t-1"/></ab>
  <ab rend="indent">I was pleased to hear good reports about <rs type="pers"
    key="903">Breitner</rs> for once.<anchor n="2" xml:id="note-t-2"
  /> As you know, the last impressions that I had of him were rather
  unfavourable as a consequence of 3 large canvases that I saw at his
  place and in which I literally saw nothing that one could locate
  either in reality or in an imaginary world.<anchor n="3"
  xml:id="note-t-3"/> Still, a few watercolours he was working on
  at the time – horses in the dunes – were better, although very
  sketchy.<anchor n="4" xml:id="note-t-4"/> And I saw in them things
  that mean I can readily understand that the painting you talk
  about is good.</ab>
```

My dear Theo,

Many thanks for your letter and 200 francs enclosed.

Thanks for letting me know the size of the frame, for which I'm thinking of making a little woman spinning, after the large study. ¹

I was pleased to hear good reports about [Breitner](#) for once. ² As you know, the last impressions that I had of him were rather unfavourable as a consequence of 3 large canvases that I saw at his place and in which I literally saw nothing that one could locate either in reality or in an imaginary world. ³ Still, a few watercolours he was working on at the time – horses in the dunes – were better, although very sketchy. ⁴ And I saw in them things that mean I can readily understand that the painting you talk about is good.

As to the drawing society: ⁵ first, I'd utterly forgotten about it because I was painting those particular figures:

451

Br. 1990: 454 | CL: 372


From: Vincent van Gogh

To: Theo van Gogh

Date: Nuenen, on or about

Wednesday, 2 July 1884

[more...](#)

¹. This 'large' painting of a woman spinning must be the same one referred to in [letters 449](#) and [450](#). The 'little woman spinning' may have been the woman spinning that is under *The parsonage garden in the snow* (F 194 / JH 603 ), which measures 59 x 78 cm. For an X-ray image of this see exhib. cat. Vienna 1996, p. 172. Cf. also [letter 466](#).

². [Breitner](#) stayed in Paris, where he met Theo, from June to November 1884. See [Hefting 1970](#), p. 75.

³. See for these three large works by [Breitner](#), one of which

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!</emphasis></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>`

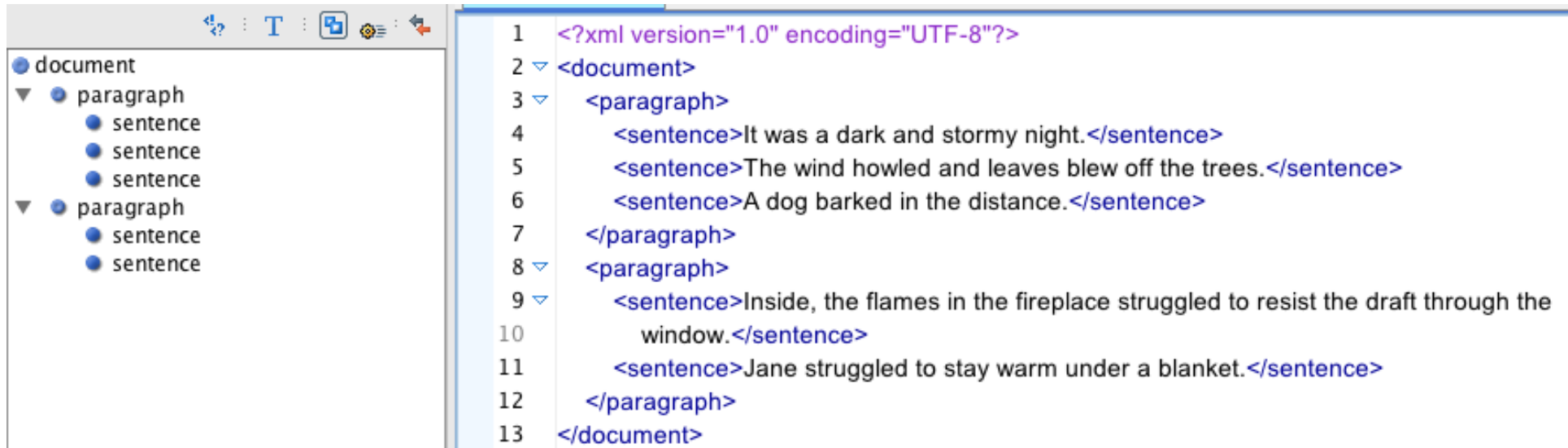
Valid: `<sentence type="interrogative" xml:lang="en">Is this is a sentence?</sentence>`

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!*



The image shows a screenshot of an XML editor. On the left, there is a tree view showing the document structure. The root node is 'document', which contains two 'paragraph' nodes. The first 'paragraph' node contains three 'sentence' nodes, and the second 'paragraph' node contains two 'sentence' nodes. On the right, the XML code is displayed with line numbers 1 through 13. The code is as follows:

```
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?