Comp 336/436 - Markup Languages

Fall Semester 2017 - Week 14

Dr Nick Hayward

Contents

- Final demo and presentation
- Final report
- Group updates
- Text Encoding Initiative - part 2

Final Demo and Presentation

- continue to develop your application/project concept and prototypes
- working app (as close as possible...)
- or project research and development
- clearly detail
- technology choices
- metadata schemes &c.
- manipulation and rendering of markup language
- choice of semantic organisation and usage
- schema usage
- ...
- show and explain implemented differences from DEV week
- where and why did you update the app?
- perceived benefits of the updates?
- how did you respond to peer review?
- ...

n.b. present your own work contributed to the project, and its development...

Final Report

Report due on 13th December 2017 by 4.15pm

- final report outline - coursework section of website
  - *PDF*
  - *group report*
  - ***extra individual report*** *- optional*

Group Updates

- what is currently working?
- which technologies?
- processing options and rendering choices?
  - *e.g. XSLT to HTML5*
- any external or custom schemas?
- what is left to add or fix?
- who is working on what?
  - *design, testing, research...*

Groups A, C, D, E, & F

Text Encoding Initiative - general usage

Exercise - part 1

```
<TEI>
  <teiHeader>
    <fileDesc>
      <titleStmt>
        <title>
          <!--Title-->
        </title>
      </titleStmt>
      <publicationStmt>
        <p>
          <!--Publication Information-->
        </p>
      </publicationStmt>
      <sourceDesc>
        <p>
          <!--Information about the source-->
        </p>
      </sourceDesc>
    </fileDesc>
  </teiHeader>
  <text>
    <body>
      <!--Some structural division, paragraph, line group, speech, ...-->
    </body>
  </text>
</TEI>
```

- using the above template
  - *add some of the missing information - as commented in the XML*
  - *write a simple XSL stylesheet to render the content and metadata to HTML*
  - *test in a browser*

Text Encoding Initiative - general usage

- within the `<teiHeader>` we could add
  - *the title of the text*
  - *`<respStmt>`*
  - *`<publicationStmt>`*
  - *`<sourceDesc>`*
- within `<text>` we could add
  - *`<front>` with a dedication and table of contents...*
  - *`<body>` with transcription etc of the text*
  - *`<back>` with a div of type 'colophon'...*

Text Encoding Initiative - general usage

<teiHeader> example

```
<teiHeader>
  <fileDesc>
    <titleStmt>
      <title>The Epic of Gilgamesh</title>
      <respStmt>
        <resp>editor</resp>
        <name xml:id="AL">Alasdair Livingstone</name>
      </respStmt>
    </titleStmt>
    <publicationStmt>
      <p>Not for distribution.</p>
    </publicationStmt>
    <sourceDesc>
      <p>Transcribed from the diaries of Professor Wilfred Lambert</p>
    </sourceDesc>
  </fileDesc>
</teiHeader>
```

Text Encoding Initiative - general usage

<text> example

```
<text>
  <front>
    <div type="dedication">
      <p>personal dedication...</p>
    </div>
    <div type="contents">
      <head>Table of Contents</head>
      <list>
        <item>1. No.1...</item>
        <item>2. No.2...</item>
        <item>3. No.3...</item>
      </list>
    </div>
 </front>
 <body>
    <p>some body text goes here....</p>
  </body>
 <back>
    <div type="colophon">
      <p>Physical book conditions...</p>
    </div>
 </back>
</text>
```

Text Encoding Initiative - textual phenomena

structural features

- Prose elements, e.g.

```
<p> <div> <head> <list> <item> <q> <pb> <seg> <figure> <tables>
```

- Verse elements, e.g.

```
<lg> <l>
```

- Drama elements, e.g.

```
<div> <sp> <p> <lg> <l> <seg>
```

Text Encoding Initiative - textual phenomena

structural features - examples

- list

```
<list>
  <item></item>
</list>
```

- line groupings

```
<lg type="free">
  <l></l>
</lg>
```

- segments

```
<seg type="preamble">
  <seg></seg>
  <seg></seg>
</seg>
```

- quotations

```
<q></q>
```

- an example with 'quoted' q

Text Encoding Initiative - textual phenomena

title pages

- `<titlePage>` used for transcription and encoding of physical title page
- within `<front>` or `<back>` using `<titlePage>`
- `<titlePage>` commonly includes `<docTitle>`
- `<docTitle>` may include subsections, `<titlePart>`
- `<docAuthor>` for name of the author
- `<byline>` for primary statement of responsibility for a work
- other elements can include
  - *`<docEdition>`*
  - *`<docDate>`*
  - *`<docImprint>` may contain*
  - *`<pubPlace>`*
  - *`<docDate>`*
  - *`<publisher>`*
  - …
  - *`<epigraph>`*
  - *`<imprimatur>`*
  - *`<graphic/>`*

Text Encoding Initiative - textual phenomena

title pages - examples

- <docTitle> can consist of several subsections
- <titlePart>, which may itself use a @type attribute documenting their role.

```
<docTitle>
  <titlePart type="main">main title...</titlePart>
  <titlePart type="sub">sub-title</titlePart>
</docTitle>
```

- an inline graphic, illustration, or figure <graphic/>

```
<figure>
 <graphic url="fig1.png"/>
 <head>.</head>
 <figDesc>.</figDesc>
</figure>
```

Exercise - part 2 - Front & Title Page with Image

Use the image on the next slide to encode a title page in the front part of a text.

Encode the following:

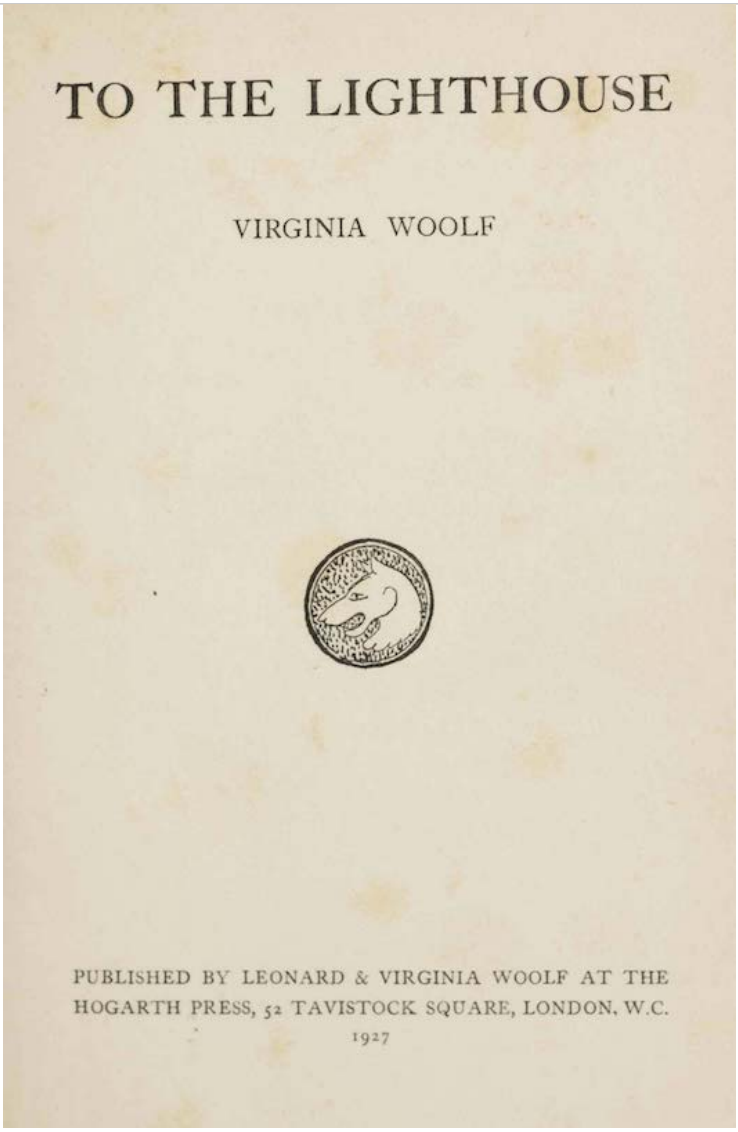
1. title and author information
2. any stylistic considerations in the font, size, spacing...
3. the graphic
4. any other appropriate information such as edition, publication information...

Please feel free to consult the TEI guidelines where necessary.

- ~ 15 minutes

Image - To The Lighthouse

front page



To the Lighthouse - GB 1st Edition

front & title page example

A simple example of encoding a front and title page

```
<TEI xmlns="http://www.tei-c.org/ns/1.0">
  <teiHeader>
    <fileDesc>
      <titleStmt>
        <title/>
      </titleStmt>
      <publicationStmt>
        <publisher/>
      </publicationStmt>
      <sourceDesc>
        <p/>
      </sourceDesc>
    </fileDesc>
  </teiHeader>
  <text>
    <front>
      <titlePage>
        <docTitle>
          <titlePart type="main">TO THE LIGHTHOUSE</titlePart>
        </docTitle>
        <byline>
          <docAuthor>Virginia Woolf</docAuthor>
        </byline>
        <figure>
          <graphic url="hogarth-logo.jpg"/>
          <figDesc>Black and white rendition of Hogarth Logo</figDesc>
        </figure>
        <docImprint>
          <publisher>PUBLISHED BY LEONARD &AMP; VIRGINIA WOOLF AT THE HOGARTH PRESS</publisher>,
          <pubPlace>52 TAVISTOCK SQUARE, LONDON. W.C.</pubPlace>,
          <docDate>1927</docDate>
        </docImprint>
      </titlePage>
    </front>
    <body>
    </body>
  </text>
</TEI>
```

n.b. in the <docImprint> element

- for the <publisher> element - we could be more specific and use <name> elements
- e.g. for Leonard and Virginia Woolf
- for the <pubPlace> element - we could enclose this within an <address> element
- this might help to add further structure and metadata
- for the <docDate> element - we could add a @when attribute with value set to the year 1927

- again, it helps to add further metadata

Text Encoding Initiative - textual phenomena

logical and semantic features

- encodes underlying semantics with specific elements other than `<hi>`
- highlighting may be used to show

```
<emph> <foreign> <distinct>
```

- uses of quotation

```
<said> <quote> <cit> <mentioned> <soCalled>
```

- other examples include

```
<term> <gloss>
```

- generic elements may also carry semantic and logical information

```
<title> <name> <num> <measure> <date> <address> <abbr> <expan>
```

Text Encoding Initiative - textual phenomena

logical and semantic features - referring strings

- in TEI
  - *places, people, objects can use the <name> elements with @type attribute*
  - *e.g.*

```
<name type="person">Jack</name>
```

- proper nouns can also be referred to using the <rs> element plus @type attribute
  - *e.g.*

```
<rs type="person">he</rs>
```

Exercise - part 3 - common structure and elements

Use the excerpt on the next slide to encode a structured grouping of text, e.g. a division, segment, page...

Encode the following:

1. this text as a paragraph of prose
2. indicate that the language of this paragraph is English
3. indicate that the paragraph is number 7
4. encode any abbreviations with their correct expansions
5. encode names with the appropriate element and attribute

Please feel free to consult the TEI guidelines where necessary.

- ~ 15 minutes

Text Encoding Initiative - general usage

Exercise - part 3 - text excerpt

- encode the following text excerpt

```
  "Good-evening, Mrs. McNab," she would say.
She had a pleasant way with her. The girls
all liked her. But dear, many things had changed
since then (she shut the drawer); many families
had lost their dearest. So she was dead; and
Mr. Andrew killed; and Dr. Prue dead too,
they said, with her first baby; but every one had
lost some one these years. Prices had gone up
shamefully, and didn't come down again neither.
She could well remember her in her grey cloak.
```

text excerpt - element encoding of local language

A simple example of a text excerpt with element defined local language encoding.

```
<TEI xmlns="http://www.tei-c.org/ns/1.0">
  <teiHeader>
    <fileDesc>
      <titleStmt>
        <title/>
      </titleStmt>
      <publicationStmt>
        <publisher/>
      </publicationStmt>
      <sourceDesc>
        <p/>
      </sourceDesc>
    </fileDesc>
  </teiHeader>
  <text>
    <body>
      <p xml:lang="en" n="7">"Good-evening,
        <name type="person">
          <choice>
            <abbr type="title">Mrs.</abbr>
            <expan>Mistress</expan>
          </choice>
        McNab</name>
      ," she would say.</p>
    </body>
  </text>
</TEI>
```

n.b.

- `en-GB` - the value must conform to BCP 47 (tags for identifying languages)
- the `<rs>` element with a `@type` attribute can also be used instead of `<name>`
- `<name>` is solely for proper nouns
- the rest of the paragraph in the example follows the same pattern

text excerpt - header encoding of global language

A simple example of a text excerpt with header defined global language encoding.

```
<TEI xmlns="http://www.tei-c.org/ns/1.0">
  <teiHeader>
    <fileDesc>
      <titleStmt>
        <title/>
      </titleStmt>
      <publicationStmt>
        <publisher/>
      </publicationStmt>
      <sourceDesc>
        <p/>
      </sourceDesc>
    </fileDesc>
    <profileDesc>
      <langUsage>
        <language ident="en-GB">British English</language>
      </langUsage>
    </profileDesc>
  </teiHeader>
  <text>
    <body>
      <p n="7"[5]>"Good-evening,
        <name type="person">
          <choice>
            <abbr type="title">Mrs.</abbr>
            <expan>Mistress</expan>
          </choice>
        McNab</name>
      ," she would say.</p>
    </body>
  </text>
</TEI>
```

n.b.

- header encoding of global language this helps specify the languages used throughout the text being encoded
- we could specify the language used, as in the above example, but now it would only make sense if the language was different from the specified default in the TEIHeader.

Text Encoding Initiative - textual phenomena

logical and semantic features - dates and time

- <date> and <time>
- system or calendar used may be documented using @calendar attribute
- value of date or time supplied using @when attribute
- normalised representation of <date> should conform to valid W3C datatype

```
<date when="2003-12-22" calendar="Gregorian">22 Nov 2003</date>
```

- <date> can also be used to mark a period of time using the attributes
  - *@from*
  - *@to*
  - *@notBefore*
  - *@notAfter*

Text Encoding Initiative - textual phenomena

logical and semantic features - numbers and measures

- <num>

```
<num type="percentage" value="23">23%</num>
```

- <measure>

```
<measure type="volume" quantity="1.5" unit="litre" commodity="wine">1.5L bottle of wine</measure>
```

Exercise - part 4

If **20,000 Leagues Under the Sea** was an actual recording of the distance travelled by Captain Nemo in the Nautilus, how would you specify this using TEI?

n.b. league = ~ 3 nautical miles or 3.5 standard English miles

- create a TEI encoded XML document
  - *add some details for the novel **20,000 Leagues Under the Sea** by **Jules Verne***
  - *correctly encode **20,000 leagues***
  - *write a simple XSL stylesheet to render the content and metadata to HTML*
  - *test in a browser*
- ~ 15 minutes

Text Encoding Initiative - textual phenomena

logical and semantic features - addresses

- physical and digital addresses can be encoded
  - *<address>*, *<email>*
- <email> can also use a @type attribute
- <addrLine> within <address>
  - *<name>*, *<street>*, *<postCode>*, *<postBox>*

Text Encoding Initiative - textual phenomena

addresses - examples

```
<address>
  <addrLine></addrLine>
</address>
```

- you can also be specific, and use semantically rich elements

```
<name>
<street>
<postCode>
<postBox>
```

Text Encoding Initiative - textual phenomena

Exercise - part 5

- how would you encode **Alice's** address in **Wonderland** using TEI?
- create example TEI XML and encode this **fictional** address
- write a simple XSL stylesheet to render the content and metadata to HTML
- test in a browser
- ~ 10 minutes

Text Encoding Initiative - textual phenomena

logical and semantic features - abbreviations and expansions

- explicitly encode
  - *<abbr>*
  - *<expan>*
- `@type` attribute may be used with `<abbr>`
  - *e.g.*

```
<abbr type="title">Dr</abbr>
```

- `<expan>` used with `<abbr>` within `<choice>`

Text Encoding Initiative - textual phenomena

abbreviations and expansions - example

```
<choice>
  <abbr type="title">Dr</abbr>
  <expan>Doctor</expan>
</choice>
```

Text Encoding Initiative - textual phenomena

Exercise - part 6

- how might you use the `choice` element to encode **PhD**?
- use TEI encoded XML
- ~ 10 minutes

Text Encoding Initiative - textual phenomena

analytical features - notes and annotations

- `<note>` can be used to record a textual annotation
- `@type` attribute used to differentiate notes
- `@resp` attribute used to assign responsibility
- position of the note can be referenced using the `@place` attribute

e.g.

```
<note n="1" place="foot" type="editorial" resp="NJH">a new note...</note>
```

Text Encoding Initiative - textual phenomena

analytical features - index entries

- pre-existing in <front> or <back>
  - *using <list> inside *
- new index using <term> inside <index>
  - *add at the location of the index item*
  - *e.g.*

```
<index>
  <term>new index term...</term>
</index>
```

Text Encoding Initiative - textual phenomena

analytical features - errors

- indicated using `<sic>`
- corrected using `<corr>`
- combine `<sic>` and `<corr>` within `<choice>`
- use `@cert`and `@resp` attributes to encode degree of certainty and editor responsible
- e.g.

```
<corr cert="high" resp="#NJH">correction...</corr>
```

- **n.b.** *hash* in the `@resp` attribute value
  - *a pointer to a name element in the `<teiHeader>`*
  - *e.g.*

```
<name xml:id="NJH">...</name>
```

Text Encoding Initiative - textual phenomena

Exercise - part 7

- how would you add a pointer to a name element in the `<teiHeader>`?
- use TEI encoded XML
- ~ 10 minutes

<respStmt> - example

- place this name within the <respStmt> element as follows,

```
<respStmt>
  <resp>editor</resp>
  <name xml:id="NJH">Nicholas J Hayward</name>
</respStmt>
```

- then reference this editor as necessary in the encoded document

Text Encoding Initiative - textual phenomena

analytical features - regularisation

- <reg> element for regularisation
- <orig> for original, non-normalised form
- use <reg> in isolation or combined with <orig> within <choice>
- e.g.

```
<choice>
  <orig>thou</orig>
  <reg resp="#NJH">you</reg>
</choice>
```

Text Encoding Initiative - textual phenomena

analytical features - additions, deletions & omissions

- <gap> element used for omission, both material and editorial
- @reason attribute used to indicate reason for omission
- @extent and @unit attributes can be used to record extent of omission
- editorial omissions should be recorded using <editorialDecl>
  - *add inside <editorialDesc> in <teiHeader>*
- <gap> may be empty or include a <desc> of the material omitted
- <add> and <del> may also be used for words and phrases
- @rend attribute may also be used with <add> and <del>
- <addSpan/> and <delSpan/> for longer passages
- <subst> to contain <add> and <del> with causal relationship
- <unclear> with @reason attribute for difficult to read deletions in the text

Text Encoding Initiative - non-textual phenomena

graphics &c.

- graphics such as illustrations, diagrams, drawings, artwork...
- anchor in a text using `<graphic/>` and optional `@url` attribute
- e.g.

```
<graphic url="http://www.somewhere.com/image.jpg"/>
```

- use `<figure>` element as parent to create a full listing for a graphic

Text Encoding Initiative - Attributes

global Attributes

Global Attributes are currently as follows:

- @cert
  - *provided by* `responsibility` *subclass*
- @n
- @rend
  - *provided by* `rendition` *subclass*
- @rendition
  - *provided by* `rendition` *subclass*
- @resp
  - *provided by* `responsibility` *subclass*
- @source
  - *provided by* `source` *subclass*
- @style
  - *provided by* `rendition` *subclass*
- @xml:base
- @xml:id
- @xml:lang
- @xml:space

Text Encoding Initiative - Further Examples

bibliography listing

- Sample from Verne Digital Corpus

Text Encoding Initiative - Further Examples

regularise

- regularise with soundex and metaphone
- regularise with soundex and metaphone - basic compare
- regularise with soundex and metaphone - array compare

Text Encoding Initiative - Further Examples

working with images

- working with image and transcription
  - *representation of primary sources*
- TEI
  - *Facsimile - guidelines & examples*
  - *Surface - guidelines & examples*
  - *Zone - guidelines & examples*

References

- TEI P5 Guidelines
  - *representation of primary sources*
  - *facsimile*
  - *surface*
  - *zone*