Comp 336/436 - Markup Languages

Fall Semester 2019 - Week 12

Dr Nick Hayward

Text Encoding Initiative - Consortium

- maintains a technical standard
- set of guidelines
- user wiki
- set of tools for development and processing
- currently predominant in social sciences and humanities, in particular
  - *textual studies*
  - *literary studies*
  - *linguistic studies*

Text Encoding Initiative - Contribute

- membership
- Special Interest Groups (SIG)
  - *currently 12 active groups*
  - *wiki and mailing list*
- Activities
  - *various boards and membership groups*
  - *conferences*
  - *workgroups*
  - *project information*

Text Encoding Initiative - Guidelines

a gentle introduction

- guidelines for electronic text encoding and interchange
- define and document a markup language for representing
  - *structural*
  - *rendition*
  - *logical & semantic*
  - *analytic*
- rules and recommendations rather than standard
- wide variety of possible solutions for encoding material
- freedom of expression of personal textual theory
- current version = P5
- ~ 569 elements, ~ 214 attributes...
- customisations such as TEI Lite

Text Encoding Initiative - Text Structure

- structural features
  - *organisation of information in text*
- TEI defines overall text structure using
  - *front*
  - *body*
  - *group*
  - *back*
- a consideration of rendition features

Text Encoding Initiative - Text Structure

some examples

Example structural features commonly found in prose, verse, and drama.

Prose

- paragraphs <p>
- divisions
- headings <head>
- lists <list>
- list item <item>
- quotations <q>
- page breaks <pb>
- segments <seg>
- figures <figure>
- tables <table>

Verse

- linegroups <lg>
- lines <l>

Drama

- divisions
- speeches <sp>
- paragraphs <p>
- linegroups <lg>
- lines <l>

- segments <seg>

Text Encoding Initiative - Text Structure

conceptual usage

- how might we initially consider structuring a document
- a book
  - *mainly consists of chapters, sections, paragraphs...*
- poetry
  - *commonly organised in poems, stanzas, lines...*
- performance texts
  - *often think in terms of scenes, acts, or parts of speech*

Text Encoding Initiative - Rendition Features

- rendition features such as
  - *distinct fonts*
  - *colours*
  - *alignments*
  - *italics, underline, bold*
  - *font weight...*
- highlighting

```
<hi>italic words...</hi>

<hi rend="italic">italic words...</hi>
```

- Malory Project - XML example

Text Encoding Initiative - Logical and Semantic features

- logical and semantic features such as
  - *emphasis*
  - *foreign words*
  - *linguistically distinct words, phrases*
- we can also consider
  - *quotation marks*
  - *quotes*
  - *cited quotation...*

Text Encoding Initiative - Analytical Features

- analytical features such as
  - *notes and comments*
  - *marking for indexing*
  - *regularisation*
  - *editorial statements*

Text Encoding Initiative - Modules

part 1

1. TEI Infrastructure (tei)
2. Common metadata (header)
3. Common Core (core)
4. Default text structure (textstructure)
5. Character and Glyph Documentation (gaiji)

- TEI guidelines - module list
- TEI Header example

Text Encoding Initiative - Modules

part 2

6. Verse (verse)
7. Performance texts (drama)
8. Transcribed speech (spoken)
9. Print dictionaries (dictionaries)
10. Manuscript description (msdescription)

Text Encoding Initiative - Modules

part 3

11. Transcription of primary sources (transcr)
12. Text criticism (textcrit)
13. Names, Dates, People, and Places (namesdates)
14. Tables, Formulae, Figures (figures)
15. Metadata for language corpora (corpus)

Text Encoding Initiative - Modules

part 4

16. Linking, segmentation, and alignment (linking)
17. Analysis and interpretation (analysis)
18. Feature structures (iso-fs)
19. Graphs, networks, and trees (nets)
20. Certainty and Uncertainty (certainty)
21. Documentation elements (tagdocs)

Text Encoding Initiative - getting started

Common Data Structure & Elements

- expressed using XML
- elements for information
- attributes for additional information
- comments use standard delimiters

```
<!--this is a comment in TEI-->
```

Text Encoding Initiative - getting started

TEI Document

- `<teiHeader>` - documents all metadata for the TEI document
- `<text>` - contains the document proper
- these elements are mandatory for all TEI documents
- this structure is contained within the `<TEI>` element

```
<TEI>
  <teiHeader>...</teiHeader>
  <text>...</text>
</TEI>
```

Text Encoding Initiative - getting started

TEI Header `<teiHeader>`

- mandatory for all TEI documents
- `<fileDesc>` - contains a description of the electronic file
- `<fileDesc>` consists of at least three mandatory elements
  - *`<titleStmt>`*
  - *`<publicationStmt>`*
  - *`<sourceDesc>`*

Text Encoding Initiative - getting started

TEI Header `<teiHeader>` example

```
<teiHeader>
  <fileDesc>
    <titleStmt>
      <title>Around the World in Eighty Days</title>
      <respStmt>
        <resp>editor</resp>
        <name xml:id="NJH">Nicholas J Hayward</name>
      </respStmt>
    </titleStmt>
    <publicationStmt>
      <p>Not for distribution.</p>
    </publicationStmt>
    <sourceDesc>
      <p>Transcribed from 1873 English Edition</p>
    </sourceDesc>
  </fileDesc>
</teiHeader>
```

Text Encoding Initiative - getting started

global attribute usage - `xml:id`

```
<!-- transcriber note -->
<note resp="#NJH">Possible mis-interpretation of author correction.</note>

<!-- editorial note -->
<note resp="#AL">correction meant as addition and not correction</note>
```

Text Encoding Initiative - getting started

<text> & <body>

```
<text>
    <body>
```

- minimally contains <body>
- <body> contains lower level text structures such as
  - *<p>*
  - *or different structures for genres other than prose*

```
<text>
  <body>
    <p>some text...</p>
  </body>
</text>
```

Text Encoding Initiative - getting started

<text> & <front>

```
<text>
    <front>
```

- optional element containing front matter such as
  - *title pages, headers, prefaces, dedications...*
- also consider prologues for drama, forewords and introductions for prose
- except for title, front matter should be encoded using generic standard elements

```
<front>
  <div type="dedication">
    <p>a small dedication...</p>
  </div>
</front>
```

Text Encoding Initiative - getting started

<text> & <back>

```
<text>
    <back>
```

- optional element and grouping
- may contain all back matter for a text
- either numbered or un-numbered divisions with `@type` attribute
  - *appendix*
  - *glossary - <list> of terms and their explanations*
  - *notes*
  - *bibliogr - list of bibliographical citations <listBibl>*
  - *index*
  - *colophon*

```
<back>
  <div type="colophon">
    <p>Printed and bound by.....</p>
  </div>
</back>
```

Text Encoding Initiative - getting started

<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 - getting started

unitary or composite texts

- TEI also allows us to encode composite texts
- group structurally related text in a `<group>` element within `<text>`
- organise texts as a corpus of diverse texts
  - *using `<teiCorpus>` as the parent element for the group*
- `<group>`
  - *group structurally related text*
- `<teiCorpus>`
  - *corpus of diverse texts*

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*