Talk 3: TEI Structure

March 2009

TEI Structure

This talk gives an introduction to the overall structure of a digital edition within the TEI context, and some of the more common phrase-level elements.

Digital edition

- metadata in the header
- transcription in the body, with links to
- images in a `<facsimile>` element

```
<TEI>
 <teiHeader>
<!-- … metadata describing the manuscript -->
<!-- includes a msDesc within the sourceDesc -->
 </teiHeader>
 <facsimile>
<!-- … metadata describing the digital images -->
 </facsimile>
 <text>
<!-- ... transcription of the manuscript -->
 </text>
</TEI>
```

Text groups

```
<text  xml:id="v147">
  <front>
<!-- introductory materials for volume 147 here -->
  </front>
  <group>
    <text  xml:id="v147-issue1">
      <body>
<!-- first issue of volume 147 -->
      </body>
    </text>
    <text  xml:id="v147-issue2">
      <body>
<!-- second issue of volume 147 -->
      </body>
    </text>
<!-- etc...  -->
  </group>
  <back>
<!-- volume index, appendix etc.  -->
  </back>
</text>
```

Corpus structure

```
<teiCorpus>
  <teiHeader>
<!-- shared metadata -->
  </teiHeader>
  <TEI xml:id="text1">
    <teiHeader>
<!-- specific metadata -->
    </teiHeader>
    <text>
<!-- ... -->
    </text>
  </TEI>
  <TEI xml:id="text2">
    <teiHeader>
<!-- specific metadata -->
    </teiHeader>
    <text>
<!-- ... -->
    </text>
  </TEI>
</teiCorpus>
```

Global attributes

Some features (potentially) apply to everything:

- identity
- language
- rendition

TEI provides global attributes for these:

- *@xml:id* provides a unique identifier for any element;
- *@n* provides a name or number for any element
- *@xml:lang* specifies the language of any element, using an ISO standard code
- *@rend* and *@rendition* provide ways of specifying the visual appearance (rendition) of any element

What kinds of metadata?

For any text encoding project, we need a place for such information as:

- identification of the resource itself ("what is this thing?")
- statements of responsibility ("who did what when?")
- indication of source ("where was this derived from?")
- publication statement ("how is this item distributed and by whom?")
- declaration of encoding practice ("what do the codes we added mean?")

The TEI Header supports all these, and more.

The TEI Header

The TEI header was designed with two needs in mind:

- bibliographers and librarians trying to document ‘electronic books’
- text analysts trying to document ‘coding practices’ within digital resources

On the one hand, the Librarian's header:

- uses standard bibliographic concepts
- respects established mappings to other such records (e.g. MARC)
- has a preference for structured data over loose prose

On the other, Everyman's header:

- Supports a huge range of very miscellaneous information, organized in fairly ad hoc ways
- Unpredictable combinations of narrowly encoded documentation systems and loose prose descriptions

TEI Header Structure

The TEI header has four main components:

- `<fileDesc>` (file description) contains a full bibliographic description of an electronic file.
- `<encodingDesc>` (encoding description) documents the relationship between an electronic text and the source or sources from which it was derived.
- `<revisionDesc>` (revision description) summarizes the revision history for a file.
- `<profileDesc>` (text-profile description) provides a detailed description of non-bibliographic aspects of a text, specifically the languages and sublanguages used, the situation in which it was produced, the participants and their setting. (just about everything not covered in the other header elements

Only `<fileDesc>` is required; the others are optional.

Example minimal structure

```
<teiHeader>
 <fileDesc>
  <titleStmt>
   <title>[Title of manuscript]</title>
  </titleStmt>
  <publicationStmt>
   <distributor>[name of data provider]</distributor>
   <idno>[project-specific identifier]</idno>
  </publicationStmt>
  <sourceDesc>
   <msDesc xml:id="ex1" xml:lang="en">
<!-- [full manuscript description ]-->
   </msDesc>
  </sourceDesc>
 </fileDesc>
 <revisionDesc>
  <change when="2008-01-01">[revision
information]</change>
 </revisionDesc>
</teiHeader>
```

Inside <text>

Inside the <text> element there are ways to group together chunks of texts:

- : provides nested division into chapters, sections, sub-sections, etc.
- <p>: marks paragraphs
- <ab>: marks paragraph-like 'anonymous blocks'
- <sp>: marks dramatic speeches
- <lg>: marks groups of verse lines
- <u>: marks transcribed utterances

Phrase-level elements

Within the elements already introduced, TEI offers plenty of scope for mark-up of smaller components. For example:

- boundaries, such as page, column, or line breaks
- highlighting, emphasis and quotation
- editorial changes such as correction, normalization etc.
- names, numbers, dates, addresses...
- links and cross-references
- notes, annotation, indexing
- graphics
- bibliographic citations
- words and other analyses

Highlighting

By highlighting we mean any combination of typographic features which distinguishes the text from its surroundings. This may be for many reasons...

- to mark foreign, archaic, technical usages
- for emphasis when spoken
- to show something is not part of the text.. (e.g. cross references, titles, headings)
- or is attributed to some other agency inside or outside the text (e.g. direct speech, quotation)

TEI provides both a generic `<hi>` tag and a large number of specific ones...

A few highlighting examples

- <hi> (highlighted: reason unknown or unimportant)

```
<p>[The rest of this communication is omitted owing to considerations of
space. —<hi rend="sc">Ed</hi>.]</p>
```

- <emph> (emphasized)

```
<said>'E won't bite yer <emph>if you buy 'im</emph> guv'ner.</said>
```

- <title> and <foreign>:

```
<p>
  <foreign xml:lang="fr">À propos</foreign> of Oxford, it is a question
whether that extremely amusing book <title>Verdant Green</title> is still much
read by freshers.
</p>
```

Quotation

Quotation marks can similarly be used to set off text for many reasons:

- <q> (used if the reason is unknown or unimportant)
- <said> (speech or thought)
- <quote> (attributed to an external source)
- <mentioned> and <soCalled> (nuances of narrative status)

```
<p>
  <said  who="#Celia">I know a lovely tin of potted
    grouse,</said> said Celia, and she went off to cut some sandwiches.
</p>
```

Quotation (continued)

Note that these elements can nest within one another:

```
<p>The poet returned to his work.  <said>
   <quote>In
       tooth and claw,</quote>
  </said> he muttered to himself,
<said>
   <quote>In tooth and claw.</quote>
  </said>
</p>
```

Editorial intervention

As a simple example, consider: ‘Excuse me sir, but would you like to buy a nice little dawg?’.
We can:

- use <orig> to show that "dawg" is what it says, even though this is a nonstandard spelling
- use <reg> to show that "dog" is an editorially-supplied regularisation of what it says
- or provide both within a <choice> element to say either is a valid encoding:

```
...a nice little <choice>
  <orig>dawg</orig>
  <reg>dog</reg>
</choice>?
```

Abbreviation

Abbreviations are highly characteristic of manuscript materials of all kinds. Western MSS traditionally distinguish:

Suspensions the first letter or letters of the word are written, generally followed by a point, or other marker: for example e.g. for exempla gratia

Contractions both first and last letters are written, generally with some other mark of abbreviation such as a superscript stroke, or, less commonly, a point or points: e.g. Mr. for Mister

Brevigraphs Special signs or tittels, such as the Tironian nota used for 'et', the letter p with a barred tail commonly used for per, the letter c with a circumflex used for cum (ĉ) etc.

Superscripts Superscript letters (vowels or consonants) are often used to indicate various kinds of contraction: e.g. w followed by superscript ch for which.

Encoding abbreviations (1)

TEI proposes two levels of encoding:

- the whole of an abbreviated word and the whole of its expansion: <abbr> and <expan>
- abbreviatory signs or characters and the 'invisible' characters they imply: <am> and <ex>

Encoding abbreviations (2)

Depending on editorial policy, we might represent this combination in any one of the following ways:

- ```
  <abbr>D<am/>ni</abbr>
  ```
- ```
  <expan>D<ex>omi</ex>ni</expan>
  ```
- ```
  <choice>
    <abbr>D<am/>ni</abbr>
    <expan>D<ex>omi</ex>ni</expan>
  </choice>
  ```

One can also provide the Unicode character inside the <am> element if available.
  ```

Supplied text

Sometimes, a transcript may need to include words not visibly present in the source:

- because the carrier has been damaged or is barely legible
- because of (assumed) scribal error

The <supplied> element is provided for use in either situations; the *@reason* attribute is used to distinguish them.

```
…Dragging the worst
among<supplied reason="omitted">s</supplied>t us…
```

Names of persons, places, things...

- <name> (a name in the text, contains a proper noun or noun phrase)
- <persName> (a name of a person)
- <placeName> (a name of a place)
- <rs> (a general-purpose name or referencing string)
- <title> (any form of title)

The *@type* attribute is useful for categorizing these, and they both also have *@key*, *@ref*, and *@nymRef* attributes.

Name example

```
<persName ref="#Radz01">
  <forename n="1">Georgio</forename>
  <forename n="2">Nicolai</forename>
  <surname>Radzivil</surname>
  <roleName type="office">capitaneo <placeName type="settlement">
     <choice>
       <abbr>Grodnen<am/>
       </abbr>
       <expan>Grodnen<ex>si</ex>
       </expan>
     </choice>
   </placeName>
  </roleName>
</persName>
```

Dates

- <date> contains a date and time in any format
- For processing it is convenient to add a normalized version, using the *@when* attribute
- Uncertain dates and times can be indicated by other attributes: *@notBefore*, *@notAfter*, or ranges with *@from@to*

```
<date  when="1518-04">anno <expan>D<ex>omi</ex>ni</expan>
  <expan>mille<ex>ssi</ex>mo</expan>
  <expan>quingente<ex>ssi</ex>mo</expan> decimo octauo, menſe Aprilis</date>
```

Cross references

A cross reference is a link from one point in a text (the source) to another (the target).
TEI provides generic elements <ptr> and <ref> for this purpose.
If the linking text can be automatically generated use <ptr>; otherwise use <ref>.
The source is the location of the <ptr> or <ref>; the target is specified by the *@target* attribute, in the form of a URI reference.

```
See <ref  target="#Section12">section 12 on page 34</ref>.
```

```
See <ptr  target="#Section12"/>.
```

Bibliographic Citations

TEI provides special elements for bibliographic citations or references:

- `<bibl>` (loosely structured)
- `<biblStruct>` (standard bibliographic structure)
- `<listBibl>` (encloses a bibliography)

These are typically used in preparing bibliographies, or in footnotes.

Exercise 3

You will be provided with a handout which walks you through embedding your manuscript description from exercise 2 into a full TEI document. This is intended partly to give you more practice with the oXygen editor, and partly to allow you to explore the encoding possibilities of the ENRICH schema. If you have any questions while you are working on it, just raise your hand.