

Talk 5: The <teiHeader>

Dot Porter

July 2009

The TEI header

The TEI header is the bibliographic record for the electronic file, storing information about the file itself, how it is was made, categorisations and analytical information about the text, and revision control information. It is one of the most important aspects of any TEI file.

Why do we need a robust TEI header?

Without proper metadata:

- no one can find your text
- no one knows how you made it
- no one knows why you made it
- no one knows what the text is
- no one knows what they are allowed to do with it

The TEI header provides a structure for documenting this metadata.

The TEI header

The TEI header was designed with two goals in mind

- needs of bibliographers and librarians trying to document 'electronic books'
- needs of text analysts trying to document 'coding practices' within digital resources

The result is that discussion of the header tends to be pulled in two directions...

The Librarian's Header

- Conforms to standard bibliographic model, using similar terminology
- Organized as a single source of information for bibliographic description of a digital resource, with established mappings to other such records (e.g. MARC)
- Emerging code of best practice in its use, endorsed by major digital collections
- Pressure for greater and more exact constraints to improve precision of description: preference for structured data over loose prose

Everyman's Header

- Gives a polite nod to common bibliographic practice, but has a far wider scope
- Supports a (potentially) huge range of very miscellaneous information, organized in fairly ad hoc ways
- Many different codes of practice in different user communities
- 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 Header: Minimal required header

```
<teiHeader>
  <fileDesc>
    <titleStmt>
      <title>A title?</title>
    </titleStmt>
    <publicationStmt>
      <p>Who published?</p>
    </publicationStmt>
    <sourceDesc>
      <p>Where from?</p>
    </sourceDesc>
  </fileDesc>
</teiHeader>
```


Example Header: TEI corpus

```
<teiCorpus>
  <teiHeader type="corpus">
    <!-- corpus-level metadata here -->
  </teiHeader>
  <TEI>
    <teiHeader type="text">
      <!-- metadata specific to this text here -->
    </teiHeader>
    <text>
      <!-- ... -->
    </text>
  </TEI>
  <TEI>
    <teiHeader type="text">
      <!-- metadata specific to this text here -->
    </teiHeader>
    <text>
      <!-- ... -->
    </text>
  </TEI>
</teiCorpus>
```

Types of content in the TEI header

- free prose
 - prose description: series of paragraphs
 - phrase: character data, interspersed with phrase-level elements, but not paragraphs
- grouping elements: specialised elements recording some structured information
- declarations: Elements whose names end with the suffix Decl (e.g. `subjectDecl`, `refsDecl`) enclose information about specific encoding practices applied in the electronic text.
- descriptions: Elements whose names end with the suffix Desc (e.g. `<settingDesc>`, `<projectDesc>`) contain a prose description, possibly, but not necessarily, organised under some specific headings by suggested sub-elements.

File Description

- has some mandatory parts:
 - `<titleStmt>`: provides a title for the resource and any associated statements of responsibility
 - `<sourceDesc>`: documents the sources from which the encoded text derives (if any)
 - `<publicationStmt>`: documents how the encoded text is published or distributed
- and some optional ones:
 - `<editionStmt>`: yes, electronic texts have editions too
 - `<seriesStmt>`: and they also fit into "series".
 - `<extent>`: how many floppy disks, gigabits, files?
 - `<notesStmt>`: notes of various types

NB A "file" may actually correspond with several operating system files.

<fileDesc> components

```
<fileDesc>
  <titleStmt>
  <!-- ... -->
  </titleStmt>
  <editionStmt>
  <!-- ... -->
  </editionStmt>
  <extent>
  <!-- ... -->
  </extent>
  <publicationStmt>
  <!-- ... -->
  </publicationStmt>
  <seriesStmt>
  <!-- ... -->
  </seriesStmt>
  <notesStmt>
  <!-- ... -->
  </notesStmt>
  <sourceDesc>
  <!-- ... -->
  </sourceDesc>
</fileDesc>
```

The File Description

- `<titleStmt>`: contains a mandatory `<title>` which identifies the electronic file (not its source!)
- optionally followed by additional titles, and by 'statements of responsibility', as appropriate, using `<author>`, `<editor>`, `<sponsor>`, `<funder>`, `<principal>` or the generic `<respStmt>`
- `<publicationStmt>`: may contain
 - plain text (e.g. to say the text is unpublished)
 - one or more `<publisher>`, `<distributor>`, `<authority>`, each followed by `<pubPlace>`, `<address>`, `<availability>`, `<idno>`

<titleStmt> example (1)

```
<titleStmt>  
  <title>Two stories by Edgar Allen Poe: electronic version</title>  
  <author>Poe, Edgar Allen (1809-1849)</author>  
  <respStmt>  
    <resp>compiled by</resp>  
    <name>James D. Benson</name>  
  </respStmt>  
</titleStmt>
```

<titleStmt> example (1)

```
<titleStmt>
  <title>Yogadarśanam (arthāt
    yogasūtrapūphah):
    a digital edition.</title>
  <title>The Yogasūtras of Patañjali:
    a digital edition.</title>
  <funder>Wellcome Institute for the History of Medicine</funder>
  <principal>Dominik Wujastyk</principal>
  <respStmt>
    <name>Wieslaw Mical</name>
    <resp>data entry and proof correction</resp>
  </respStmt>
  <respStmt>
    <name>Jan Hajic</name>
    <resp>conversion to TEI-conformant markup</resp>
  </respStmt>
</titleStmt>
```

<publicationStmt> example

```
<publicationStmt>
  <publisher>Sigma Press</publisher>
  <address>
    <addrLine>21 High Street,</addrLine>
    <addrLine>Wilmslow,</addrLine>
    <addrLine>Cheshire M24 3DF</addrLine>
  </address>
  <date>1991</date>
  <distributor>Oxford Text Archive</distributor>
  <idno type="ota">1256</idno>
  <availability>
    <p>Available with prior consent of depositor for
      purposes of academic research and teaching only.</p>
  </availability>
</publicationStmt>
```


<notesStmt> example

<notesStmt> is pretty self-evident and contains notes on the text as a whole:

```
<notesStmt>  
  <note>Historical commentary provided by Mark Cohen.</note>  
  <note>OCR scanning done at University of Toronto.</note>  
</notesStmt>
```

The Source Description

Many electronic texts were not just 'born digital': their source/s need specification in traditional bibliographic style

- `<bibl>`, `<biblStruct>`
- (for texts which were born digital): `<biblFull>` may contain a nested `<fileDesc>`
- `<listBibl>`, `<listPerson>`, `<listPlace>` etc.
- prose description
- more specialised elements are available for spoken texts (`<recordingStmt>` etc.) and for manuscripts (`<msDescription>`)

<sourceDesc> example (1)

```
<sourceDesc>  
  <p>Born digital. </p>  
</sourceDesc>
```

```
<sourceDesc>  
  <bibl>  
    <title level="a">Enigma</title>, <title level="j">Punch: or the London  
Charivari</title>, <date when="1914-07-01">July 1, 1914</date>, 147, p.  
6</bibl>  
</sourceDesc>
```

<sourceDesc> example (2)

```
<sourceDesc>
  <biblStruct xml:lang="fr">
    <monogr>
      <author>Eugène Sue</author>
      <title>Martin, l'enfant trouvé</title>
      <title type="sub">Mémoires d'un valet de chambre</title>
      <imprint>
        <pubPlace>Bruxelles et Leipzig</pubPlace>
        <publisher>C. Muquardt</publisher>
        <date when="1846">1846</date>
      </imprint>
    </monogr>
  </biblStruct>
</sourceDesc>
```

Association between header and text

By default everything asserted by a header is true of the text to which it is prefixed. This can be over-ridden:

- as when a text header over-rides or amplifies a corpus-header setting
- when model.declarable elements are selected by means of the *@decls* attribute (available on all model.declaring elements)
- using special purpose selection/definition elements e.g. `<catRef>` and `<taxonomy>` (see below)

Most components of the encoding description are declarable.

Encoding Description

`<encodingDesc>` groups notes about the procedures used when the text was encoded, either summarised in prose or within specific elements such as

- `<projectDesc>`: goals of the project
- `<samplingDecl>`: sampling principles
- `<editorialDecl>`: editorial principals, e.g.
`<correction>`, `<normalization>`, `<quotation>`,
`<hyphenation>`, `<segmentation>`,
`<interpretation>`
- `<classDecl>`: classification system/s used
- `<tagsDecl>`: specifics about usage of particular elements

The `<encodingDesc>` can replace the user manual, or facilitate semi-automatic document management, given agreed codes of practice.

<rendition> element

- <rendition>: structured information about appearance in the source document

```
<tagsDecl>  
  <rendition xml:id="r-center" scheme="css">text-align: center;</rendition>  
  <rendition xml:id="r-small" scheme="css">font-size: small;</rendition>  
  <rendition xml:id="r-large" scheme="css">font-size: large;</rendition>  
</tagsDecl>
```

<appInfo> element

- <appInfo>: structured information about an application which has edited this TEI file

```
<appInfo>
  <application version="1.7" ident="ImageMarkupTool" notAfter="2008-06-01">
    <label>Image Markup Tool</label>
    <ptr target="#P1"/>
    <ptr target="#P2"/>
  </application>
</appInfo>
```


Profile Description

An extensible rag-bag of descriptions, categorised only as 'non-bibliographic'. Default members of the `model.profileDescPart` class include:

- `<creation>`: information about the origination of the intellectual content of the text, e.g. time and place
- `<langUsage>`: information about languages, registers, writing systems etc used in the text
- `<textDesc>` and `<textClass>`: classifications applied to the text by means of a list of specified criteria or by means of a collection of pointers, respectively
- `<particDesc>` and `<settingDesc>`: information about the 'participants', either real or depicted, in the text
- `<handNotes>`: information about the hands identified in a manuscript

<creation> example

```
<creation>  
  <date when="1992-08">August 1992</date>  
  <rs type="city">Taos, New Mexico</rs>  
</creation>
```

Language and character set usage

The `<langUsage>` element is provided to document usage of languages in the text. Languages are identified by their ISO codes:

```
<langUsage>  
  <language ident="en">English</language>  
  <language ident="bg-cy">Bulgarian in Cyrillic characters </language>  
  <language ident="bg">Romanized Bulgarian</language>  
</langUsage>
```

Classification Methods

`<textClass>` provides a classification (by domain, medium, topic...) for the whole of a text expressed in one or more of the following ways:

- direct reference to a locally defined category (using `<catRef>`)
- reference to an externally defined category (using `<classCode>`)
- documented by `<keywords>`

Example

```
<textClass>
  <catRef target="#X123"/>
  <classCode scheme="DD12">001.9</classCode>
  <keywords scheme="DD">
    <term>End of the World</term>
    <term>Day of Judgment</term>
    <term>Apocalypse</term>
  </keywords>
</textClass>
<classDecl>
  <taxonomy>
    <category xml:id="X1">
      <catDesc>Homiletic writing</catDesc>
      <category xml:id="X123">
        <catDesc>Day of Judgment</catDesc>
      </category>
    </category>
  </taxonomy>
</classDecl>
```

Detailed characterization of a text

`<textDesc>` provides a description of a text in terms of its 'Situational parameters'

```
<textDesc n="novel">  
  <channel mode="w">print; part issues</channel>  
  <constitution type="single"/>  
  <derivation type="original"/>  
  <domain type="art"/>  
  <factuality type="fiction"/>  
  <interaction type="none"/>  
  <preparedness type="prepared"/>  
  <purpose type="entertain" degree="high"/>  
  <purpose type="inform" degree="medium"/>  
</textDesc>  
<!-- These subelements constitute the class model.textDescPart: redefine that  
to roll your own. -->
```

<particDesc> example (1)

```
<particDesc xml:id="p2">  
  <p>Female informant, well-educated, born in Shropshire UK, 12 Jan  
    1950, of unknown occupation. Speaks French fluently.  
    Socio-Economic status B2 in the PEP classification scheme. </p>  
</particDesc>
```

<particDesc> example (2)

```
<particDesc>
  <listPerson>
    <person xml:id="p123">
      <!-- More details on this person -->
    </person>
    <person xml:id="p234">
      <!-- More details on this person -->
    </person>
  </listPerson>
</particDesc>
```


<settingDesc> example (1)

<settingDesc>

<p>The time is early spring, 1989. P1 and P2 are playing on the rug of a suburban home in Bedford. P3 is doing the washing up at the sink. P4 (a radio announcer) is in a broadcasting studio in London. **</p>**

</settingDesc>

<settingDesc> example (2)

```
<settingDesc>
  <setting who="#p1 #p2">
    <name type="city">Bedford</name>
    <name type="region">UK: South East</name>
    <date>early spring, 1989</date>
    <locale>rug of a suburban home</locale>
    <activity>playing</activity>
  </setting>
  <setting who="#p3">
    <name type="city">Bedford</name>
    <name type="region">UK: South East</name>
    <date>early spring, 1989</date>
    <locale>at the sink</locale>
    <activity>washing-up</activity>
  </setting>
<!-- ... -->
</settingDesc>
```

<handNotes> example

```
<handNotes>  
  <handNote xml:id="H1" script="copperplate" medium="brown-ink"> Carefully  
written with regular descenders</handNote>  
  <handNote xml:id="H2" script="print" medium="pencil"> Unschooled  
scrawl</handNote>  
</handNotes>
```

Revision Description

A list of `<change>` elements, each with a *@date* and *@who* attributes, indicating significant stages in the evolution of a document. Most recent first.

<revisionDesc> example

```
<revisionDesc>
  <change when="2006-08-09" who="#LB">handedits following
newhrdgen. xsl</change>
  <change when="2000-10-11" who="#OUCS">Final manual corrections for
BNC- W</change>
  <change when="2000-10-18" who="#OUCS">Further manual corrections for
BNC- W</change>
  <change when="2000-01-08" who="#OUCS">Manually changed catdescriptions etc.
for BNC- W</change>
  <change when="1994-11-30" who="#OUCS">First release for BNC-1</change>
</revisionDesc>
```

<revisionDesc> example

```
<revisionDesc>
  <change>
    <date>$LastChangedDate: 2009-07-03 21:45:33 +0100 (Fri, 3 July 2009)
$. </date>
    <name>$LastChangedBy: dporter $</name>
    <note>$LastChangedRevision: 7892 $</note>
  </change>
</revisionDesc>
```

Bibliographies

The TEI provides numerous ways to provide bibliographic citations, from the highly flexible to the highly structured.

Bibliographic Citations

- `<bibl>` (loosely structured bibliographic citation)
- `<biblStruct>` (structured bibliographic citation)
- `<listBibl>` (a list of bibliographic citations such as a bibliography)
- The 'header' module also includes `<biblFull>` (fully-structured bibliographic citation based on the TEI `fileDesc` element)

Simple <bibl> Example

```
<p>In Punch there is an brief note which could be misconstrued as a slur upon  
Canadians. <note>  
  <bibl>  
    <title level="a">Men for the Antarctic</title>  
    <title level="j">Punch: or the London Charivari</title>,  
    <biblScope>p. 6</biblScope>  
  </bibl>  
</note> It should not be understood as such. </p>
```

Another <bibl> Example

Keble is, of course, named after the hymn-writer and divine; and Balliol, where C. S. C. played the wag so divertingly, after Balliol. `<hi rend="it">À propos</hi>` of Oxford, it is a question whether that extremely amusing book, `<bibl>`
`<title>Verdant Green</title>`
`</bibl>`, is still much read by freshers.

Simple <biblStruct> Example

'Enigma', Punch: or the London Charivari, July 1, 1914, 147, pp. 1-20

```
<biblStruct>
  <analytic>
    <title level="a">Enigma</title>
  </analytic>
  <monogr>
    <title level="j">Punch: or the London Charivari</title>
    <imprint>
      <pubPlace>London</pubPlace>
      <date when="1914-07-01">July 1, 1914</date>
      <biblScope type="vol">147</biblScope>
      <biblScope type="pp">1-20</biblScope>
    </imprint>
  </monogr>
</biblStruct>
```

Another <biblStruct> Example

```
<biblStruct>
  <monogr>
    <title>Magnalia Christi Americana: or, The
      ecclesiastical history of New-England, ...</title>
    <author>Mather, Cotton (1663-1728)</author>
    <imprint>
      <publisher>Printed for Thomas Parkhurst, at the
        Bible and Three Crowns in Cheapside.</publisher>
      <pubPlace>London</pubPlace>
      <date when="1702">MDCCII</date>
    </imprint>
  </monogr>
</biblStruct>
```

<biblFull> example (1)

```
<biblFull>
  <titleStmt>
    <title>The Feminist Companion to Literature in English: women
writers from the middle ages to the present</title>
    <author>Blain, Virginia</author>
    <author>Clements, Patricia</author>
    <author>Grundy, Isobel</author>
  </titleStmt>
  <editionStmt>
    <edition>UK edition</edition>
  </editionStmt>
  <extent>1231 pp</extent>
  <publicationStmt>
    <publisher>Yale University Press</publisher>
    <pubPlace>New Haven and London</pubPlace>
    <date>1990</date>
  </publicationStmt>
  <sourceDesc>
    <p>No source: this is an original work</p>
  </sourceDesc>
</biblFull>
```

<biblFull> example (2)

```
<biblFull>
  <titleStmt>
    <title>Envisioning Information</title>
    <author>Tufte, Edward R[olf]</author>
  </titleStmt>
  <extent>126 pp. </extent>
  <publicationStmt>
    <publisher>Graphics Press</publisher>
    <pubPlace>Cheshire, Conn. USA</pubPlace>
    <date>1990</date>
  </publicationStmt>
</biblFull>
```

Conclusion

Without proper metadata:

- no one can find your text
- no one knows how you made it
- no one knows why you made it
- no one knows what the text is
- no one knows what they are allowed to do with it