TEI Header and Metadata

TEI@Oxford

February 2010

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>
```

The TEI supports two ‘levels’ or types of header

- *corpus level* metadata sets default properties for everything in a corpus
- *text level* metadata sets specific properties for one component text of a 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>
```

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.

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>

A real minimal header

```
<fileDesc>
 <titleStmt>
  <title>Punch, or the London Charivari: an electronic
edition</title>
  <respStmt>
   <resp>TEI version</resp>
   <name>TEI@Oxford team</name>
  </respStmt>
 </titleStmt>
 <publicationStmt>
  <p>Unpublished</p>
 </publicationStmt>
 <sourceDesc>
  <p>Recoded from the Project Gutenberg versions
      Recoded from the Project Gutenberg versions</p>
 </sourceDesc>
</fileDesc>
```

Title- and Responsibility- statements...

There may be many of them:

```
<title>Artamene</title>
<title type="alt">Le Grand Cyrus</title>
<title type="sub">Digital Edition</title>
```

Amongst the guilty parties:

```
<author>Scudery, Madeleine de</author>
<principal>Geffin, Alexandre</principal>
<funder>Fonds Nationale Suisse de la Recherche Scientifique
Fonds Nationale Suisse de la Recherche Scientifique</funder>
<respStmt>
 <resp>Encoding check</resp>
 <name>Jean Untel</name>
</respStmt>
```

<titleStmt> example (1)

```
<titleStmt>
 <title>Yogadarśanam (arthāt yogasūtrapūphah): a digital edition.
    Yogadarśanam (arthāt yogasūtrapūphah): a digital edition.
    Yogadarśanam (arthāt yogasūtrapūphah): a digital
edition.</title>
 <title>The Yogasūtras of Patañjali: a digital edition.
    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>TEI Consortium</publisher>
 <distributor>Oxford Text Archive</distributor>
 <idno type="ota">1256</idno>
 <availability>
   <p>Available under the terms of a Creative Commons Attribution
and Share Alike licence.
      Available under the terms of a Creative Commons Attribution
and Share Alike licence.</p>
 </availability>
</publicationStmt>
```

<notesStmt> example

<notesStmt> can contain notes on almost any aspect:

```
<notesStmt>
 <note>Material prepared for the TEI@Oxford Summer School.</note>
</notesStmt>
```

The Source Description

Few electronic texts are original 'born digital' works: their source/s therefore need to be described using traditional bibliographic practice

- prose description
- <bibl> : contains free text or any pixture of bibliographic elements such as <author>, <publisher> etc.
- <biblStruct> contains effectively the same elements but constrained in various ways according to bibliographic standards
- <biblFull> special-cases texts which were born TEI by replicating an embedded <fileDesc>
- A <listBibl> may be used for lists of such descriptions
- Specialised elements are available for spoken texts (<recordingStmt> etc.) and for manuscripts or other text-bearing objects (<msDesc>)
- Authority lists for e.g people (<listPerson>) or places (<listPlace>) can be included.

<sourceDesc> examples (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>
```

<bibl> vs. <biblStruct> Example

```
<bibl>
 <title level="a">Enigma</title>, in <title level="j">Punch: or
the London Charivari
    Punch: or the London Charivari</title> (July 1, 1914), vol 147,
pp. 1-20
</bibl>
```

```
<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>
```

<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.

Sample encoding description 1

```
<encodingDesc>
 <projectDesc>
  <p>The Imaginary Punch Project aims to ....
  </p>
 </projectDesc>
 <samplingDecl>
  <p>All pages containing editorial text have been transcribed in
full. Pages containing only advertisements or illustrations have
been omitted.
      All pages containing editorial text have been transcribed in
full. Pages containing only advertisements or illustrations have
been omitted.
      All pages containing editorial text have been transcribed in
full. Pages containing only advertisements or illustrations have
been omitted.</p>
 </samplingDecl>
 <editorialDecl>
  <hyphenation>
   <p>Original spelling has been retained, except that words
hyphenated across line breaks have been silently re-assembled. The
hyphen has been retained only where there exist cases of the same
word being hyphenated in mid-line position.
         Original spelling has been retained, except that words
hyphenated across line breaks have been silently re-assembled. The
hyphen has been retained only where there exist cases of the same
word being hyphenated in mid-line position.
```

Sample encoding description 2

```
<encodingDesc>
<!-- ... -->
 <classDecl>
  <taxonomy xml:id="size">
   <category xml:id="large">
    <catDesc>story occupies more than half a page</catDesc>
   </category>
   <category xml:id="medium">
    <catDesc>story occupies between quarter and a half page
         story occupies between quarter and a half page</catDesc>
   </category>
   <category xml:id="small">
    <catDesc>story occupies less than a quarter page
         story occupies less than a quarter page</catDesc>
   </category>
<!-- etc -->
  </taxonomy>
  <taxonomy xml:id="topic">
   <category xml:id="politics-domestic">
    <catDesc>Refers to domestic political events
         Refers to domestic political events</catDesc>
   </category>
   <category xml:id="politics-foreign">
    <catDesc>Refers to foreign political events</catDesc>
   </category>
   <category xml:id="social-women">
```

Sample encoding description 3

```
<encodingDesc>
<!-- ... -->
 <tagsDecl>
  <namespace name="http://www.tei-c.org/ns/1.0">
   <tagUsage gi="cit" occurs="410"/>
   <tagUsage gi="div" occurs="115"/>
   <tagUsage gi="gap" occurs="3"/>
   <tagUsage gi="head" occurs="156"/>
   <tagUsage gi="hi" occurs="147"/>
   <tagUsage gi="l" occurs="2"/>
   <tagUsage gi="lg" occurs="1"/>
   <tagUsage gi="p" occurs="680"/>
   <tagUsage gi="quote" occurs="3"/>
   <tagUsage gi="s" occurs="2415"/>
   <tagUsage gi="w" occurs="41799"/>
  </namespace>
  <namespace name="http://www.ipp.org/ns/1.0">
   <tagUsage gi="citCom" occurs="417"/>
  </namespace>
 </tagsDecl>
</encodingDesc>
```

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="fr">French</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:

using <catRef> direct reference to a locally defined (e.g. in the corpus header) category

using <classCode> reference to some commonly agreed and externally defined category (e.g. UDC)

using <keywords> assign arbitrary descriptive terms taken from a bibliographic controlled vocabulary or a tag cloud

BNC Example

```
<profileDesc>
 <creation>
  <date when="1962"/>
 </creation>
 <textClass>
  <catRef
     target="#WRI #ALLTIM1 #ALLAVA2 #ALLTYP3 #WRIDOM5 #WRILEV2
#WRIMED1 #WRIPP5 #WRISAM3 #WRISTA2 #WRITAS0#WRI #ALLTIM1 #ALLAVA2
#ALLTYP3 #WRIDOM5 #WRILEV2 #WRIMED1 #WRIPP5 #WRISAM3 #WRISTA2
#WRITAS0#WRI #ALLTIM1 #ALLAVA2 #ALLTYP3 #WRIDOM5 #WRILEV2 #WRIMED1
#WRIPP5 #WRISAM3 #WRISTA2 #WRITAS0#WRI #ALLTIM1 #ALLAVA2 #ALLTYP3
#WRIDOM5 #WRILEV2 #WRIMED1 #WRIPP5 #WRISAM3 #WRISTA2 #WRITAS0#WRI
#ALLTIM1 #ALLAVA2 #ALLTYP3 #WRIDOM5 #WRILEV2 #WRIMED1 #WRIPP5
#WRISAM3 #WRISTA2 #WRITAS0#WRI #ALLTIM1 #ALLAVA2 #ALLTYP3 #WRIDOM5
#WRILEV2 #WRIMED1 #WRIPP5 #WRISAM3 #WRISTA2 #WRITAS0#WRI #ALLTIM1
#ALLAVA2 #ALLTYP3 #WRIDOM5 #WRILEV2 #WRIMED1 #WRIPP5 #WRISAM3
#WRISTA2 #WRITAS0#WRI #ALLTIM1 #ALLAVA2 #ALLTYP3 #WRIDOM5 #WRILEV2
#WRIMED1 #WRIPP5 #WRISAM3 #WRISTA2 #WRITAS0 #WRI #ALLTIM1 #ALLAVA2
#ALLTYP3 #WRIDOM5 #WRILEV2 #WRIMED1 #WRIPP5 #WRISAM3 #WRISTA2
#WRITAS0#WRI #ALLTIM1 #ALLAVA2 #ALLTYP3 #WRIDOM5 #WRILEV2 #WRIMED1
#WRIPP5 #WRISAM3 #WRISTA2 #WRITAS0#WRI #ALLTIM1 #ALLAVA2 #ALLTYP3
#WRIDOM5 #WRILEV2 #WRIMED1 #WRIPP5 #WRISAM3 #WRISTA2 #WRITAS0"/>
  <classCode scheme="DLEE">W nonAc: humanities arts</classCode>
  <keywords scheme="COPAC">
   <term>History, Modern - 19th century</term>
   <term>Capitalism - History - 19th century</term>
```

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.
   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.
   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="HanBISM">
      <persName>
        <forename>Hannah</forename>
        <forename>Leopoldine</forename>
        <forename>Alice</forename>
        <surname>von Bismarck-Schönhausen</surname>
      </persName>
      <birth  when="1893-05-11">1893</birth>
      <death>1971</death>
    </person>
    <person  xml:id="JLOW">
      <persName>James William Lowther</persName>
      <persName  type="title"  from="1921-07-08">Viscount Ullswater</persName>
      <occupation  from="1905-06-08"  to="1921-04-28">Speaker of the House of
Commons</occupation>
      <birth  when="1855-04-01">1855</birth>
      <death  when="1949-03-27">1947</death>
      <note>
        <ref
            target="http://www.oxforddnb.com/view/article/34615?docPos=2">DNB entry</ref>
      </note>
      <note>
        <ref
            target="http://hansard.millbanksystems.com/people/mr-james-lowther">Hansard
entry</ref>
      </note>
    </person>
  </listPerson>
</particDesc>
```

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.
- Can be maintained manually, but better done by means of a CMS (change management system)

```
<revisionDesc>
 <change>
  <date>$LastChangedDate: 2010-02-06 19:38:40 +0000 (Sat, 06 Feb
2010) $.</date>
  <name>$LastChangedBy: rahtz $</name>
  <note>$LastChangedRevision: 9288 $</note>
 </change>
</revisionDesc>
```

Some more Acronym soup

Some significant metadata related acronyms:

DCMI: Dublin Core Metadata Initiative Very simple standard for describing web resources: 15 ‘lowest common denominator’ fields

RDF: Resource Description Framework W3C Standard for representing any kind of resource description using object oriented concepts: basis of the ‘semantic web’

OAIS: Open Archival Information System well developed abstract model for any archival system: ISO standard

EAD: Encoded Archival Description International Standard for describing archival collections

METS: Metadata Encoding and Transcription Standard generalised method of integrated different metadata systems

TEI provides a richer vocabulary than EAD or DCMI, and is less abstract than RDF or METS

The future

- The TEI header was originally conceived as something for non-specialist usage
- It lacks "application profiles" for particular uses
- Standard codes of practice or ways of using have been developed by particular user communities (e.g. digital librarians, corpus linguists)
- As a ‘primary source of information’ it remains an essential framework for documenting:
  - what your text is
  - where it came from
  - how you encoded it
  - how it may be used (technically)
  - how it may be used (legally)