Metadata: The TEI Header and Manuscript Description

TEI@Oxford

2010-07

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.
- <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
- <revisionDesc> (revision description) summarizes the revision history for a file.

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

Corpus Header Example

```
<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

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 minimal header for Punch

```
<fileDesc>
  <titleStmt>
    <title>Punch, or the London Charivari: an electronic
        edition</title>
    <editor>Owen Seaman (1861-1936)</editor>
    <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</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</funder>
<respStmt>
  <resp>Encoding check</resp>
  <name>Jean Untel</name>
</respStmt>
```

<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.</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

All electronic works need to indicate their source, even if it is just to say that it is 'born digital'. There are variety of ways to do this:

- prose description
- <bibl> : contains free text or any mixture 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 for spoken texts (<recordingStmt> etc.) and for manuscripts (<msDesc>) Discussed later!
- Authority lists for e.g people (<listPerson>) or places (<listPlace>) can be included.

<sourceDesc> examples

```
<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</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>
```

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.

<encodingDesc> Example (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.</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. </p>
    </hyphenation>
<!-- ... -->
  </editorialDecl>
<!-- ... -->
</encodingDesc>
```

<encodingDesc> Example (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</catDesc>
   </category>
   <category xml:id="small">
    <catDesc>story occupies less than a quarter page</catDesc>
   </category>
  </taxonomy>
  <taxonomy xml:id="topic">
   <category xml:id="politics-domestic">
    <catDesc>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">
    <catDesc>refers to role of women in society</catDesc>
   </category>
   <category xml:id="social-servants">
    <catDesc>refers to role of servants in society</catDesc>
```

Profile Description

A collection 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

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"/>
  <classCode scheme="DLEE">W nonAc: humanities arts</classCode>
  <keywords scheme="COPAC">
   <term>History, Modern - 19th century</term>
   <term>Capitalism - History - 19th century</term>
   <term>World, 1848-1875</term>
  </keywords>
 </textClass>
</profileDesc>
```

This categorization applies to the whole text. For more fine grained classification, use *@decls* on e.g. a element.

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: 2010-06-28 09:14:36 +0100 (Mon, 28 Jun
        2010) $.</date>
    <name>$LastChangedBy: lou $</name>
    <note>$LastChangedRevision: 10346 $</note>
  </change>
</revisionDesc>
```

Manuscript Description

Why are manuscripts special?

- Manuscripts are *unique objects*, often of great cultural or political value.
- Books, by contrast, exist in multiple copies, and can be described adequately by well-established and formalised bibliographic conventions.
- For manuscripts, there are several traditions, often descriptive or belle lettriste, and little consensus.

Similar concerns apply to other text-bearing objects.

Objectives of <msDesc>

The TEI <msDesc> element is intended for several different kinds of applications:

- standalone database of library records (*finding aid*)
- discursive text collecting many records (*catalogue raisonné*)
- metadata component within a digital surrogate (*electronic edition*)
- tool for 'quantitative codicology'

Catalogue Raisonné

An <msDesc> can appear anywhere a <p> paragraph can

```
<div>
  <head>The Arnamagnæan Manuscript Collection</head>
  <p>The Arnamagnæan Collection is widely recognised as one of the
     most significant collections of early Scandinavian manuscripts in
     the world…</p>
  <p>Among its more important holdings are:
<msDesc xml:id="AM02-0101" xml:lang="en">
<!-- …-->
   </msDesc>
  </p>
  <p>In the following manuscript….
<msDesc xml:id="AM04-0595" xml:lang="en">
<!-- …-->
   </msDesc>
  </p>
</div>
```

Having one's cake and eating it

Two conflicting desires:

- preserve (or perpetuate) existing descriptive prose
- reliable search, retrieval, and analysis of data

The <msDesc> tries, wherever possible, to do both of these things.

Components of a manuscript description

Within the <msDesc> element come a required <msIdentifier> element, which groups information identifying the manuscript, followed by an optional <head>, which can be used to provide in a brief, unstructured way information on the manuscript's contents etc. These are then followed either by one or more paragraphs (<p>), or one or more of the following specialised elements:

- <msContents>: an itemised list of the intellectual content of the manuscript, with transcriptions of rubrics, incipits, explicits etc, as well as primary bibliographic references
- <physDesc>: groups information concerning all physical aspects of the manuscript, its material, size, format, script, decoration, binding, marginalia etc.
- <history>: provides information on the history of the manuscript, its origin, provenance and acquisition by its holding institution

Components of a manuscript description (cont.)

- <additional>: groups other information about the manuscript, in particular, administrative information relating to its availability, custodial history, surrogates etc.
- <msPart>: contains in essence a nested <msDesc>, in cases of composite manuscripts now regarded as constituting a single unit but made up of two or more parts which were originally physically distinct.

Within each of these elements a number of sub-elements is available; <msContents>, for example, will normally consist of one or more <msItem> elements, each in turn containing specific elements for <rubric>, <incipit>, <explicit> and <colophon>, as well as the standard TEI elements <author>, <title> and <bibl> for bibliographic references. As with <msDescription> itself, however, the contents of these first-level and second-level elements need not be this structured, since there is also the option of using paragraphs.

Identification (1)

The <msIdentifier>

Traditional three part specification:

- place (<country>, <region>, <settlement>)
- repository (<institution>, <repository>)
- identifier (<collection>, <idno>)

```
<msIdentifier>
  <country>Canada</country>
  <settlement>Ottawa</settlement>
  <repository>Library and Archives Canada</repository>
  <collection>E.W.B. Morrison</collection>
  <idno>MG 30 E 81 v. 16</idno>
</msIdentifier>
```

Identification (2)

Alternative or additional names can also be included:

```
<msIdentifier>
  <country>Danmark</country>
  <settlement>København</settlement>
  <repository> Det ArnamagnæanskeInstitut </repository>
  <idno>AM 45 fol.</idno>
  <msName  xml:lang="la">Codex Frisianus</msName>
  <msName  xml:lang="is">Fríssbók</msName>
</msIdentifier>
```

Intellectual Content

- May simply use paragraphs of text…
- … or a tree of <msItem> elements
- … optionally preceded by a prose summary

We can describe the content in general terms:

```
<msContents>
  <p>An extraordinary charivari of heroic deeds and improving tales,
     including an early version of <title>Guy of Warwick</title> and
     several hymns.</p>
</msContents>
```

or we can provide detail about each distinct item:

```
<msContents>
  <summary>An extraordinary charivari of heroic deeds, improving
     tales, and hymns.</summary>
  <msItem>
<!-- details of Guy of Warwick here -->
  </msItem>
  <msItem>
<!-- other items here -->
  </msItem>
</msContents>
```

The <msItem> element

Manuscripts contain identifiable items, usually physically tied to a *locus*.

- <locus>, if present, must be given first
- then any of the following, in a specified order:
  - <author>, <respStmt>
  - <title>, <rubric>, <incipit>, <explicit>, <colophon>, <finalRubric>
  - <quote>, <textLang>, <decoNote>, <bibl>, <listBibl>, <note> …
  - … or nested <msItem>s

<msContents> with multiple <msItem>s

```
<msContents>
  <msItem n="1">
    <locus from="5r" to="7v">fols. 5r-7v</locus>
    <title type="supplied">An ABC</title>
  </msItem>
  <msItem n="2">
    <locus from="7v" to="8v">fols. 7v-8v</locus>
    <title type="uniform" xml:lang="fr">Lenvoy de Chaucer a
        Scogan</title>
  </msItem>
<!-- …further items here… -->
  <msItem n="6">
    <locus from="14r" to="126v">fols. 14r-126v</locus>
    <title type="uniform">Troilus and Criseyde</title>
    <note>Bk. 1:71-Bk. 5:1701, with additional losses due to
        mutilation throughout</note>
  </msItem>
</msContents>
```

Physical Description

An artificial (but helpful) grouping of many distinct items.

You can simply supply paragraphs of prose, covering such topics as

- <objectDesc>: the physical carrier
- <handDesc>: what is carried on it
- <musicNotation>, <decoDesc>, <additions>
- <bindingDesc> and <sealDesc>
- <accMat>: accompanying material

Or, group your discussion within the specific elements mentioned above.

Similarly, within the specific elements, you can supply paragraphs of prose, or further specific elements.

The carrier 1

The <objectDesc> can contain just paragraphs, or <supportDesc> and <layoutDesc>

```
<objectDesc form="codex">
  <supportDesc material="mixed">
    <p>Early modern <material>parchment</material> and
    <material>paper</material>.</p>
  </supportDesc>
  <layoutDesc>
    <layout columns="1" ruledLines="25 32"/>
  </layoutDesc>
</objectDesc>
```

The carrier 2

A more complex substructure with specific elements for <support>, <extent>, <foliation>, <collation>, <condition>.
Multiple layouts may also be specified:

```
<layoutDesc>
  <layout ruledLines="25" columns="1">
    <p>
      <locus from="1r" to="202v"/>
      <locus from="210r" to="212v"/> Between 25 and 32 ruled
        lines.</p>
  </layout>
  <layout ruledLines="34 50" columns="1">
    <p>
      <locus from="203r" to="209v"/>Between 34 and 50 ruled
        lines.</p>
  </layout>
</layoutDesc>
```

<handDesc> and <decoDesc>

- <handNote> (note on hand) describes a particular style or hand distinguished within a manuscript.
- <decoNote> contains a note describing either a decorative component of a manuscript or a fairly homogenous class of such components.

<handDesc> example (1)

```
<handDesc hands="2">
  <p>The manuscript is written in two contemporary hands, otherwise
    unknown, but clearly those of practised scribes. Hand I writes
    ff.1r-22v and hand II ff. 23 and 24. Some scholars, notably
    Verner Dahlerup and Hreinn Benediktsson, have argued for a third
    hand on f. 24, but the evidence for this is insubstantial.</p>
</handDesc>
```

<handDesc> example (2)

```
<handDesc hands="2">
  <handNote xml:id="Eirsp-1" scope="minor" script="other">
    <p>The first part of the manuscript, <locus from="1v" to="72v:4">fols
1v-72v:4</locus>, is written in a practised Icelandic
       Gothic bookhand. This hand is not found elsewhere.</p>
  </handNote>
  <handNote xml:id="Eirsp-2" scope="major" script="other">
    <p>The second part of the manuscript,
<locus from="72v:4" to="194v">fols 72v:4-194</locus>, is written in a hand
       contemporary with the first; it can also be found in a
       fragment of <title>Knýtlinga saga</title>, <ref>AM 20b II
          fol.</ref>.</p>
  </handNote>
</handDesc>
```

<additions>

The <additions> element can be used to list or describe any additions to the manuscript, such as marginalia, scribblings, doodles, etc., which are considered to be of interest or importance.

```
<additions>
  <p>The text of this manuscript is not interpolated with sentences
     from Royal decrees promulgated in 1294, 1305 and 1314. In the
     margins, however, another somewhat later scribe has added the
     relevant paragraphs of these decrees, see pp. 8, 24, 44, 47
     etc.</p>
  <p>As a humorous gesture the scribe in one opening of the
     manuscript, pp. 36 and 37, has prolonged the lower stems of one
     letter f and five letters þ and has them drizzle down the
     margin.</p>
</additions>
```

<accMat>

<accMat> (accompanying material) contains details of any significant additional material which may be closely associated with the manuscript being described, such as non-contemporaneous documents or fragments bound in with the manuscript at some earlier historical period.

```
<accMat> A copy of a tax form from 1947 is included in the envelope
with the letter. It is not catalogued separately. </accMat>
```

<history>

- <origin>: where it all began
- <provenance>: everything in between
- <acquisition>: how you acquired it

<origin> is *datable* element and thus has attributes *@notBefore* and *@notAfter*, *@when* etc.

<history> Example

```
<history>
  <origin>
    <p>Written in <origPlace>England</origPlace> in the
<origDate notAfter="1300" notBefore="1200">13th cent. </origDate>
    </p>
  </origin>
  <provenance>
    <p>On fol. 54v very faint is <q>Iste liber est fratris guillelmi
          de buria de <gap reason="illegible"/> Roberti ordinis
          fratrum Pred<ex>icatorum</ex>
     </q>, 14th cent. (?):
    <q>hanauilla</q> is written at the foot of the page (15th
        cent.).</p>
  </provenance>
  <acquisition>
    <p>Bought from the Rev. <name type="person">W. D. Macray</name>
        on <date when="1863-03-17">March 17, 1863</date>, for 1 pound
        10s.</p>
  </acquisition>
</history>
```

<additional> information

- <adminInfo> : administrative information
- <surrogates> : information about other surrogates, i.e. photographs, digital images etc.
- <accMat> : accompanying material
- <listBibl> : bibliography

Administrative information

- record history
- availability
- custodial history
- miscellaneous remarks

```
<adminInfo>
  <custodialHist>
    <custEvent  type="conservation"  notBefore="1961-03"  notAfter="1963-02">
      <p>Conserved between March 1961 and February 1963 at Birgitte
            Dalls Konserveringsværksted.</p>
    </custEvent>
    <custEvent  type="photography"  notBefore="1988-05-01"  notAfter="1988-05-30">
      <p>Photographed in May 1988 by AMI/FA.</p>
    </custEvent>
  </custodialHist>
</adminInfo>
```

And finally

A <msDesc> can contain <msPart>, essentially a nested <msDesc>, where originally distinct manuscripts or parts of a manuscripts have been brought together to form a composite manuscript.

```
<msDesc>
  <msIdentifier>
    <settlement>Amiens</settlement>
    <repository>Bibliothèque Municipale</repository>
    <idno>MS 3</idno>
    <msName>Maurdramnus Bible</msName>
  </msIdentifier>
<!-- other elements here -->
  <msPart>
    <altIdentifier>
      <idno>MS 6</idno>
    </altIdentifier>
<!-- other information specific to this part here -->
  </msPart>
<!-- other msParts here -->
</msDesc>
```

Conclusions

- The TEI header was originally conceived as something for non-specialist usage but has everything needed for rigorous bibliographic description
- It provides detailed methods for encoding specialist items such as manuscript descriptions or details concerning spoken texts or linguist corpora
- 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)