

Metadata Matters

Using TEI Headers to Document a Text

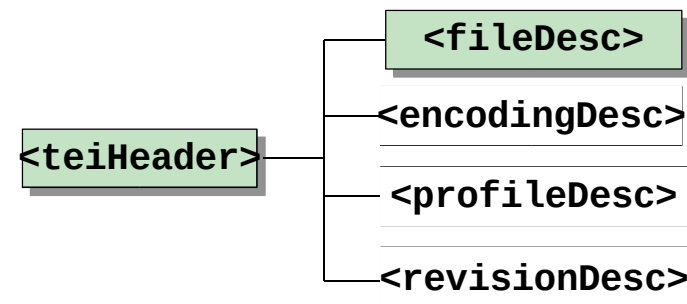
Metadata requirements: the scope

- Identification of the object
- Documentation of its structure and organization
- Statement of rights (reproduction, ownership etc.)
- Statement of intended usage
- Documentation of interpretive scheme/s applied
- Brief characterization for search engines

The TEI header

- Based on AACR2 practice, the header contains:
 - *mandatory* file description
 - *optional* encoding, profile and revision descriptions
- The TEI envisages only two “levels” of header
 - The corpus header
 - The text header

TEI Header structure



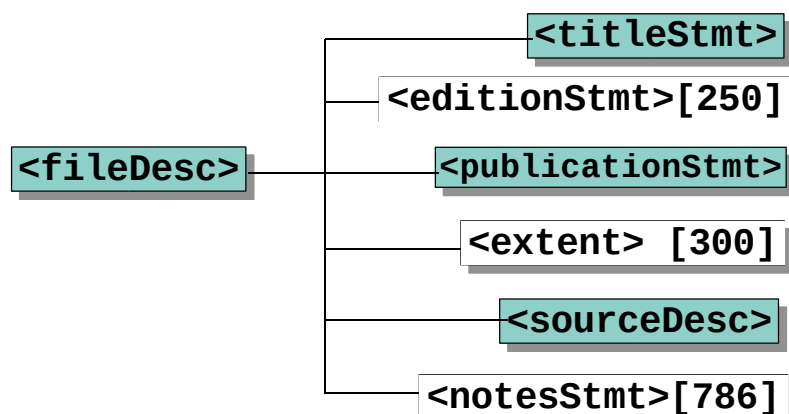
For example

```
<teiHeader><fileDesc>
<titleStmt><title>Thomas Paine: Common sense,
a machine-readable transcript</title>
<respStmt><resp>compiled by</resp>
<name>Jon K
Adams</name></respStmt></titleStmt>
<publicationStmt>
<istributor>Oxford Text
Archive</istributor> </publicationStmt>
<sourceDesc><bibl>The complete writings of
Thomas Paine,collected and edited by Phillip
S. Foner (New York, Citadel Press, 1945)
</bibl></sourceDesc>
</fileDesc><teiHeader>
```

The file description

- Mandatory
- Supplies full description of the electronic file itself, and its source/s
- Must specify at least a title, a publication statement, and a source
- Use of authority control is advisable but not required

The File Description



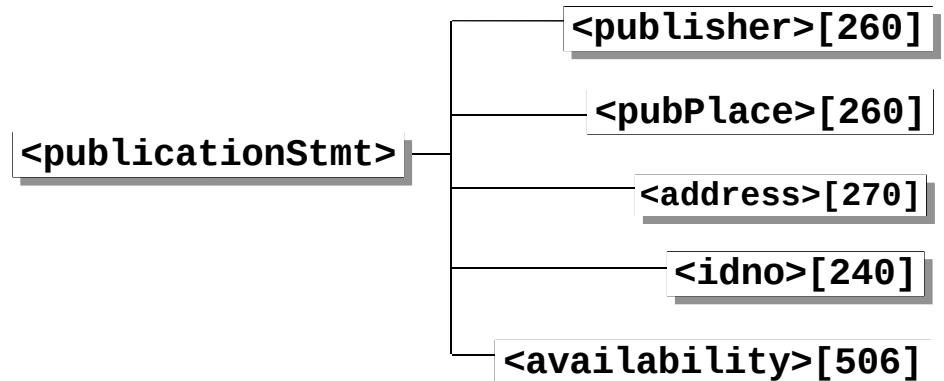
The title statement

- mandatory <title> [245]
- identifies the electronic file, not its source
- optionally followed by statements of responsibility, as appropriate, using <author>, <editor>, <respStmt> [720], <sponsor>, <funder>, <principal> [536]

A real title statement

```
<titleStmt><title>Ayala's Angel</title>
<title type="sub">a machine readable
transcription</title>
<author>Trollope, Anthony (1815-1882)
</author>
<principal>Skilton, David </principal>
</titleStmt>
```

The publication statement



Edition and extent statements

- As elsewhere, distinguishing a new “edition” from just another version may be difficult...
- Extent should be expressed in some platform independent way, e.g. words, Kb.

```
<editionStmt>
<edition>Public Domain TEI edition prepared
at the Oxford Text Archive</edition>
<respStmt><resp>enc</resp>
<name>Price-Wilkin, John</name></respStmt>
</editionStmt>
<extent>1.2 Kb 8 bit bytes</extent>
```

Sample publication statement

```
<publicationStmt>
<istributor>Oxford Text Archive</istributor>
<pubPlace>Oxford
<address><addrLine>OUCS,</addrLine>
<addrLine>13 Banbury Road</addrLine>
<addrLine>Oxford OX2 6NN</addrLine></address></pubPlace>
<idno type="OTA">1873</idno>
<idno type="HTI">TroAyala</idno>
<idno type="URL">
ftp://ota.ox.ac.uk/pub/ota/English/Trollope/ayala.1873</
idno>
<availability status="FREE"><p>Freely available for non-
commercial use provided that this header is included in
its entirety with any copy distributed
</p></availability>
<date>1996-6-4</date></publicationStmt>
```

The source description

- may contain common TEI bibliographic elements <bibl>, <biblStruct>,
 - or a nested file description <biblFull>
 - or a list <listBibl>
 - or a prose description
- specialised elements for transcribed speech
- Or (for the “ born-digital” document) simply the text “ Original”

Sample source description

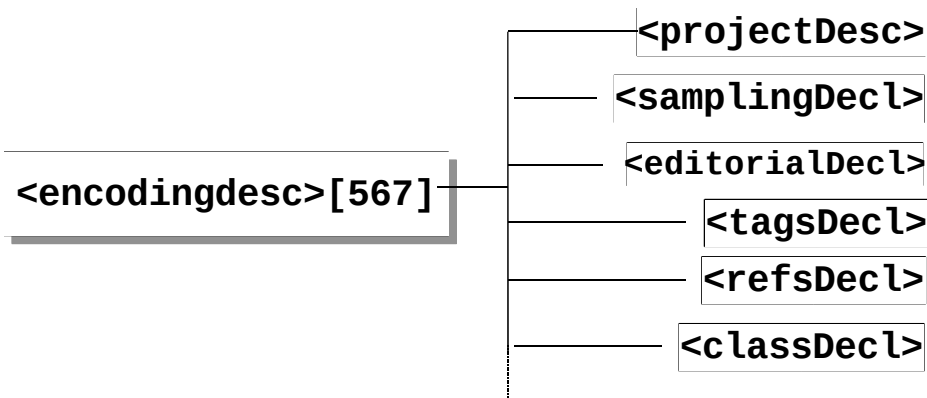
```
<sourceDesc><biblFull><titleStmt><title>Ayala's
Angel</title>
<author>Trollope, Anthony, 1815-1882</author>
<respStmt><resp>edt</resp><name>Skilton, David</name>
</respStmt>
<respStmt><resp>au</resp><name> Ellis, Alice
Thomas</name></respStmt>
<respStmt><resp>ill</resp><name> Geary,
Robert</name></respStmt></titleStmt>
<editionStmt><edition>The Folio Society Edition of the
novels of Anthony Trollope,
1989</edition></editionStmt>
<extent>xvi, 532 p. ; 23 cm</extent>
<publicationStmt><publisher>The Folio
Society</publisher>
<pubPlace>London</pubPlace><date>1989</date>
<idno type="ISBN">1870587049</idno>
</publicationStmt></biblFull>
</sourceDesc>
```

Typical BNC source statements

```
<biblStruct> <monogr> <title>It might have been
Jerusalem. </title> <author n="HealyT1"
domicile="Scotland">Healy, Thomas</author> <imprint
n="POLYG01"> <publisher>Polygon Books</publisher>
<pubPlace>Edinburgh</pubPlace> <date
value="1991">1991</date> </imprint> <biblScope
type="pp">1-81</biblScope> </monogr> </biblStruct>
```

```
<recordingStmt>
<recording n="018201" dur="322" date="1991-11-28"
time="18:15+" type="Walkman" id="KB7RE000"></recording>
<recording n="018202" dur="253" date="1991-11-28"
time="18:15+" type="Walkman" id="KB7RE001"></recording>
<!-- ... --> <recording n="018207" dur="630" date="1991-11-29"
time="10:15+" type="Walkman" id="KB7RE006"></recording>
<recording n="018301" dur="75" date="1991-11-29" time="12:15+"
type="Walkman" id="KB7RE007"> <!-- ... --> </recording>
```

The Encoding Description



documents what was done in creating the electronic form of a text

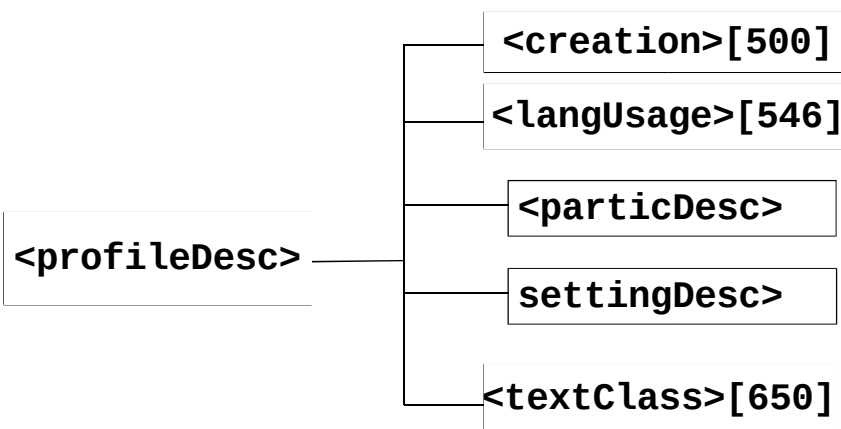
Editorial Declarations

- document policies for some key decisions:
 - <correction>, <normalization>, <quotation>, <hyphenation>, <segmentation>, <interpretation>
- declarable (can vary)
- attributes are suggested for some key aspects; can just contain prose description

Example encoding description

```
<encodingDesc><projectDesc>
<p>Archival storage of uncorrected type-setting
tapes used by the Folio
Society</p></projectDesc>
<editorialDecl><p>There are several typographic
errors in the original which have been left
unmarked</p></editorialDecl>
<tagsDecl>
<tagusage gi="p" occurs="4123"></tagusage>
<!-- etc -->
</tagsDecl>
<refsDecl><p>Each chapter is tagged as a div1 ;
the N attribute holds sequence numbers.
</p></refsDecl></encodingDesc>
```

The Profile Description



Sample profile description

```
<profileDesc
<creation><date>1846</date></creation>
<langUsage>
<language id='ENG'>English</language>
</langUsage>
<textclass><keywords><term>Novel</term>
<term>English Literature -- 19th
century</term></keywords>
<catRef
source='UDC'>822.91</catRef></textclass>
</profileDesc>
```

Language

- global lang attribute
- references a <language> element
- defines both natural language and writing system

```
<langUsage>
<language id='ENG'>English</language>
<language id='BGO'>Bulgarian in Cyrillic characters
</language>
<language id='BUL'>Romanized Bulgarian</language>
</langUsage>
```

BNC Participant Description

```
<person age="2" dialect="XLC" id="PS0M6"
role="self" sex="f" soc="C2">
<name>Pamela</name><age>32</age>
<occupation>sales assistant</occupation>
<dialect>Lancashire</dialect></person>
```

```
<person age="X" educ="X" id="PS2AY"
n="W0003" sex="m" soc="UU">
<name>Frank Harasikwa</name>
<occupation>politician</occupation>
<para>Euro candidate presenting self for
selection</para></person>
```

Setting Descriptions

```
<setting id="KDFSE002" n="063505"
who='PS0M6'>
<name>Lancashire: Morecambe </name>
<locale>at home</locale>
<activity spont="H"> watching television
</activity></setting>
```

```
<u who=PS0M8 ><unclear ></u>
<u who=PS0M6><s n="311"><w VVB>Show <w DPS>your
<w NN1>daddy<c PUN>.</u>
<u who=PS0M8><s n="312"><w NN1>Daddy<c PUN>.</u>
```

Text classification

```
<classCode scheme="DDC">821.32</classCode>
<classCode scheme="dc.Subject">Hot air
ballooning</classCode>

<catref scheme="XYZ" targets='abc1 def'>

<keywords source="LCSH">
<term>huntin;</term>
<term>shootin;</term>
<term>fishin</term>
</keywords>
```

Using a classification-1

```
<textClass>
<catRef target="alltim3 allava2 alltyp3 wriaag0
wriad0 wriase3 wriaty2 wriaud3 wridom8 wrilev2
wrimed1 wripp5 wrisam1 wrista2 writas3"/>
</textClass>
```

using codes to reference an internal closed definition

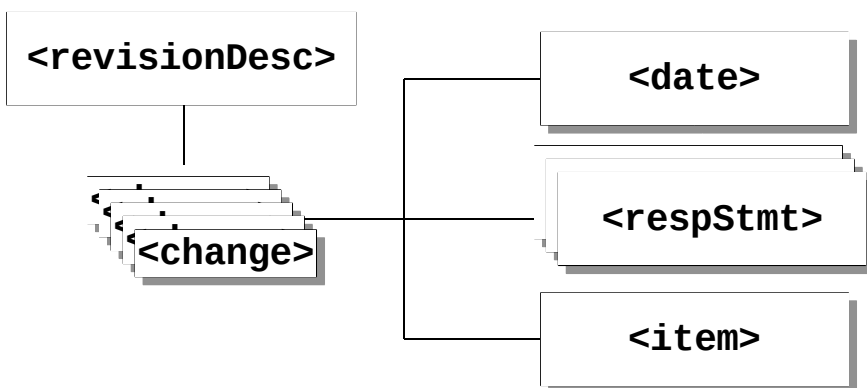
```
<category id="wridom8">
<catDesc>Informative: belief & thought
</catDesc></category>
```

Using a classification-2

```
<classCode scheme="DLee">W religion</classCode>
<keywords scheme="COPAC">
<term>Marriage - Religious aspects
-Christianity</term>
<term>Marriage - Christian viewpoints</term>
</keywords>
<keywords>
<term>Christian guide to marriage</term>
</keywords>
```

using open-ended, externally-defined criteria

The revision description



Sample revision description

```
<revisionDesc>
<change><date>1996-11-19</date>
<respStmt>
<resp>xxx</resp><name>ASM</name></respStmt>
<item>Header Edited</item></change>
<change><date>unknown</date>
<respSstmt>
<resp>xxx</resp><name>JPW</name></respSstmt>
<item>Basic Tei Lite tagging</item></change>
</revisionDesc>
```

Issues in using the Header

- originally intended for non-specialist use
- application profiles for particular uses
- independent headers
- up/down translation

Crosswalks

<title type=" main" >	DC.title.main	246
<author>	DC.creator.name	100
<publicationStmt>	DC.publisher.name	260
<sourceDesc>	DC.source	500, 534
<classDecl>	DC.subject.schema	6XX

Harmonization of practice

- TEI Header is widely deployed, but in varying dialects
- Several attempts have been made to define standard “ codes of practice” for digital libraries using it (e.g. by LC, AHDS)
- It is likely to remain the best “ source of information” for cataloguing use, whether or not it is also the best means of expression.

Creating and managing headers

- As an independent XML document
- Within a database
 - special case of common problem
- OTA experience
 - all data in XML headers
 - harvestable by Z39.50 client
 - OLAC compatibility is a SMOP