Digital Humanities Observatory
Ireland's window on humanities e-scholarship

A project of the RIA ROYAL IRISH ACADEMY
ACADAMH RÍOGA NA hÉIREANN

Introduction to the TEI header

Kevin S. Hawkins

k.hawkins@dho.ie

07.04.2010 • National University of Ireland, Galway

~ FUNDED UNDER PRTLI . CYCLE 4 2008 - 2011 ~

HEA Higher Education Authority An tÚdarás um Ard-Oideachas

NDP National Development Plan 2007 - 2013

EUROPEAN UNION STRUCTURAL FUNDS

What is the TEI header?

The TEI header (<teiHeader>) is the 'virtual title page' of a TEI document. It contains metadata (information about the TEI document).

<teiHeader> is the first, mandatory child element of the root <TEI> element; therefore, it appears at the top ('at the head') of every TEI document.

The header metadata provides:

- a bibliographic record of the electronic text as well as the source from which the electronic text is derived
- documentation of the encoding and editorial principles used in tagging the electronic text
- terms for indexing, searching, and retrieval
- a record of changes made to the electronic document

Structure of the header

The header contains many specialised elements not found anywhere in the ‘body’ of a TEI document (everything after <teiHeader>). These elements allow for highly structured descriptions of the document.

Many parts of the header allow free-form prose descriptions as an alternative to the highly structured descriptions.

Few header elements are required, so a header can be quite minimal.

The four children of <teiHeader>

1. <filedesc>: bibliographic info (*required*)
2. <encodingDesc>: description of encoding practices (*optional*)
3. <profileDesc>: search terms (*optional*)
4. <revisionDesc>: record of changes (*optional*)

The four children of <teiHeader>, plus children of <fileDesc>

1. <filedesc>: bibliographic info (*required*)
    1. <titleStmt> (*required*)
    2. <editionStmt> (*optional*)
    3. <extent> (*optional*)
    4. <publicationStmt> (*required*)
    5. <seriesStmt> (*optional*)
    6. <notesStmt> (*optional*)
    7. <sourceDesc> (*required*)
2. <encodingDesc>: description of encoding practices (*optional*)
3. <profileDesc>: search terms (*optional*)
4. <revisionDesc>: record of changes (*optional*)

The four children of <teiHeader>, plus children of <fileDesc>

1. <filedesc>: bibliographic info (*required*)
    1. <titleStmt> (*required*)
    2. <editionStmt> (*optional*)
    3. <extent> (*optional*)
    4. <publicationStmt> (*required*)
    5. <seriesStmt> (*optional*)
    6. <notesStmt> (*optional*)
    7. <sourceDesc> (*required*) ← description of the source
2. <encodingDesc>: description of encoding practices (*optional*)
3. <profileDesc>: search terms (*optional*)
4. <revisionDesc>: record of changes (*optional*)

All other elements describe the TEI document itself.

The four children of <teiHeader>, plus children of <fileDesc>

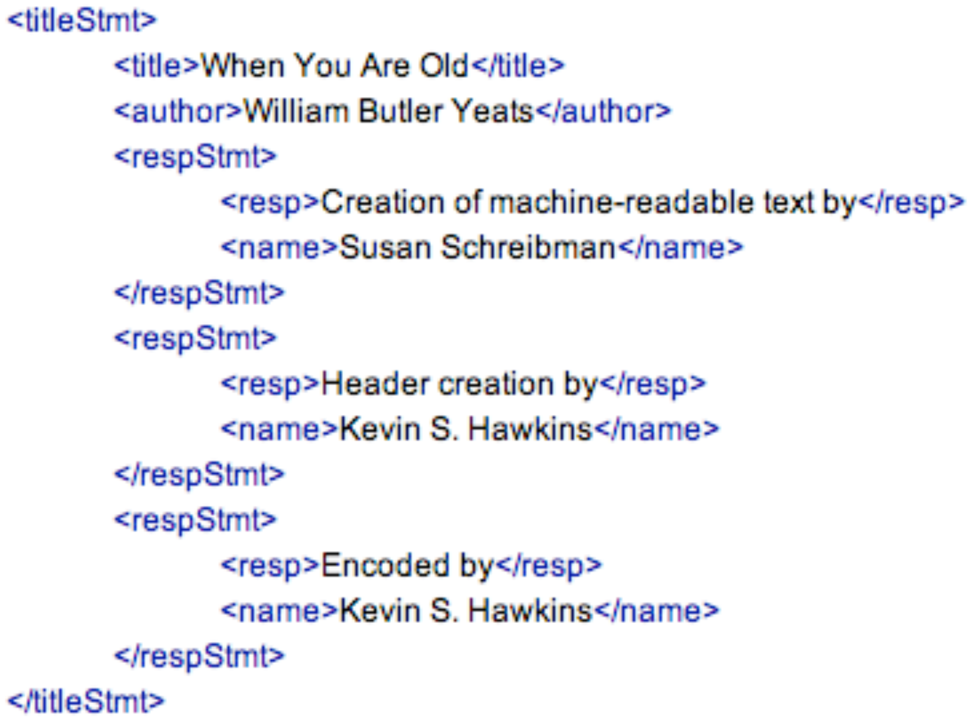
1. <filedesc>: bibliographic info (*required*)
    1. <titleStmt> (*required*)
    2. <editionStmt> (*optional*)
    3. <extent> (*optional*)
    4. <publicationStmt> (*required*)
    5. <seriesStmt> (*optional*)
    6. <notesStmt> (*optional*)
    7. <sourceDesc> (*required*) ← description of the source
2. <encodingDesc>: description of encoding practices (*optional*)
3. <profileDesc>: search terms (*optional*)
4. <revisionDesc>: record of changes (*optional*)

All other elements describe the TEI document itself.

Only these three elements are required!

Children of <fileDesc> (1)

<titleStmt>: title and info about those responsible for intellectual content of the TEI document (*required*)

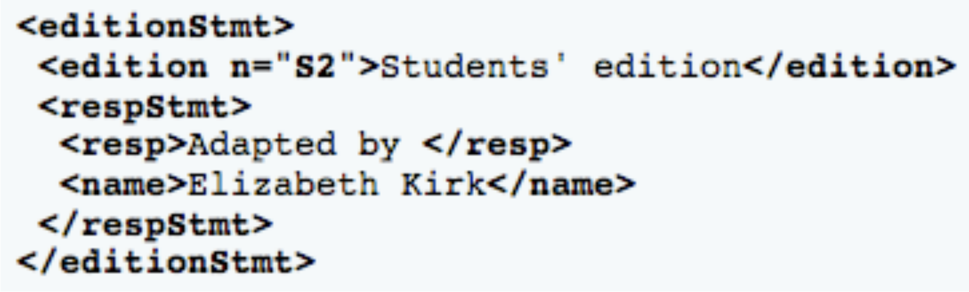


```
<titleStmt>
        <title>When You Are Old</title>
        <author>William Butler Yeats</author>
        <respStmt>
                <resp>Creation of machine-readable text by</resp>
                <name>Susan Schreibman</name>
        </respStmt>
        <respStmt>
                <resp>Header creation by</resp>
                <name>Kevin S. Hawkins</name>
        </respStmt>
        <respStmt>
                <resp>Encoded by</resp>
                <name>Kevin S. Hawkins</name>
        </respStmt>
</titleStmt>
```

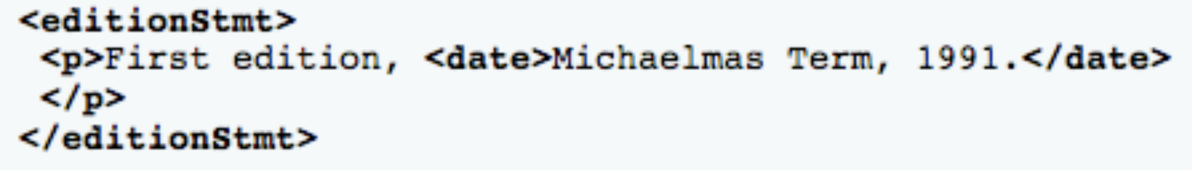
Children of <fileDesc> (2)

<editionStmt>: edition number or other description of the edition (*optional*)

Examples from the TEI Guidelines:



```
<editionStmt>
 <edition n="S2">Students' edition</edition>
 <respStmt>
  <resp>Adapted by </resp>
  <name>Elizabeth Kirk</name>
 </respStmt>
</editionStmt>
```



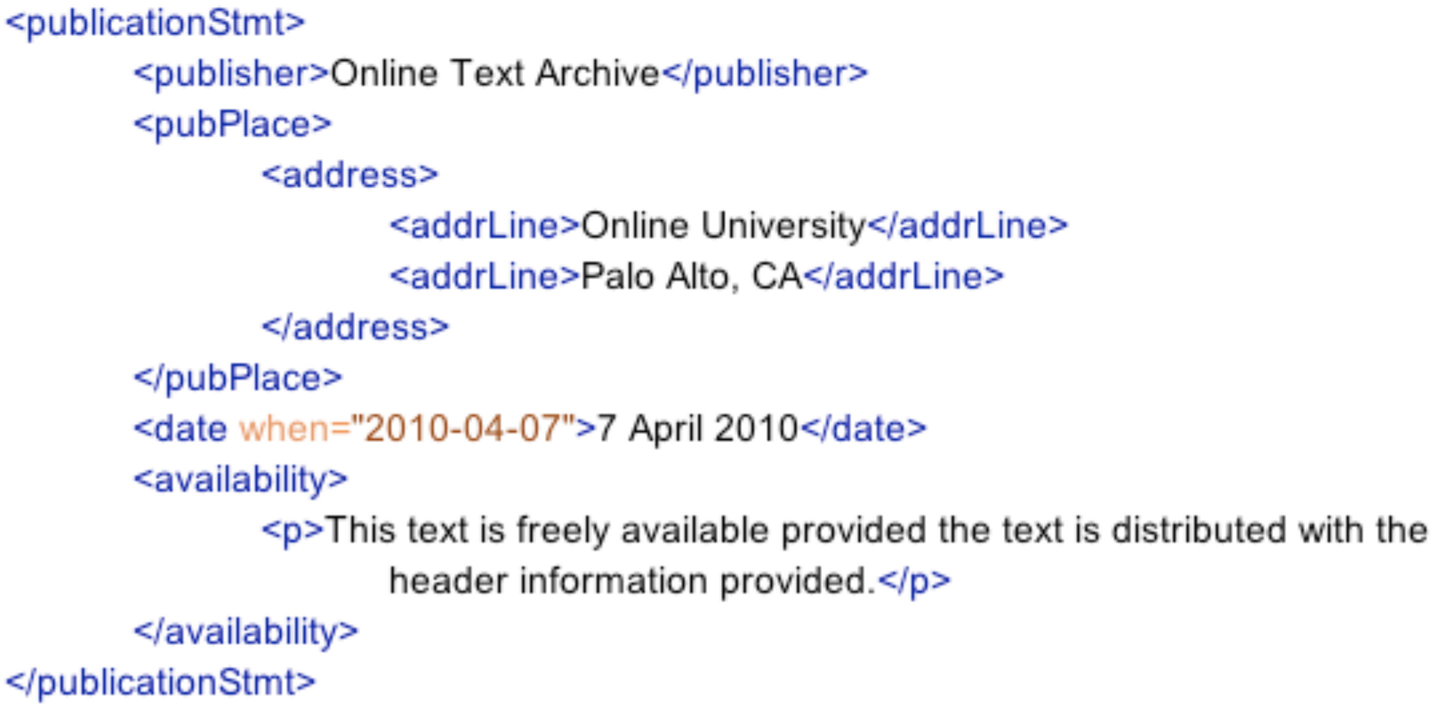
```
<editionStmt>
 <p>First edition, <date>Michaelmas Term, 1991.</date>
 </p>
</editionStmt>
```

Children of <fileDesc> (3)

<extent>: size of the TEI document (in bytes, words, paragraphs, etc.) (*optional*)

Children of <fileDesc> (4)

<publicationStmt>: info about the publication and distribution of the TEI document (*required*)



```
<publicationStmt>
        <publisher>Online Text Archive</publisher>
        <pubPlace>
                <address>
                        <addrLine>Online University</addrLine>
                        <addrLine>Palo Alto, CA</addrLine>
                </address>
        </pubPlace>
        <date when="2010-04-07">7 April 2010</date>
        <availability>
                <p>This text is freely available provided the text is distributed with the
                        header information provided.</p>
        </availability>
</publicationStmt>
```

Children of <fileDesc> (5)

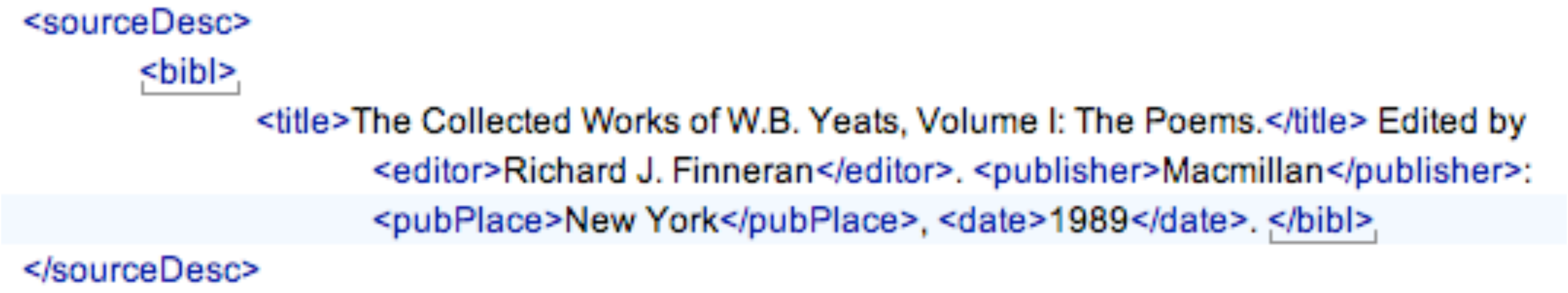
<seriesStmt>: info about the series of which the TEI document is a part (*optional*)

<noteStmt>: bibliographic notes (info that doesn't fit elsewhere) (*optional*)

Children of <fileDesc> (6)

<sourceDesc>: describes the source from which the TEI document was created. Can be:

- a short statement (‘This is a born-digital document.’)
- a semi-structured citation (as below)
- something as detailed as a <fileDesc>.



```
<sourceDesc>
        <bibl>
                <title>The Collected Works of W.B. Yeats, Volume I: The Poems.</title> Edited by
                        <editor>Richard J. Finneran</editor>. <publisher>Macmillan</publisher>:
                        <pubPlace>New York</pubPlace>, <date>1989</date>. </bibl>
</sourceDesc>
```

<encodingDesc>

<encodingDesc> ‘describes the relationship between an electronic text and its source or sources. It allows for detailed description of whether (or how) the text was normalized during transcription, how the encoder resolved ambiguities in the source, what levels of encoding or analysis were applied, and similar matters’ (from the TEI Guidelines)

Can contain a prose description or use up to seven specialised child elements …

Children of <encodingDesc> (1)

- <projectDesc> describes the overall project purpose and process
- <samplingDecl> documents rationale for text sampling or selection in case parts of text or corpus have been omitted

Children of <encodingDesc> (2)

- <editorialDecl> explains editorial principles of encoding or transcribing texts. Can contain a prose description or use up to seven specialised child elements to describe:
  - corrections or normalisation performed during the transcription
  - handling of quotation marks and hyphenation
  - any standardisation of dates or numbers performed
  - analytic or interpretive information added to the text

Children of <encodingDesc> (3)

- <tagsDecl> records how tags are used and how their content should be displayed by default
- <refsDecl> specifies how canonical references are constructed in the TEI document
- <classDecl> gives information about any systems for classification used in the TEI document
- <appInfo> can be used to record information about programs which have acted upon the TEI document

<profileDesc>

<profileDesc> contains ‘classificatory and contextual information about the text, such as its subject matter, the situation in which it was produced, the individuals described by or participating in producing it, and so forth. Such a text profile is of particular use in highly structured composite texts such as corpora or language collections, where it is often highly desirable to enforce a controlled descriptive vocabulary or to perform retrievals from a body of text in terms of text type or origin. The text profile may however be of use in any form of automatic text processing’ (from the TEI Guidelines)

Can contain a number of specialised child elements …

Children of <profileDesc> (1)

- <creation> contains info about the origin of a text, such as its date and place of creation (when this information isn't clear from the bibliographic info)
- <langUsage> describes languages used in a text.
- <textClass> allows you to assign terms for classification (such as subject headings and other controlled vocabularies) to a text

Children of <profileDesc> (2)

There are other elements for use with:

- linguistic corpora – to describe the linguistic context)
- manuscript transcriptions – to describe the ‘hands’ identified in the manuscript

<revisionDesc>

<revisionDesc> ‘allows the encoder to provide a history of changes made during the development of the electronic text. The revision history is important for version control and for resolving questions about the history of a file.’ (from the TEI Guidelines)

This contains individual <change> elements, each of which describes a change and indicates who made it.

This looks like a lot of work …

Creating good, consistent metadata for a collection of documents is hard, and it's not something most of us find interesting.

However, digital texts, just like the primary source material we all study, often end up being studied in ways that the authors never intended or even imagined. It's good to give as much context about the text as is feasible to help others make use of the TEI document in the future.

How much detail? (1)

There’s no one answer to this question.

If something is easy to identify, take a bit of extra time to do it.

If you would have to do research to know the answer, think about how easily someone might be able to do the same research in the future.

- Is the answer available in reference works, or is it only determinable by working with primary source materials such as the ones you’re encoding? If the latter, that sounds like something worth identifying.

How much detail? (2)

Some header elements date to an earlier era, when files and the systems they are stored in were less integrated. There’s some information which you might not bother recording in the header if the data is reliably stored elsewhere. For example:

- <extent> in the <fileDesc>
- <revisionDesc>

How much detail? (3)

Don’t include header elements if the information is clearly and readily reconstructable from the body of the TEI document. For example:

- <langUsage>: Only include this in the header if you want to elaborate beyond use of the xml:lang= attribute used in the body.

Also keep in mind …

Most encoding projects involve encoding more than one text. So you can use a template to create your headers since a lot of the information is the same in all of them.

Your collection may end up being aggregated with other collections at an institution. Speak to those involved to make sure you all structure your headers in a way that makes them compatible with each other:

- use the same elements in the same way
- use controlled vocabularies, thesauri, and authority lists

Controlled vocabularies, thesauri and authority files

A **controlled vocabulary** is a standard set of keywords designed to cover a particular area of study.

A **thesaurus** or **authority file** is a controlled vocabulary containing synonyms pointing to the ‘authorised’ form that you should use. Some thesauri even contain a hierarchy of terms.

Controlled vocabularies, thesauri and authority files

Some controlled vocabularies are built into the TEI (like codes for languages). Others are given in the TEI as suggestions (like Library of Congress Subject Headings).

If you use the authorized forms of names, you can disambiguate people with similar names, and your users will be able to search your materials with other materials.

There are lots of controlled vocabularies out there. Don’t ‘reinvent the wheel’!

Some examples

- Library of Congress Authorities:
  - subject headings (LCSH)
  - names of authors, editors, etc.
  - titles of well-known literary works
- Getty Thesaurus of Geographical Names
- Placenames Database of Ireland
- Northern Ireland Place-Name Project
- *Dictionary of Irish Biography*

Questions?