# Introduction to the TEI header

Creating Digital Editions:
An Introduction to the Text Encoding Initiative
January 17, 2014

# 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 may contain documentation about four things:

- Bibliographic description of the text being encoded *(required)*
- Decisions made about how to encode the text *(recommended)*
- Detailed description of relevant non-bibliographic elements of a text *(optional)*
- A record of changes made to the electronic document *(recommended)*

# The four children of <teiHeader>

1.<fileDesc>: bibliographic info (*required*)

2.<encodingDesc>: description of encoding practices (*recommended*)

3.<profileDesc>: search terms (*optional*)

4.<revisionDesc>: record of changes (*recommended*)

# Structure of the header

The header contains many specialized elements not found anywhere in the ‘body’ of a TEI document (that is, everything after the close of <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.

# Bibliographic Information <fileDesc>

You must:

- Give your TEI document a title **<titleStmt>**
- State something about the publisher/publication of the TEI document **<publicationStmt>**
- Describe the source text that you are encoding (all other parts of the header describe the electronic file, as opposed to the source) **<sourceDesc>**

You may:

Document other bibliographic details such as editions, series, and the extent of the text or file. **<editionStmt>, <seriesStmt>, <extent>, <notesStmt>**

# Decisions about encoding <encodingDesc>

describes the relationship between an electronic text and its source or sources: what did the encoder choose to include, exclude, address, ignore, or change between the source text and the TEI document?

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

http://www.tei-c.org/release/doc/tei-p5-doc/en/html/ref-encodingDesc.html

# Decisions about encoding <encodingDesc>

You may:

- Describe the overall project purpose and process <projectDesc>

-Document rationale for text sampling or selection in case parts of text or corpus have been omitted <samplingDecl>

- Explain editorial principles for transcription, encoding (such as normalization, correction, or standardization of spelling, numbers, punctuation, etc.) <editorialDecl>

- And more (<tagsDecl>, <refsDecl>, <classDecl>, <appInfo>)

# Non-Bibliographic Information <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)

http://www.tei-c.org/release/doc/tei-p5-doc/en/html/ref-profileDesc.html

# Non-Bibliographic Information <profileDesc>

You may provide information about the creation <creation>, languages <langUsage>, and classification <textClass> of the text.

There are also specialized elements for providing context for linguistic corpora or identifying the hands of copyists involved in the production of a manuscript.

# 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

http://authorities.loc.gov/

Getty Thesaurus of Geographical Names

http://www.getty.edu/research/conducting_research/vocabularies/tgn/

Art and Architecture Thesaurus

http://www.getty.edu/research/conducting_research/vocabularies/aat/

# <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)

http://www.tei-c.org/release/doc/tei-p5-doc/en/html/ref-revisionDesc.html

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

http://www.tei-c.org/release/doc/tei-p5-doc/en/html/ref-change.html

# 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)

A very minimal, valid TEI header might look something like this:

```
<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>
   <distributor>Oxford Text Archive</distributor>
  </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>
```

# How much detail? (3)

**Avoid redundancy:**

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? (4)

**Avoid redundancy:**

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

Questions?