Textual Editing in the Digital Age

Enriching your TEI Metadata

•••

James Cummings

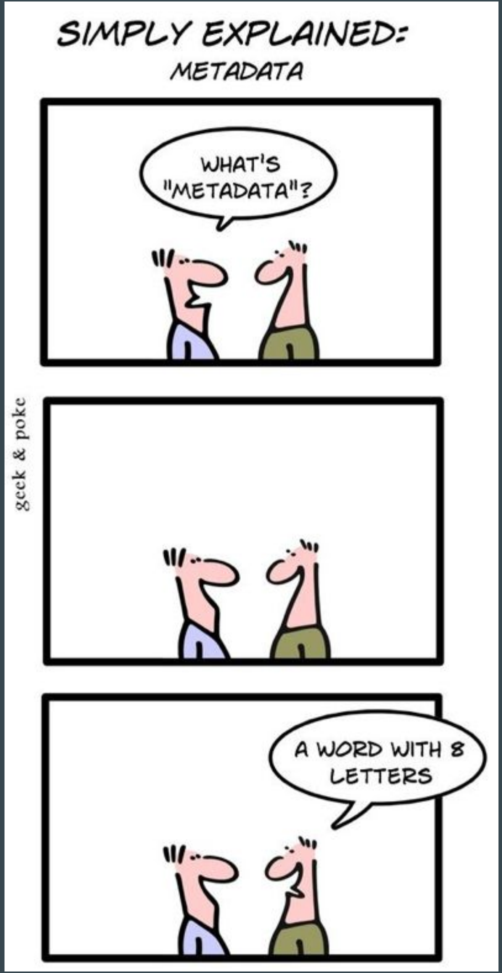
ATNU
ANIMATING TEXT
NEWCASTLE UNIVERSITY

IES INSTITUTE OF ENGLISH STUDIES
SCHOOL OF ADVANCED STUDY
UNIVERSITY OF LONDON

Newcastle University

This lecture

- Some history about the <teiHeader> (librarians vs editors)
- The structure of the <teiHeader>
- All the little boxes of metadata and what goes where
- Some cool things you can do in your <teiHeader>



What is Metadata?

- often called "data about data"
- term originally used only with electronic data but its meaning has broadened
- data about the content, context, and structure of information resources
- the catalogue record of the data/text/edition

Librarians vs Editors: Some history about the <teiHeader>

Librarian’s Header:

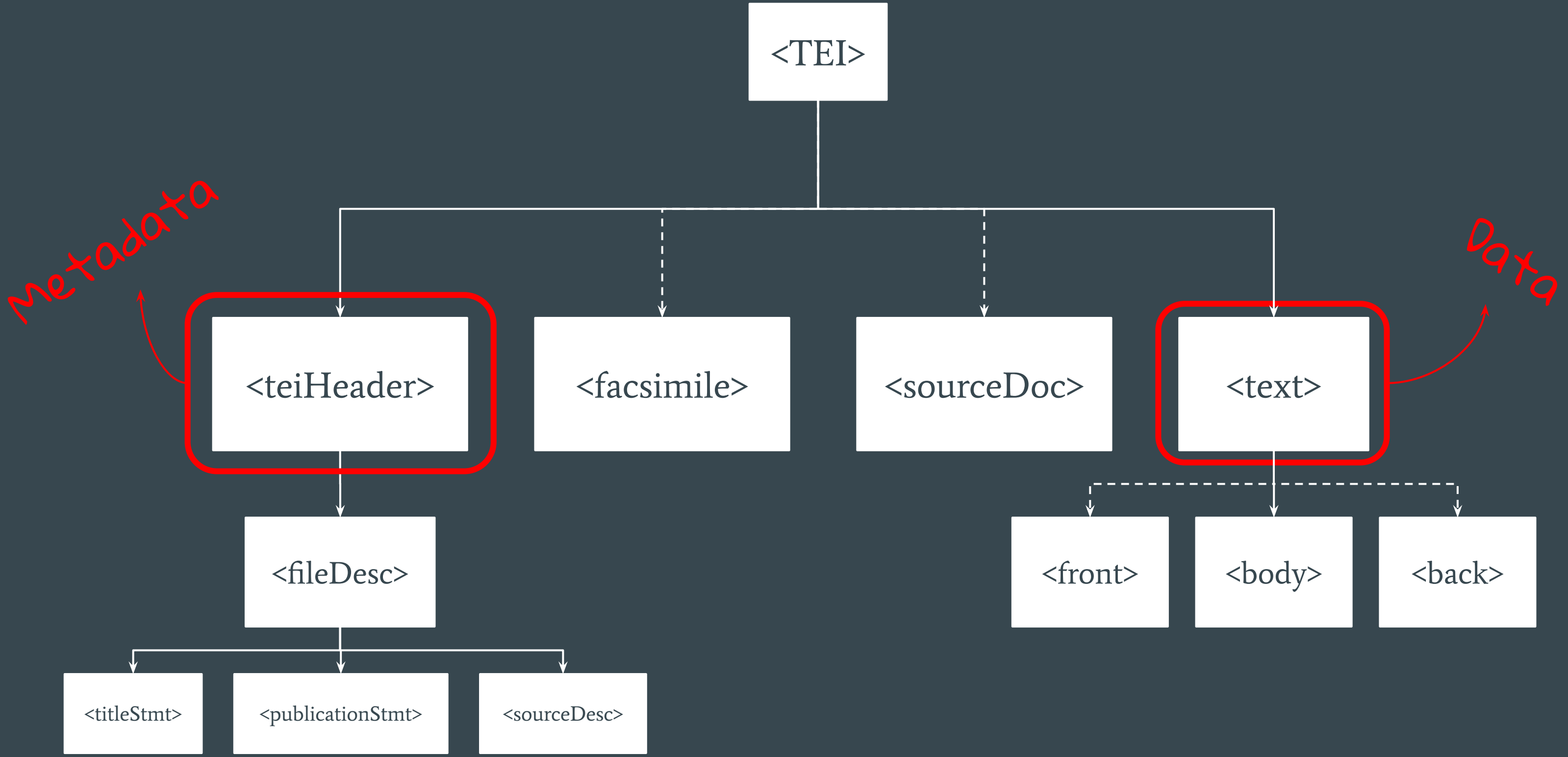
- Conforms to standard bibliographic models
- Easily mapped to METS/EAD/M and other librar
- Based on T Interest Gro
- Pressure for ific constraints
- Prefers structured data over loose prose

Editor’s Header:

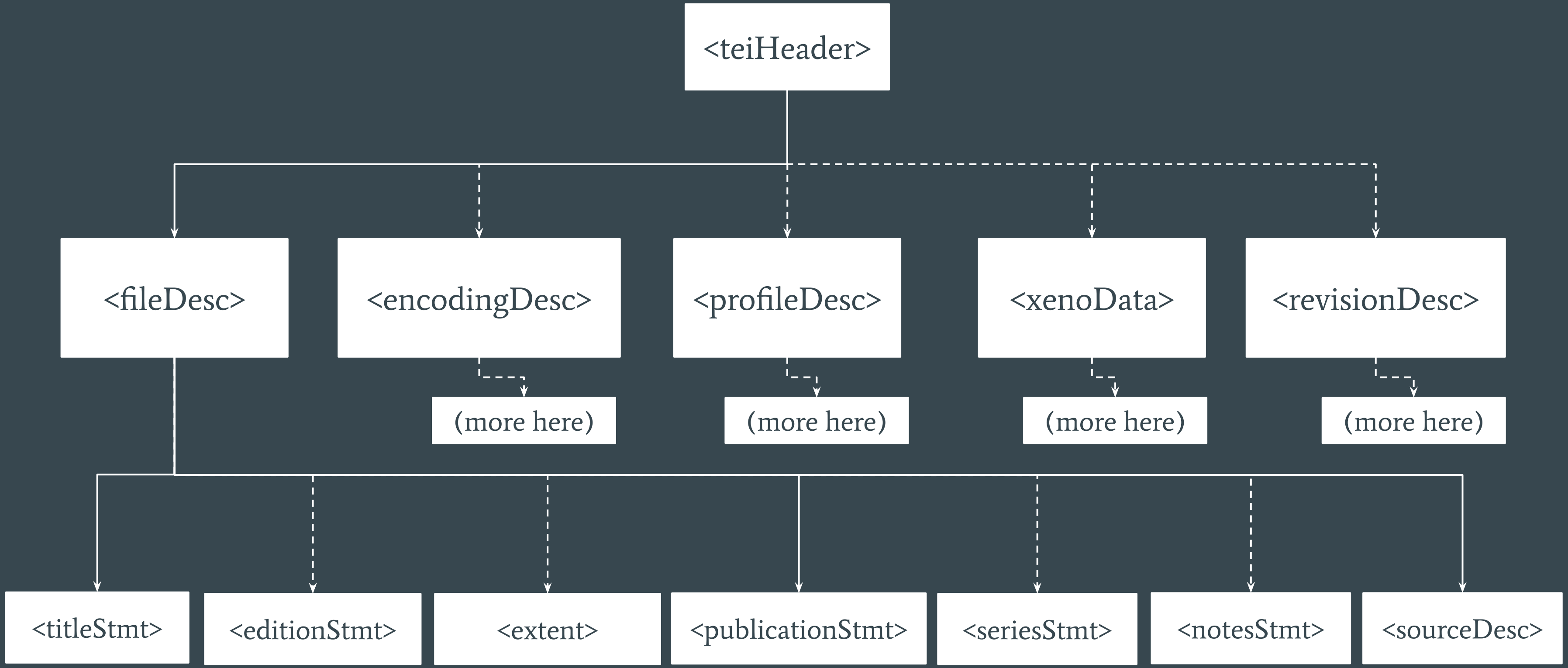
- Polite nod to bibli graphic practices
- S y) huge range of ation ctice in encoding communities
- Often concerned with editorial principles and project documentation
- Mixture of tightly described sections and loose prose

Reality: Most <teiHeader>s are somewhere between the two

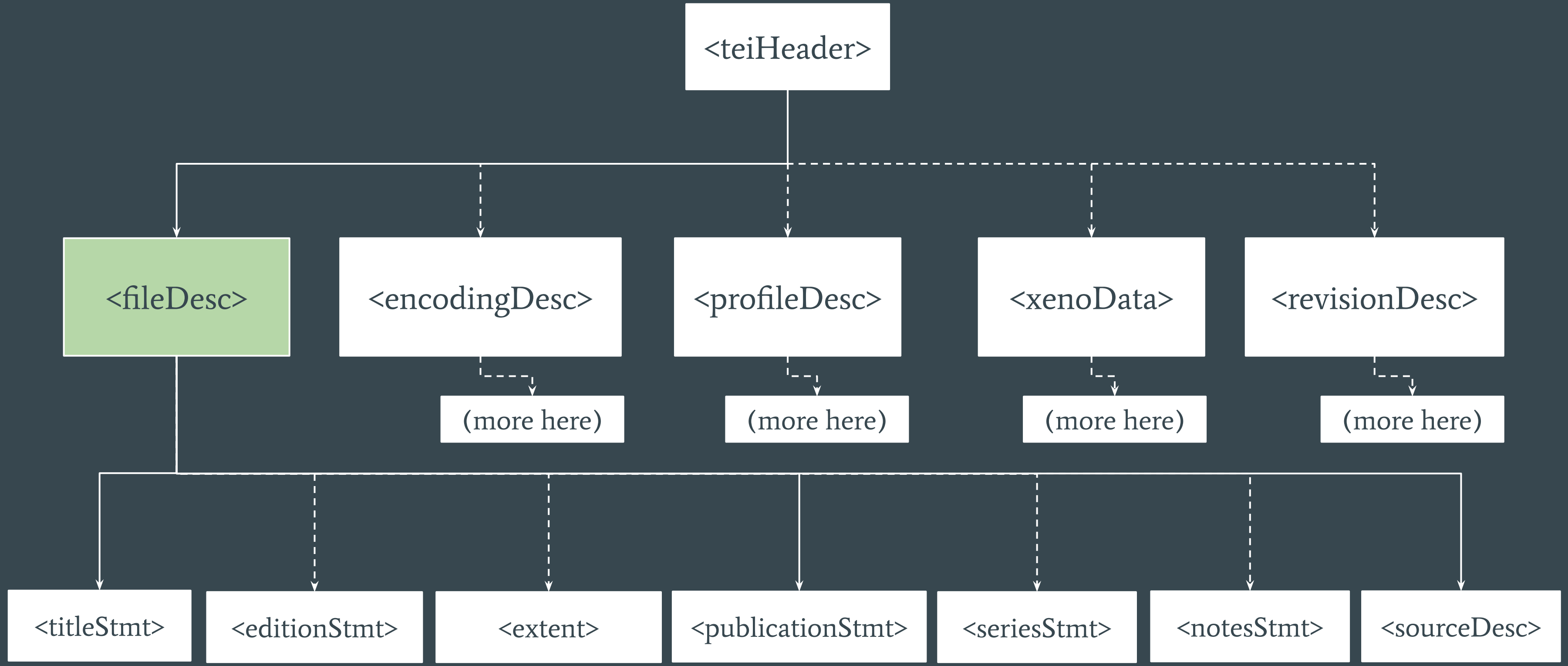
What is the structure of a TEI document?



What is the structure of a <teiHeader>?

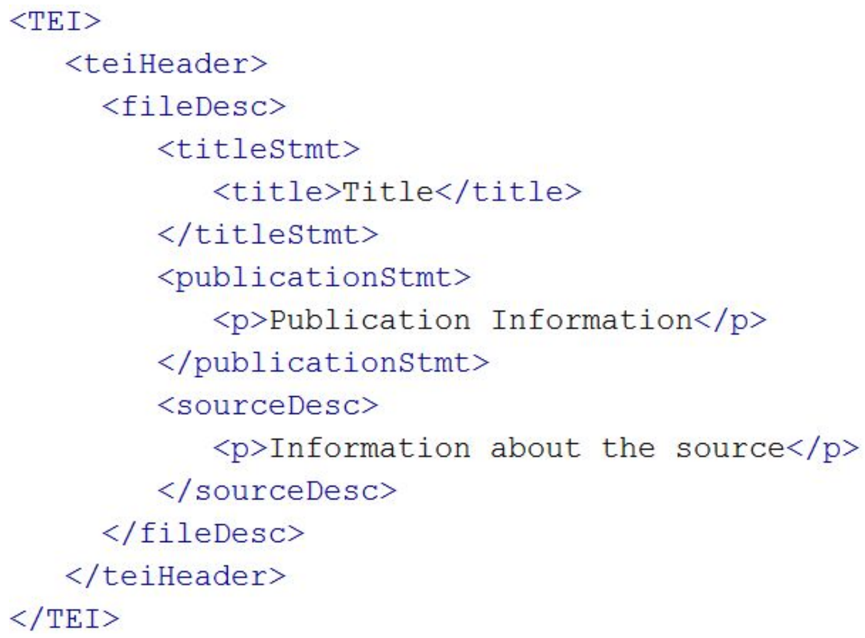


What is the structure of a <teiHeader>?



Most minimal <teiHeader>

- Added required child elements to <fileDesc>:
    - <titleStmt>
    - <publicationStmt>
    - <sourceDesc>
- Populated elements with bare minimum



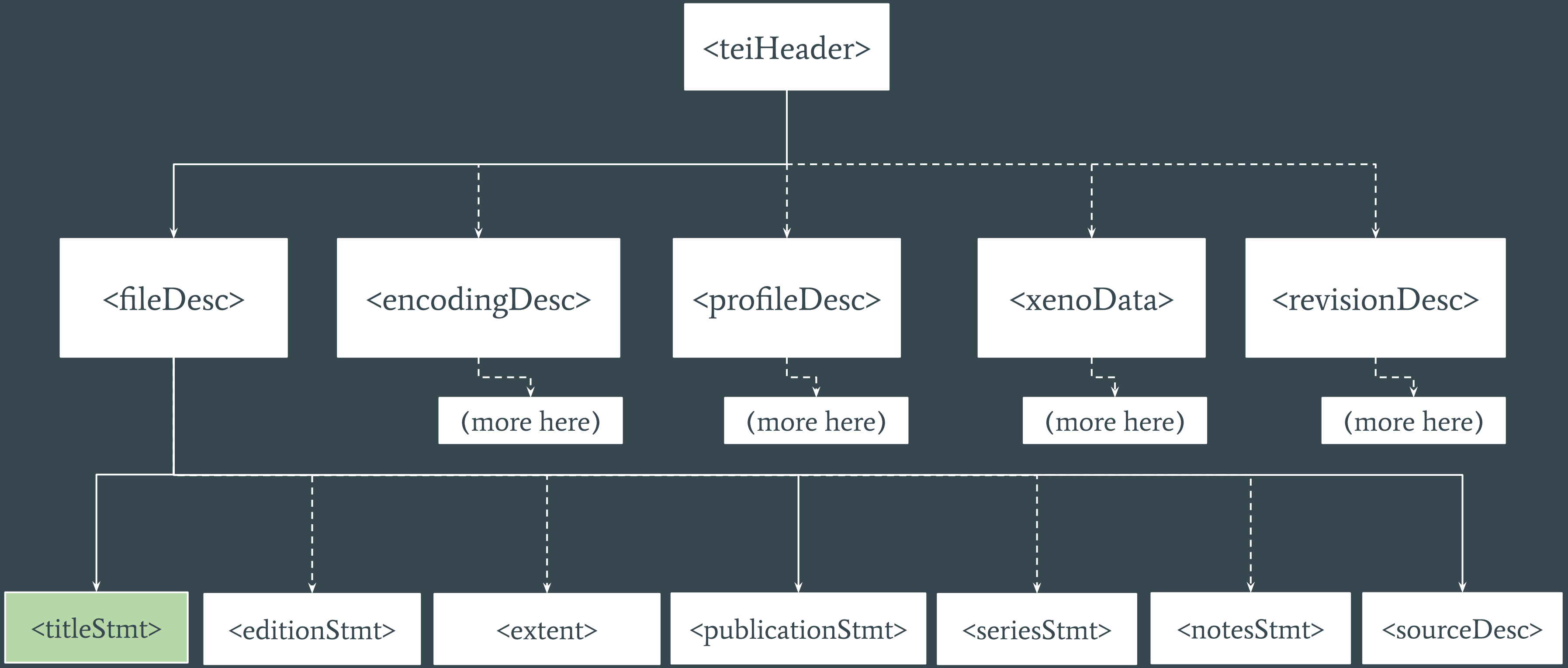
```
<TEI>
   <teiHeader>
     <fileDesc>
        <titleStmt>
           <title>Title</title>
        </titleStmt>
        <publicationStmt>
           <p>Publication Information</p>
        </publicationStmt>
        <sourceDesc>
           <p>Information about the source</p>
        </sourceDesc>
     </fileDesc>
   </teiHeader>
</TEI>
```

<fileDesc>

What is it?

- Contains bibliographic description of the electronic file, including:
  - Title Statement
  - Edition Statement (optional)
  - Extent (optional)
  - Publication Statement
  - Series Statement (optional)
  - Notes Statement (optional)
  - Source Description
- It is important to remember that this is a description of the electronic file

What is the structure of a <teiHeader>?

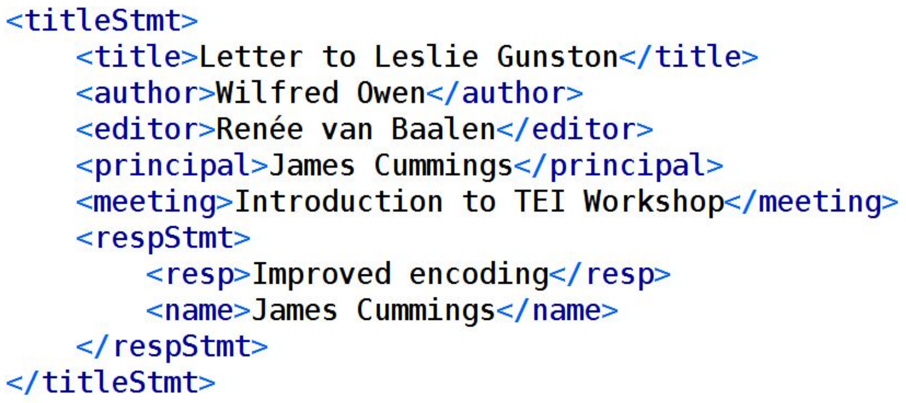


<titleStmt>

- Mandatory (must contain a title)
- <titleStmt>: contains a mandatory <title> which identifies the electronic file

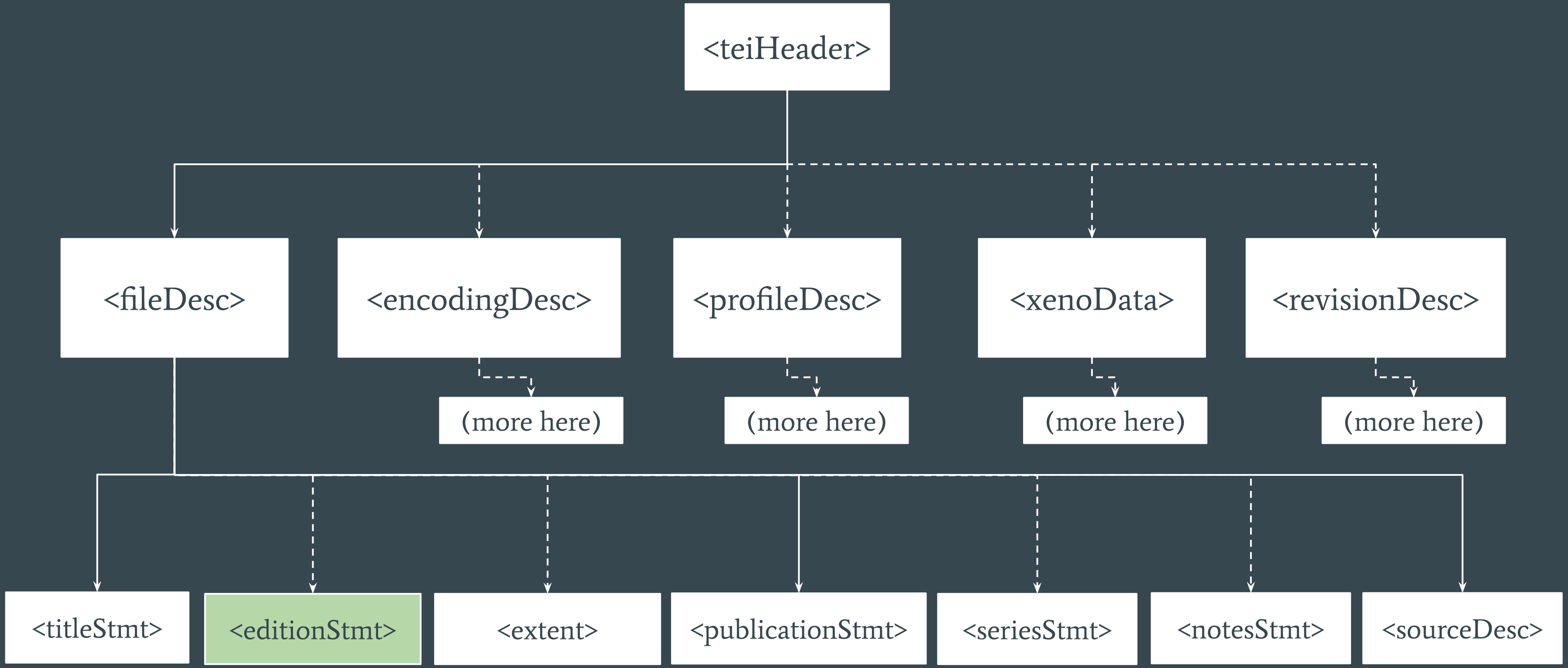
optionally followed by:

- additional titles
- statements of responsibility
  e.g. <author>, <editor>, <sponsor>, <funder>, <principal>, <meeting>, or the generic <respStmt>



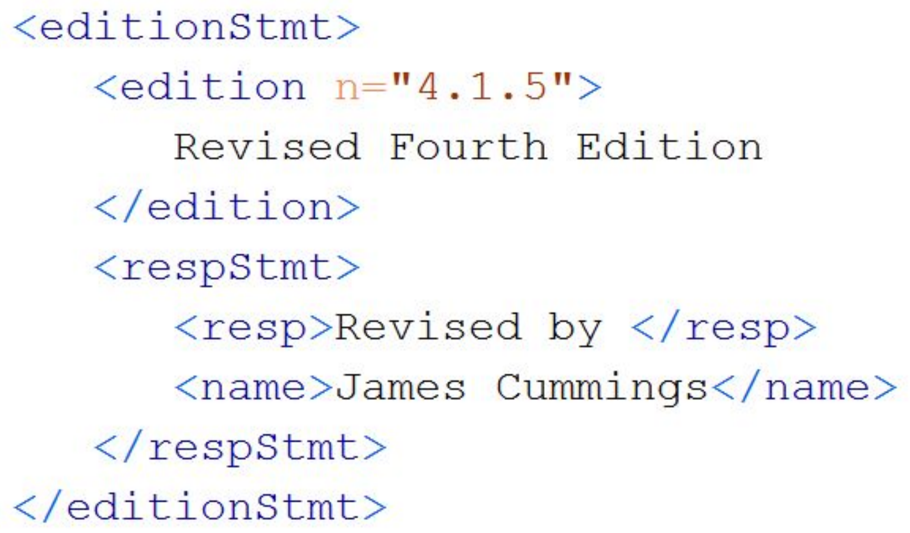
```
<titleStmt>
    <title>Letter to Leslie Gunston</title>
    <author>Wilfred Owen</author>
    <editor>Renée van Baalen</editor>
    <principal>James Cummings</principal>
    <meeting>Introduction to TEI Workshop</meeting>
    <respStmt>
        <resp>Improved encoding</resp>
        <name>James Cummings</name>
    </respStmt>
</titleStmt>
```

What is the structure of a <teiHeader>?



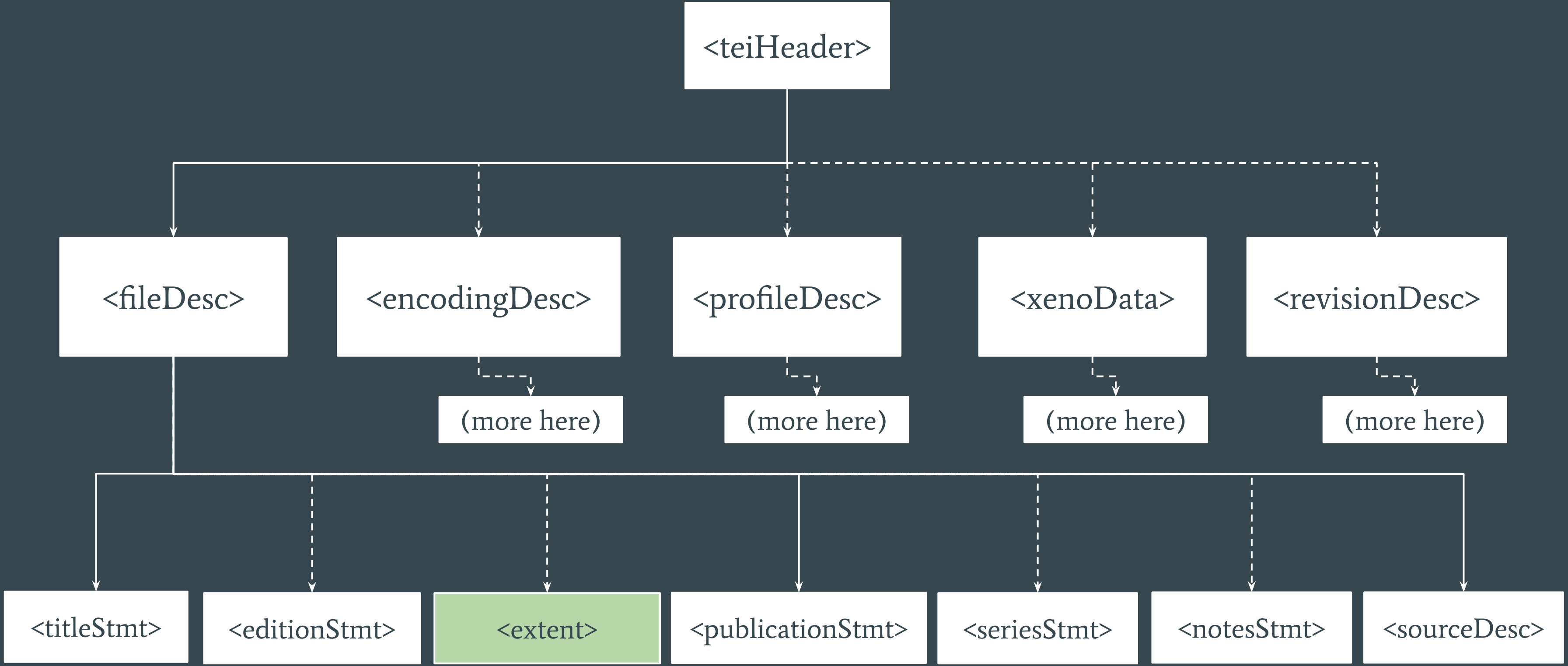
<editionStmt>

- Groups information relating to one edition of a text
- In addition to an <edition> element it may have any of the responsibility elements <author>, <editor>, <meeting>, <funder>, <principal>, <sponsor>, and general <respStmt>



```
<editionStmt>
   <edition n="4.1.5">
      Revised Fourth Edition
   </edition>
   <respStmt>
      <resp>Revised by </resp>
      <name>James Cummings</name>
   </respStmt>
</editionStmt>
```

What is the structure of a <teiHeader>?



<teiHeader>

<fileDesc>

<encodingDesc>

<profileDesc>

<xenoData>

<revisionDesc>

(more here)

(more here)

(more here)

(more here)

<titleStmt>

<editionStmt>

<extent>

<publicationStmt>

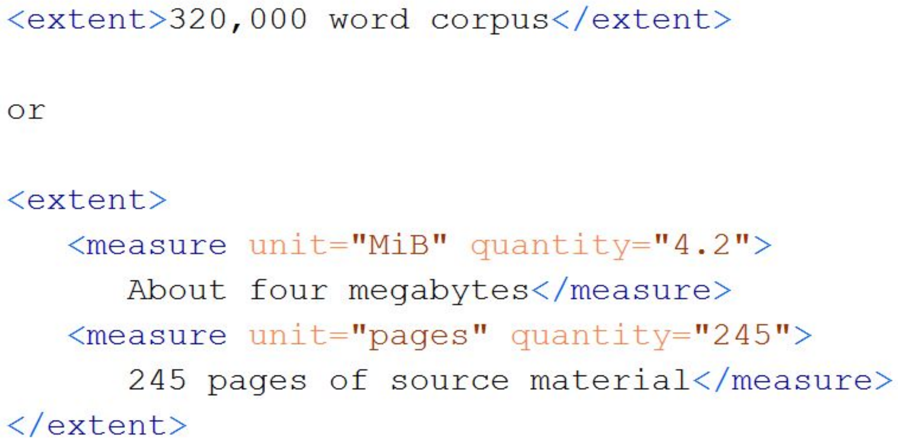
<seriesStmt>

<notesStmt>

<sourceDesc>

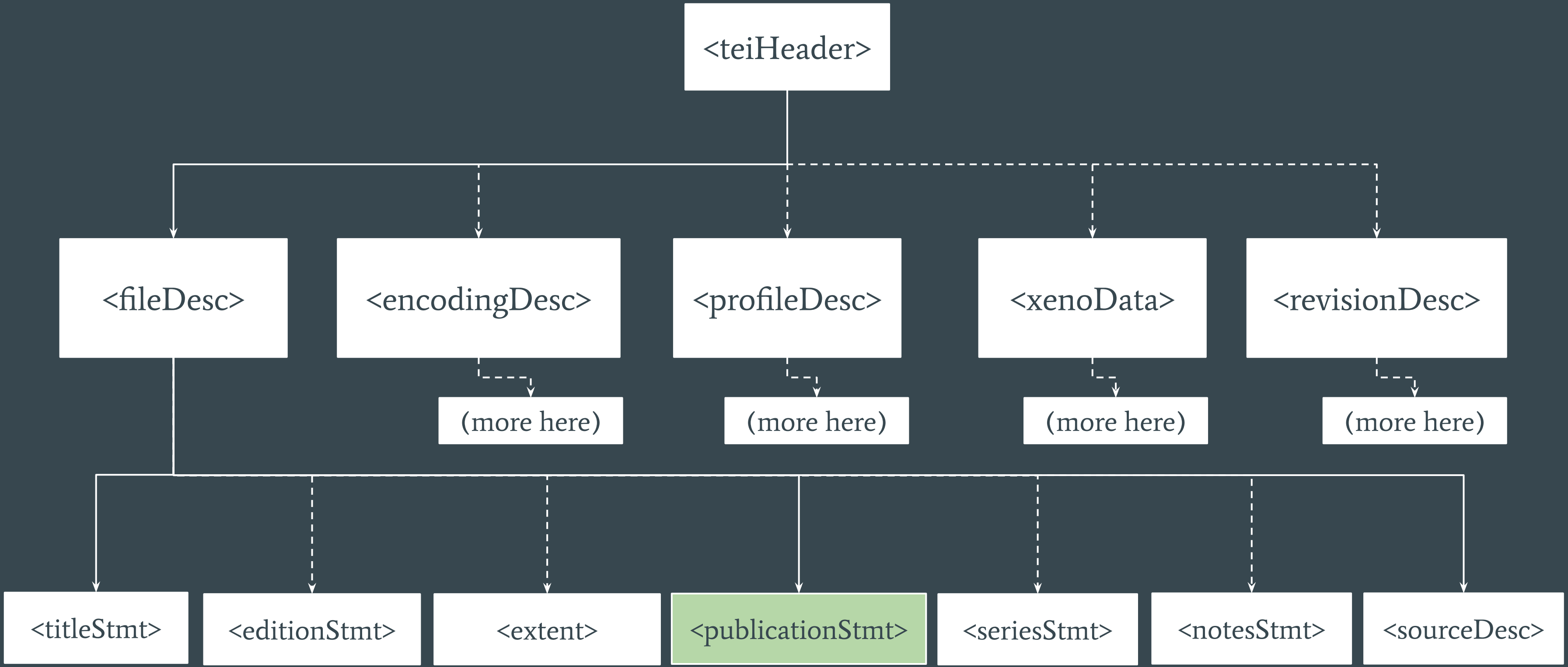
<extent>

- This is for the extent of the electronic version that this <teiHeader> describes
- Can be given in any useful figure for later use (e.g. in catalogues), for example, the number of words, gigabytes, files, entriesm interviews, chapters, stories – whatever makes sense to this form of text



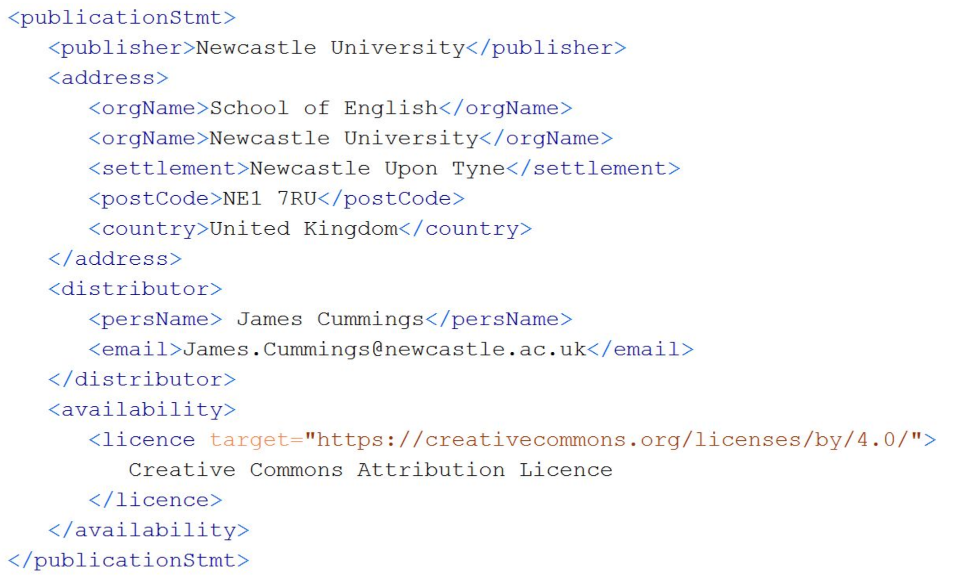
```
<extent>320,000 word corpus</extent>

or

<extent>
   <measure unit="MiB" quantity="4.2">
      About four megabytes</measure>
   <measure unit="pages" quantity="245">
      245 pages of source material</measure>
</extent>
```

What is the structure of a <teiHeader>?



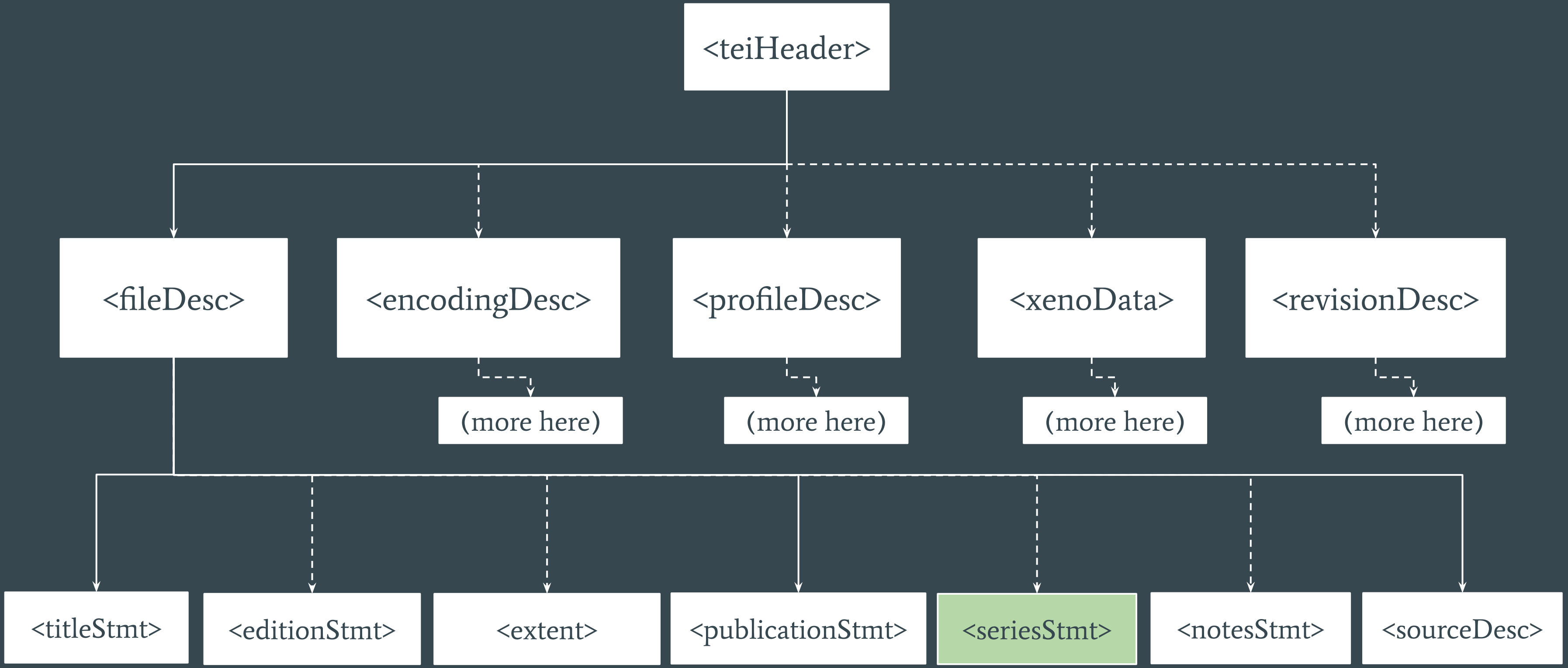
<publicationStmt>

- Mandatory element
- At least one of <publisher>,<distributor> and/or <authority> must be present unless the entire publication statement is given as prose paragraphs using <p>
- A formal license may be entered in <licence> included in <availability>
- Creation date (stored in <profileDesc>) is different from publication date



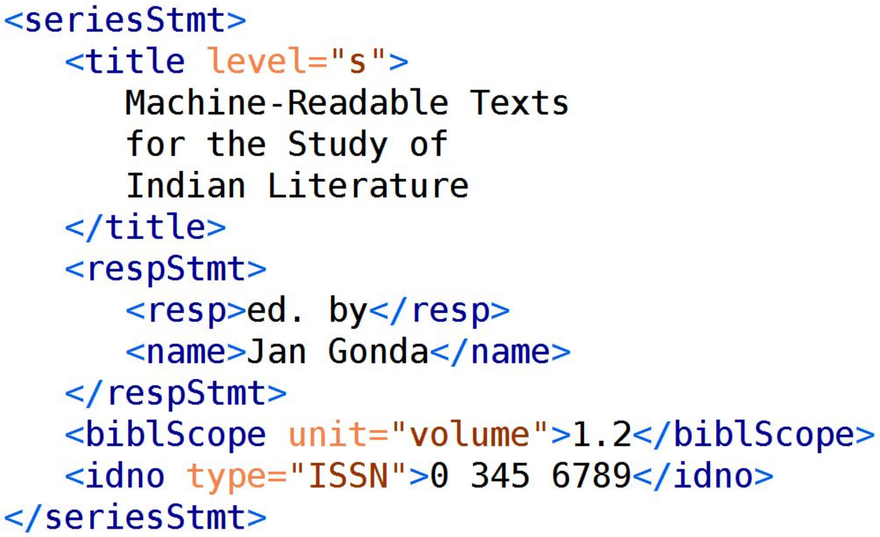
```
<publicationStmt>
   <publisher>Newcastle University</publisher>
   <address>
      <orgName>School of English</orgName>
      <orgName>Newcastle University</orgName>
      <settlement>Newcastle Upon Tyne</settlement>
      <postCode>NE1 7RU</postCode>
      <country>United Kingdom</country>
   </address>
   <distributor>
      <persName> James Cummings</persName>
      <email>James.Cummings@newcastle.ac.uk</email>
   </distributor>
   <availability>
      <licence target="https://creativecommons.org/licenses/by/4.0/">
         Creative Commons Attribution Licence
      </licence>
   </availability>
</publicationStmt>
```

What is the structure of a <teiHeader>?



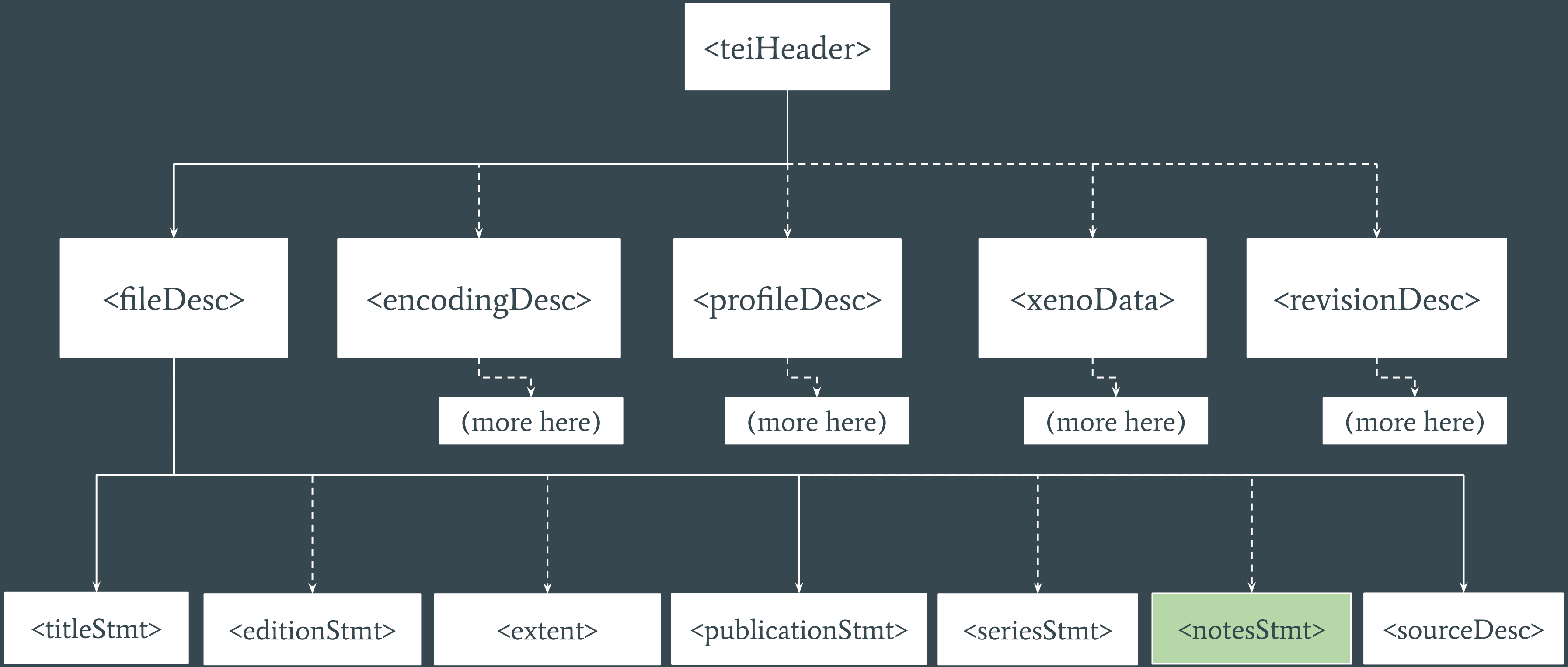
<seriesStmt>

- Groups information about the series, if any, to which the electronic file belongs
- Can contain elements like: <biblScope>, <editor>, <respStmt>, <title>, <idno>
- Or prose paragraphs
- (Yes, digital editions come in series!)



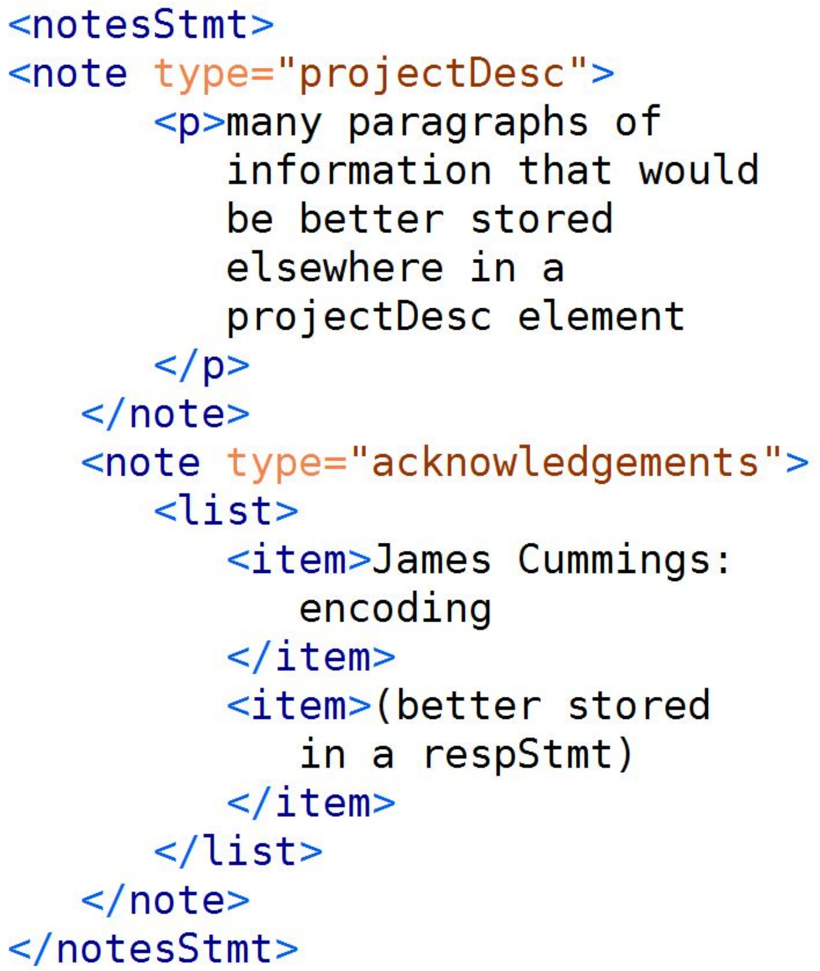
```xml
<seriesStmt>
   <title level="s">
      Machine-Readable Texts
      for the Study of
      Indian Literature
   </title>
   <respStmt>
      <resp>ed. by</resp>
      <name>Jan Gonda</name>
   </respStmt>
   <biblScope unit="volume">1.2</biblScope>
   <idno type="ISSN">0 345 6789</idno>
</seriesStmt>
```

What is the structure of a <teiHeader>?



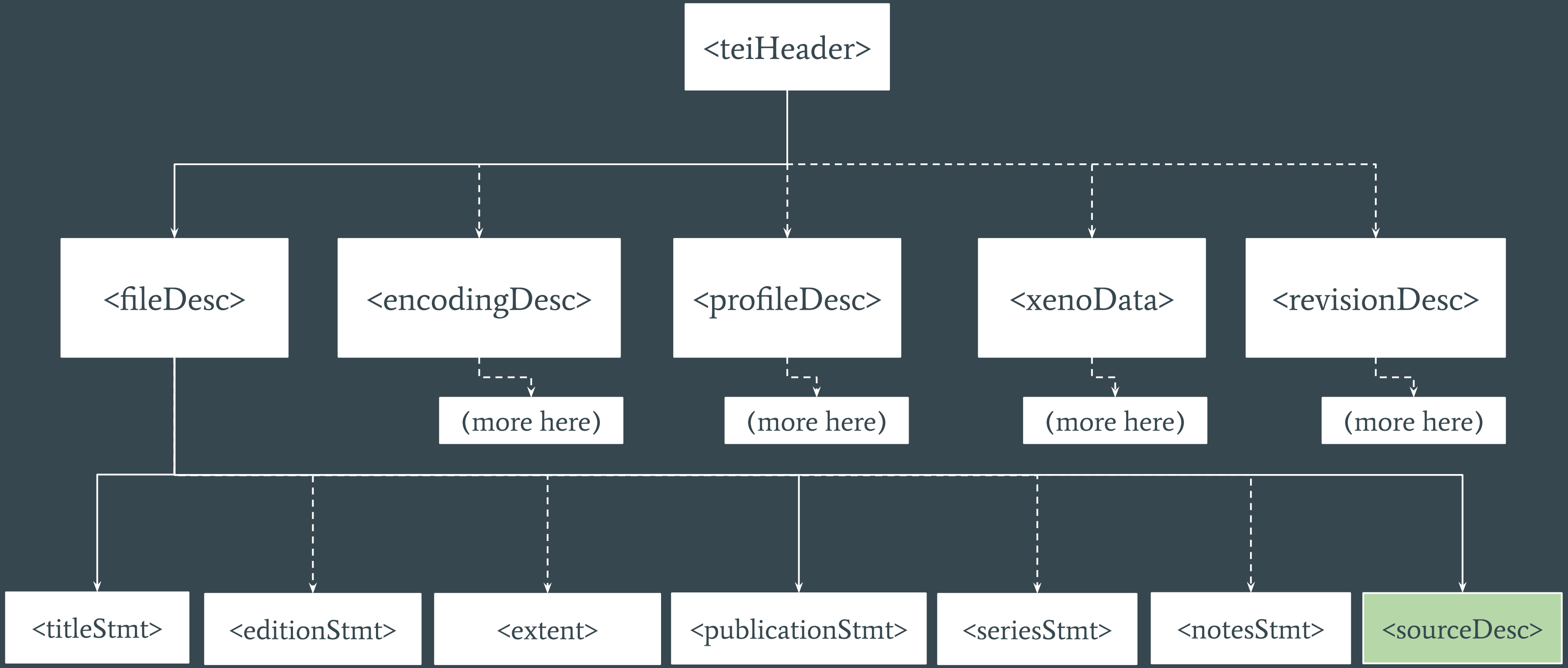
<notesStmt>

- Collects together any notes providing information about a text additional to that recorded in other parts of the bibliographic description
- Each <note> inside can just be some text or many paragraphs of text, classified with a type attribute
- In practice used by projects for all sorts of ad hoc information (which sometimes would be better stored elsewhere)



```xml
<notesStmt>
<note type="projectDesc">
      <p>many paragraphs of
         information that would
         be better stored
         elsewhere in a
         projectDesc element
      </p>
   </note>
   <note type="acknowledgements">
      <list>
         <item>James Cummings:
            encoding
         </item>
         <item>(better stored
            in a respStmt)
         </item>
      </list>
   </note>
</notesStmt>
```

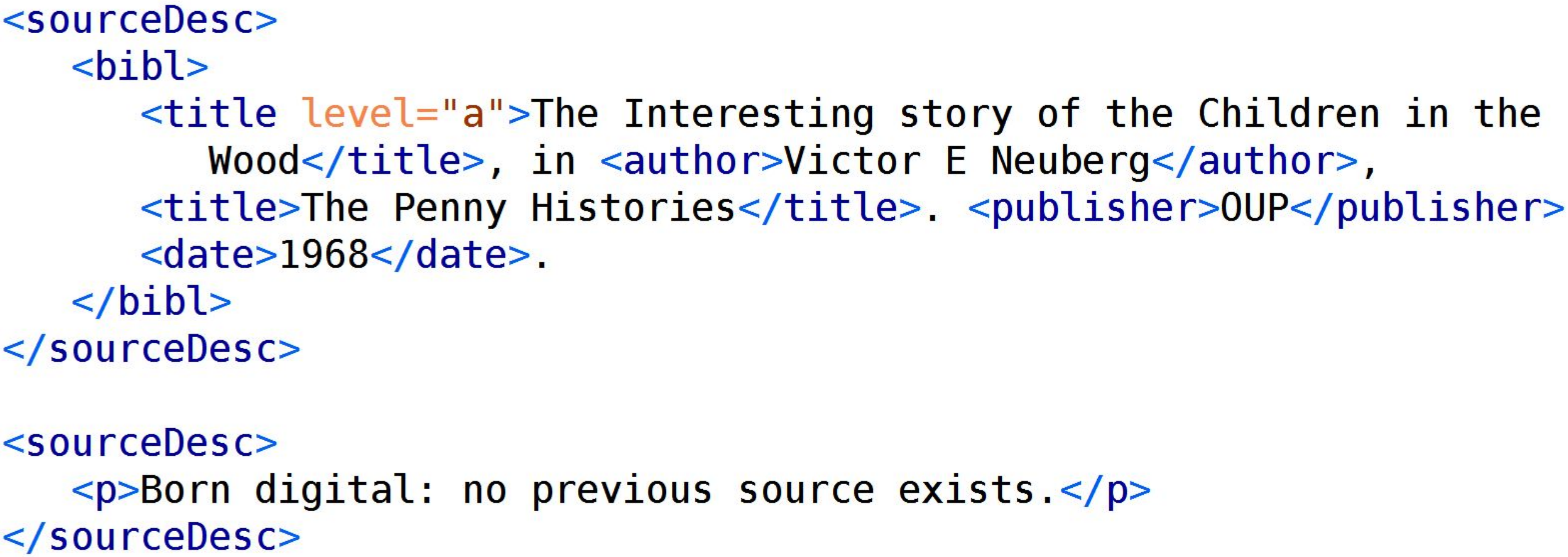
What is the structure of a <teiHeader>?



<sourceDesc>

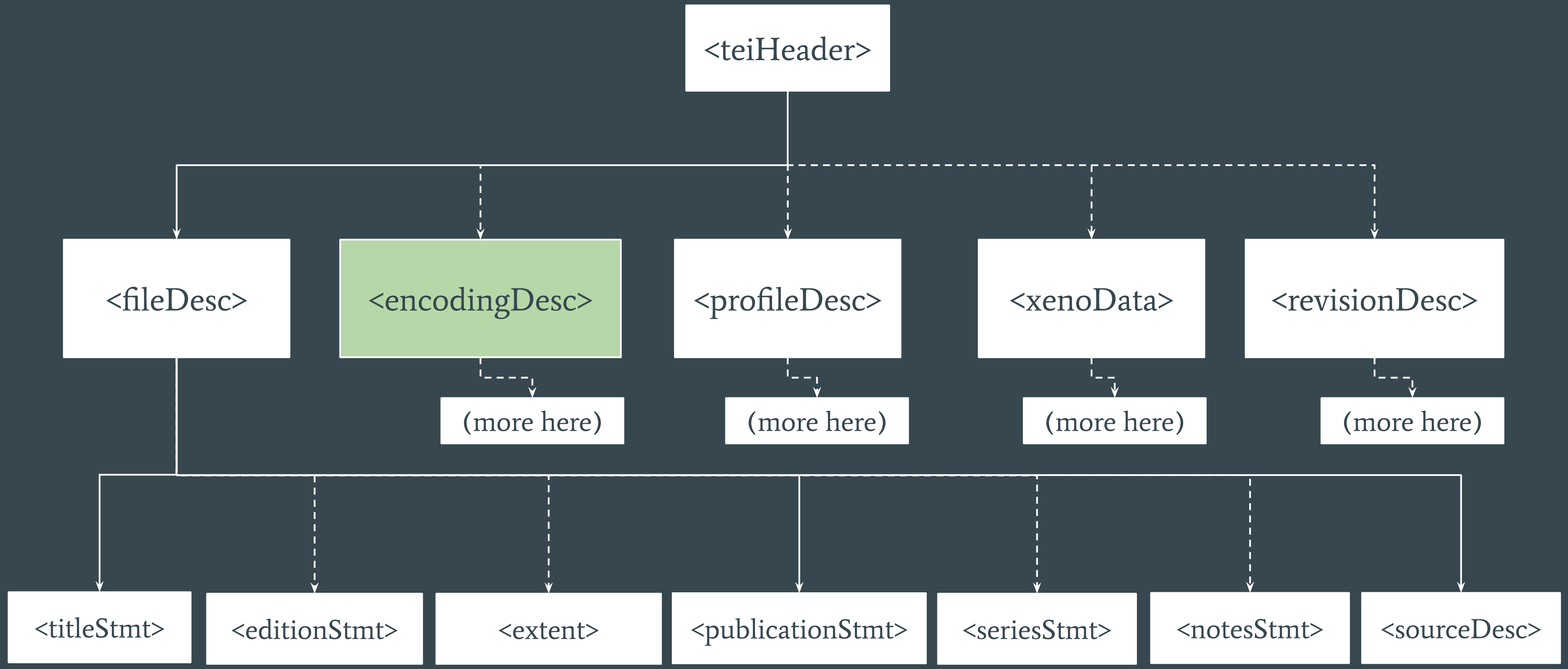
All electronic works need to document their source, even 'born digital' ones! The <sourceDesc> can have:

- prose description, just one or more <p> elements
- <bibl> (bibliographic citation): contains free text and/or any mixture of bibliographic elements such as <author>, <publisher> etc.
- <biblStruct> (structured) contains similar elements but constrained in various ways according to bibliographic standards
- A <listBibl> may be used for lists of such descriptions, e.g. bibliographies
- Specialised elements for spoken texts (<recordingStmt> etc.) and for manuscripts (<msDesc>)
- Authority lists: <listPerson>, <listPlace>, <listOrg> if not storing elsewhere



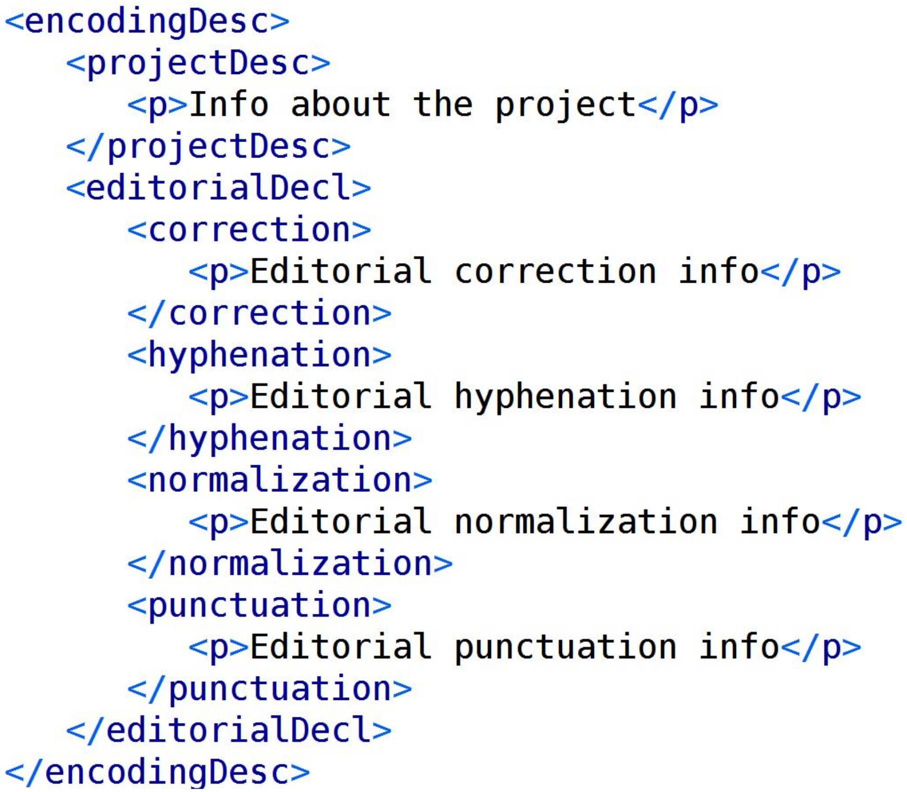
```
<sourceDesc>
   <bibl>
      <title level="a">The Interesting story of the Children in the
         Wood</title>, in <author>Victor E Neuberg</author>,
      <title>The Penny Histories</title>. <publisher>OUP</publisher>
      <date>1968</date>.
   </bibl>
</sourceDesc>

<sourceDesc>
   <p>Born digital: no previous source exists.</p>
</sourceDesc>
```

What is the structure of a <teiHeader>?



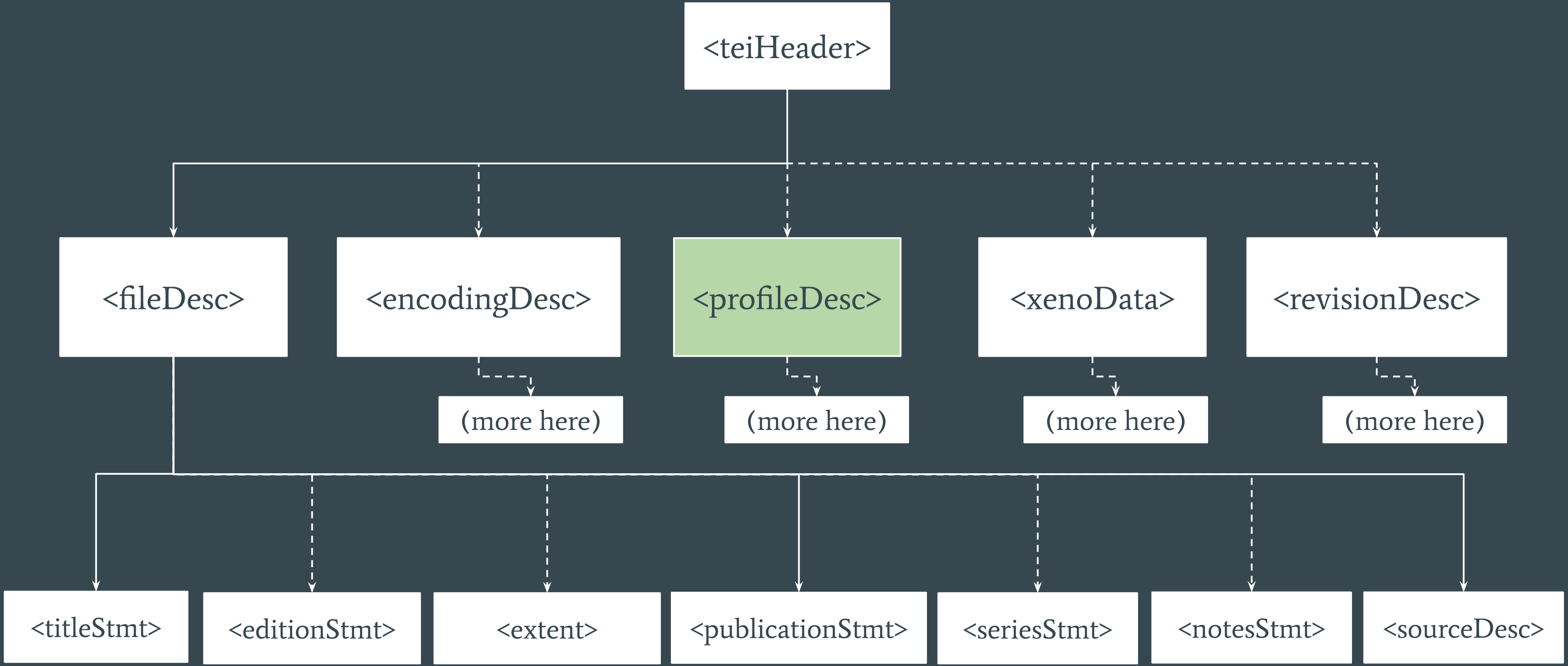
<encodingDesc>

- Groups notes about the procedures used when the text was encoded in <p> or specific elements such as:
- <projectDesc>: goals of the project
- <samplingDecl>: sampling principles
- <editorialDecl>: editorial principals,
  - e.g. <correction>, <hyphenation>, <interpretation>, <normalization>, <punctuation>, <quotation>, <segmentation>
- <classDecl>: classification system/s
- <tagsDecl>: specifics about element usage or rendition



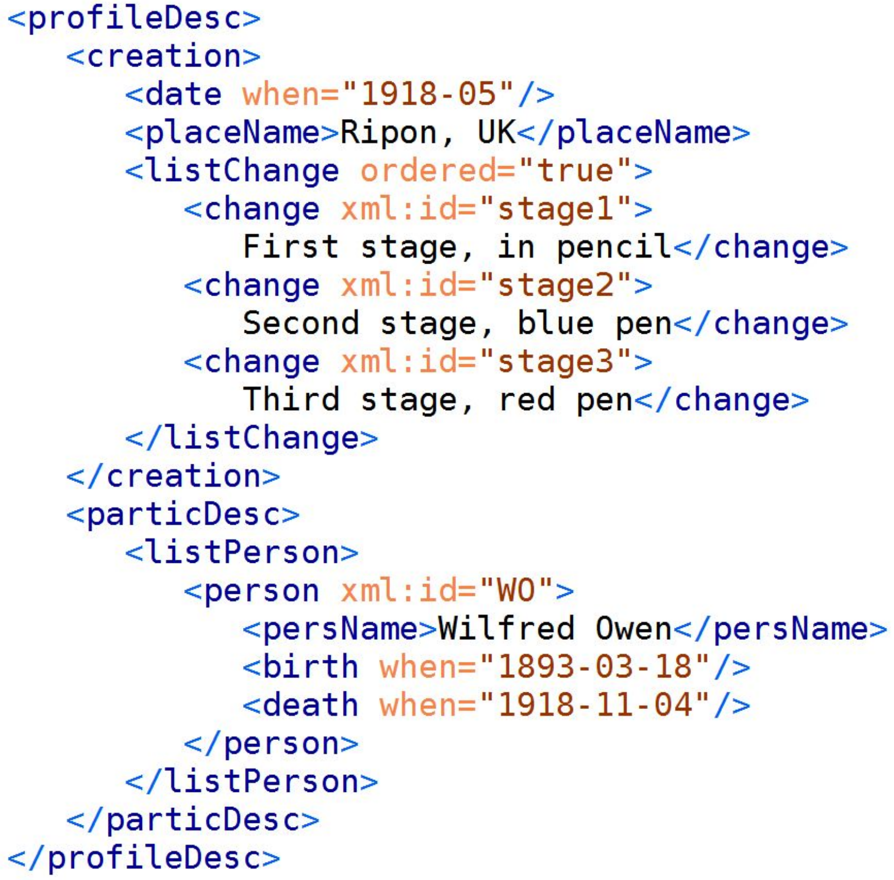
```
<encodingDesc>
   <projectDesc>
      <p>Info about the project</p>
   </projectDesc>
   <editorialDecl>
      <correction>
         <p>Editorial correction info</p>
      </correction>
      <hyphenation>
         <p>Editorial hyphenation info</p>
      </hyphenation>
      <normalization>
         <p>Editorial normalization info</p>
      </normalization>
      <punctuation>
         <p>Editorial punctuation info</p>
      </punctuation>
   </editorialDecl>
</encodingDesc>
```

What is the structure of a <teiHeader>?



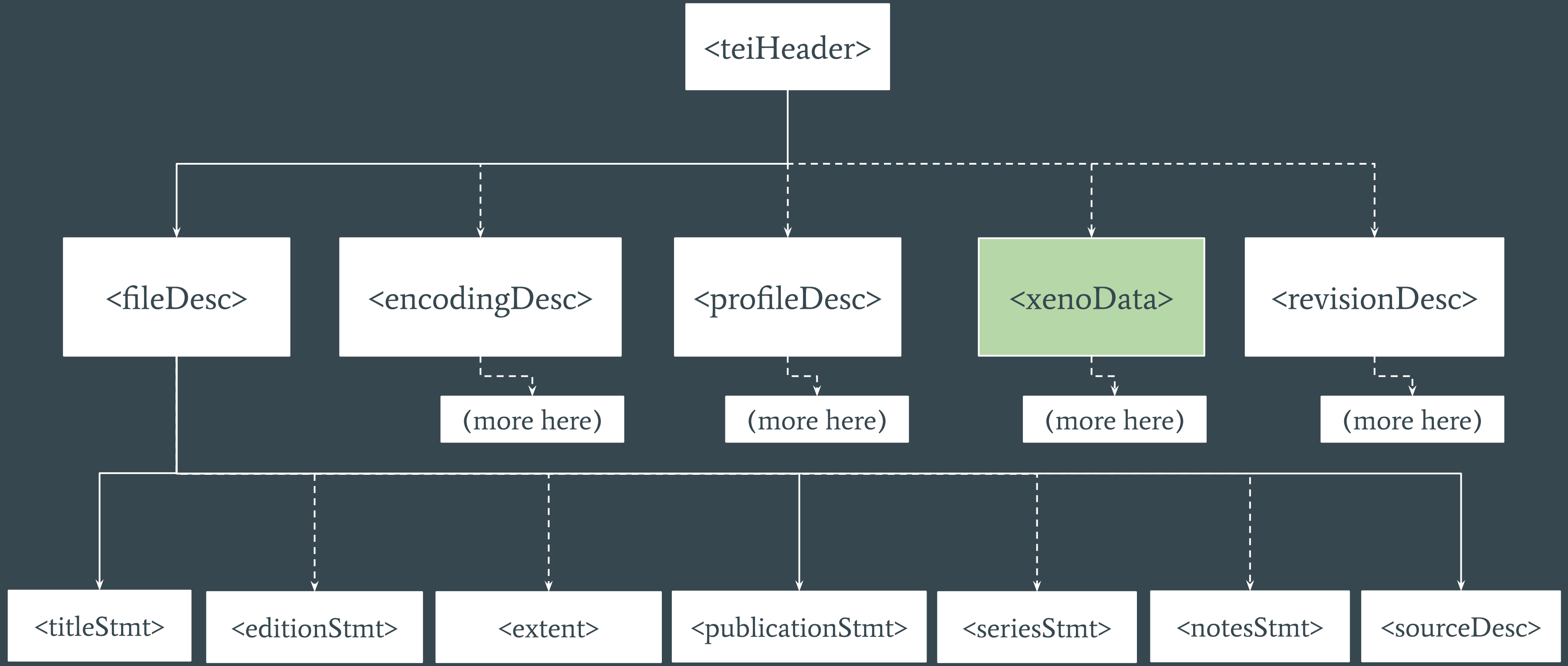
<profileDesc>

- <creation>: the creation of the text
- <langUsage>: languages, registers, writing systems
- <textDesc> and <textClass>: classifications applied to the text
- <particDesc> and <settingDesc>: details of ‘participants’, either real or depicted
- <handNotes>: hands distinguished within a manuscript when not giving full manuscript description



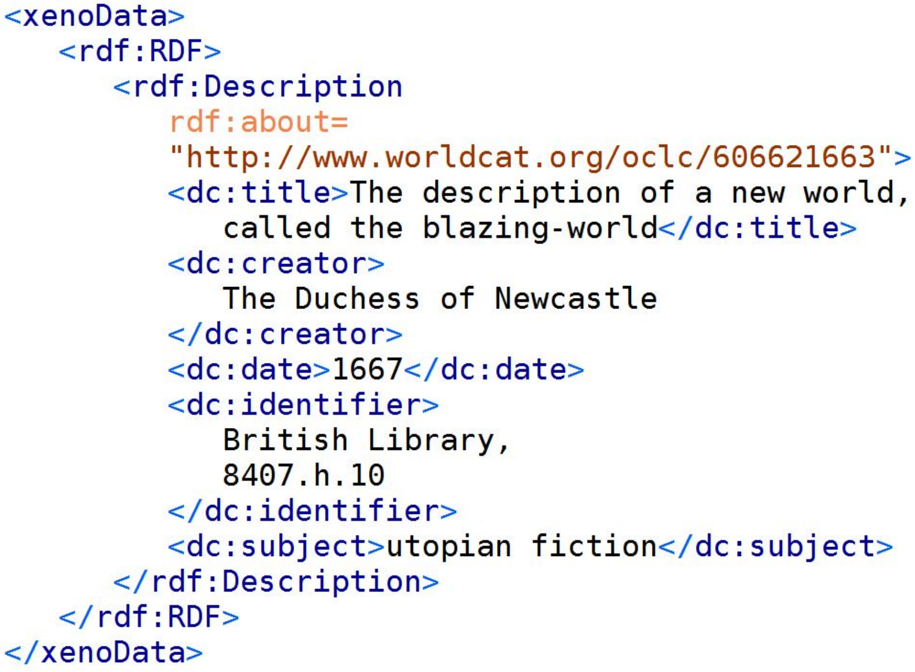
```
<profileDesc>
   <creation>
      <date when="1918-05"/>
      <placeName>Ripon, UK</placeName>
      <listChange ordered="true">
         <change xml:id="stage1">
            First stage, in pencil</change>
         <change xml:id="stage2">
            Second stage, blue pen</change>
         <change xml:id="stage3">
            Third stage, red pen</change>
      </listChange>
   </creation>
   <particDesc>
      <listPerson>
         <person xml:id="WO">
            <persName>Wilfred Owen</persName>
            <birth when="1893-03-18"/>
            <death when="1918-11-04"/>
         </person>
      </listPerson>
   </particDesc>
</profileDesc>
```

What is the structure of a <teiHeader>?



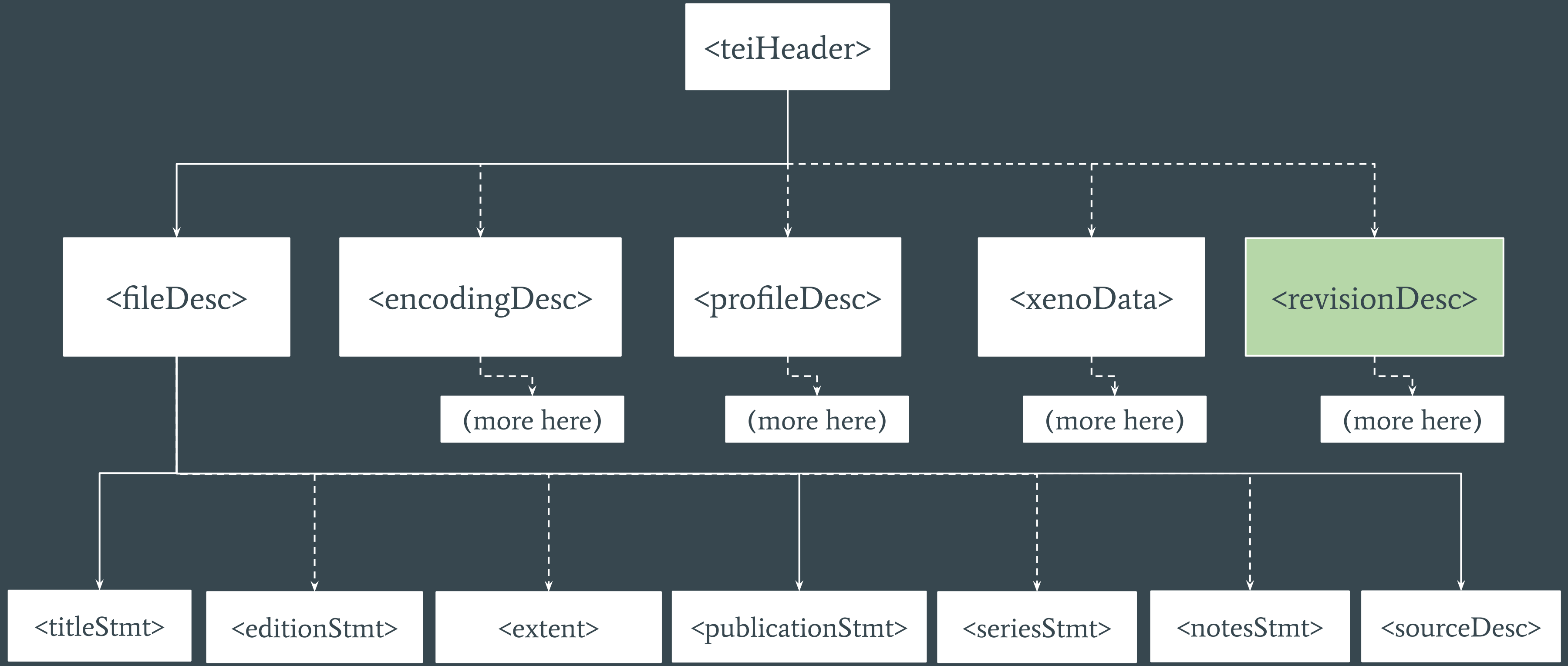
<xenoData>

- Provides a container element for non-TEI metadata in any format
- In many places these might be better stored in proper TEI locations but for convenience in processing are duplicated (or only) stored here
- One can also embed non-TEI XML anywhere in a TEI document but will have to modify the TEI customisation to enable this



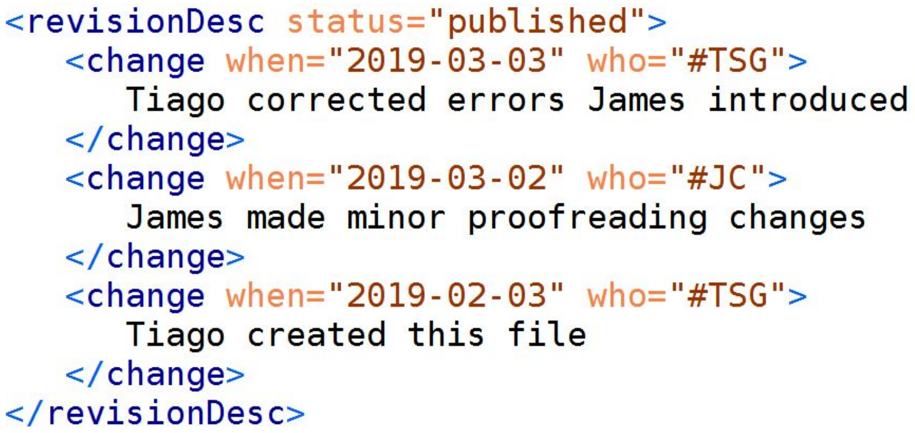
```
<xenoData>
   <rdf:RDF>
      <rdf:Description
         rdf:about=
         "http://www.worldcat.org/oclc/606621663">
         <dc:title>The description of a new world,
            called the blazing-world</dc:title>
         <dc:creator>
            The Duchess of Newcastle
         </dc:creator>
         <dc:date>1667</dc:date>
         <dc:identifier>
            British Library,
            8407.h.10
         </dc:identifier>
         <dc:subject>utopian fiction</dc:subject>
      </rdf:Description>
   </rdf:RDF>
</xenoData>
```

What is the structure of a <teiHeader>?



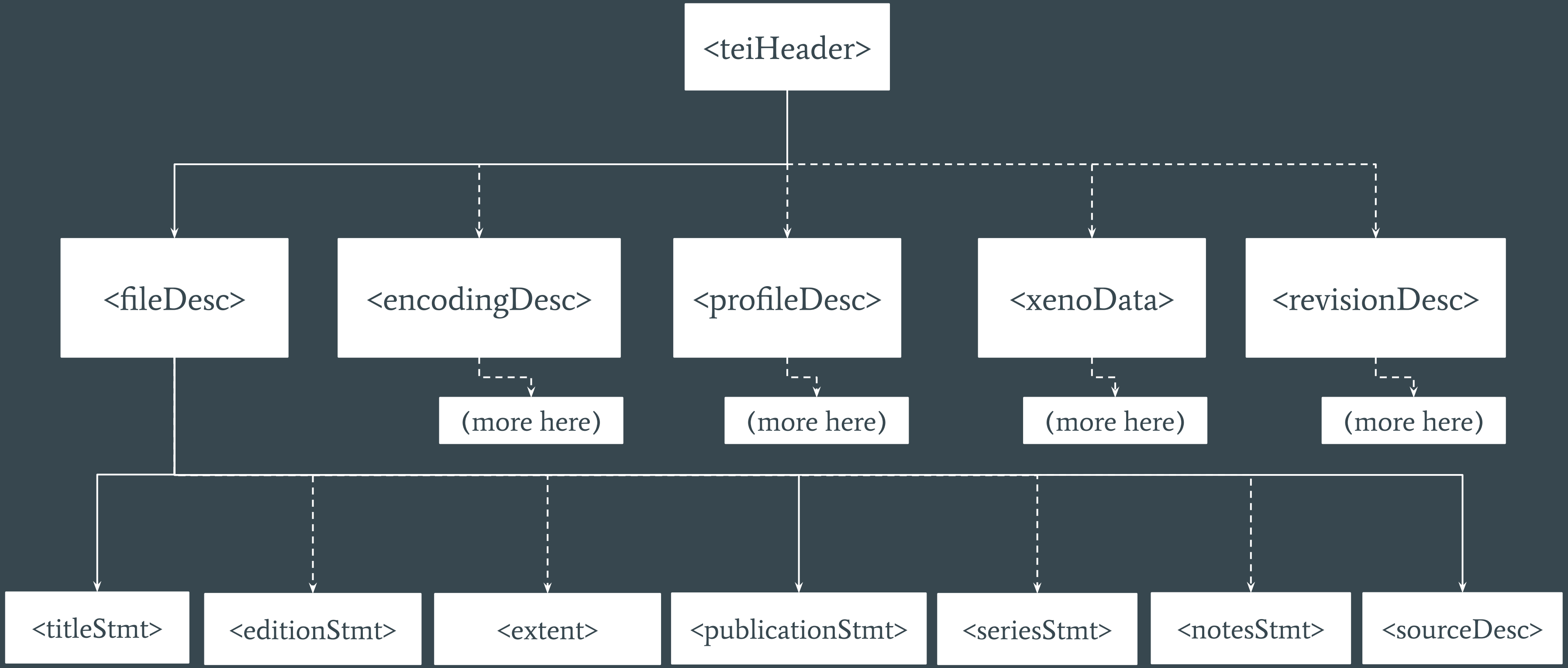
<revsionDesc>

- Contains list of <change> elements with @date and @who attributes documenting significant stages in the history of the digital document
- Conventionally, the most recent change is given first.
- Can use a <listChange>
- Can be maintained manually, or could done by means of a version control system (like Git)



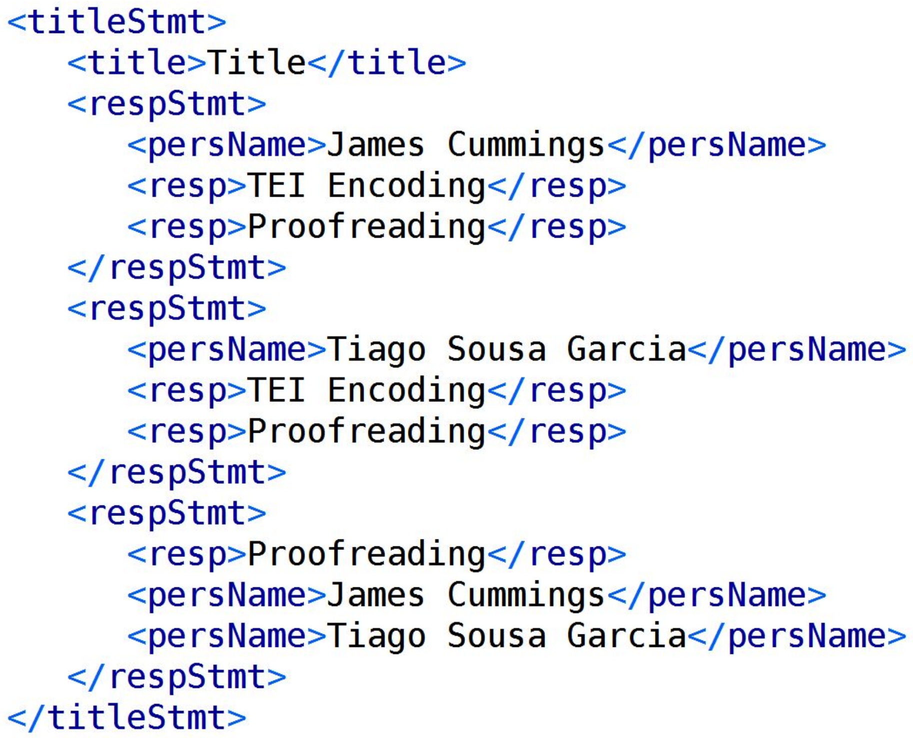
```xml
<revisionDesc status="published">
   <change when="2019-03-03" who="#TSG">
      Tiago corrected errors James introduced
   </change>
   <change when="2019-03-02" who="#JC">
      James made minor proofreading changes
   </change>
   <change when="2019-02-03" who="#TSG">
      Tiago created this file
   </change>
</revisionDesc>
```

Is that everything you can do in a <teiHeader>?



Some common <teiHeader> tasks: responsibility

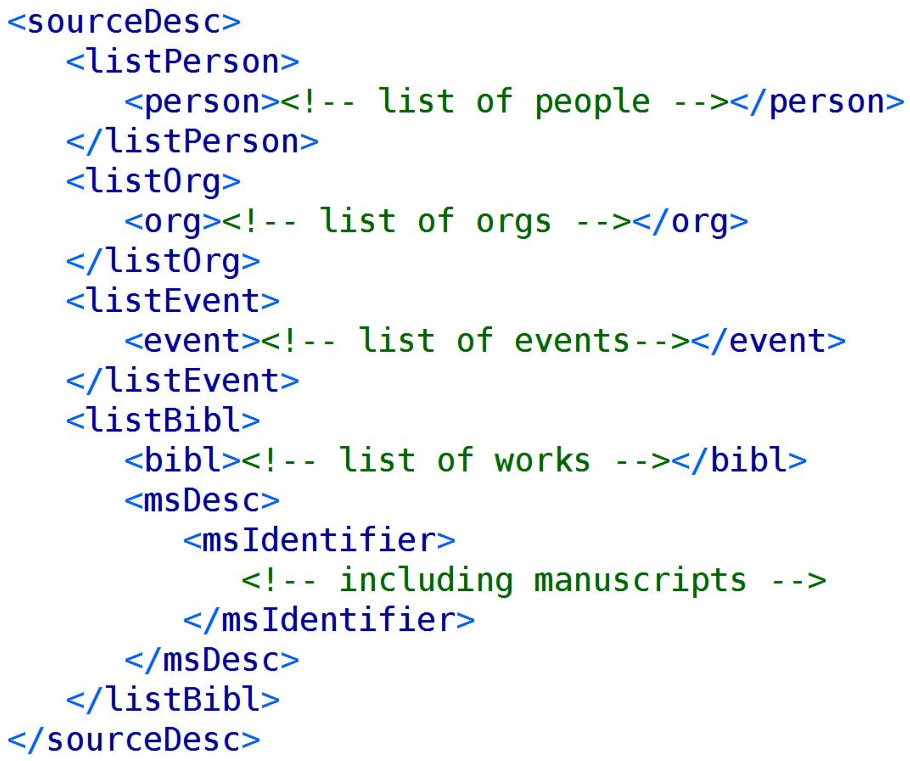
- TEI gives a general purpose way to give acknowledgement on anything in addition to <author>, <editor>, <funder>, <principal>, etc.
- One can either give a <respStmt> per individual, or group individuals under specific tasks



```
<titleStmt>
   <title>Title</title>
   <respStmt>
      <persName>James Cummings</persName>
      <resp>TEI Encoding</resp>
      <resp>Proofreading</resp>
   </respStmt>
   <respStmt>
      <persName>Tiago Sousa Garcia</persName>
      <resp>TEI Encoding</resp>
      <resp>Proofreading</resp>
   </respStmt>
   <respStmt>
      <resp>Proofreading</resp>
      <persName>James Cummings</persName>
      <persName>Tiago Sousa Garcia</persName>
   </respStmt>
</titleStmt>
```

Some common <teiHeader> tasks: named entities

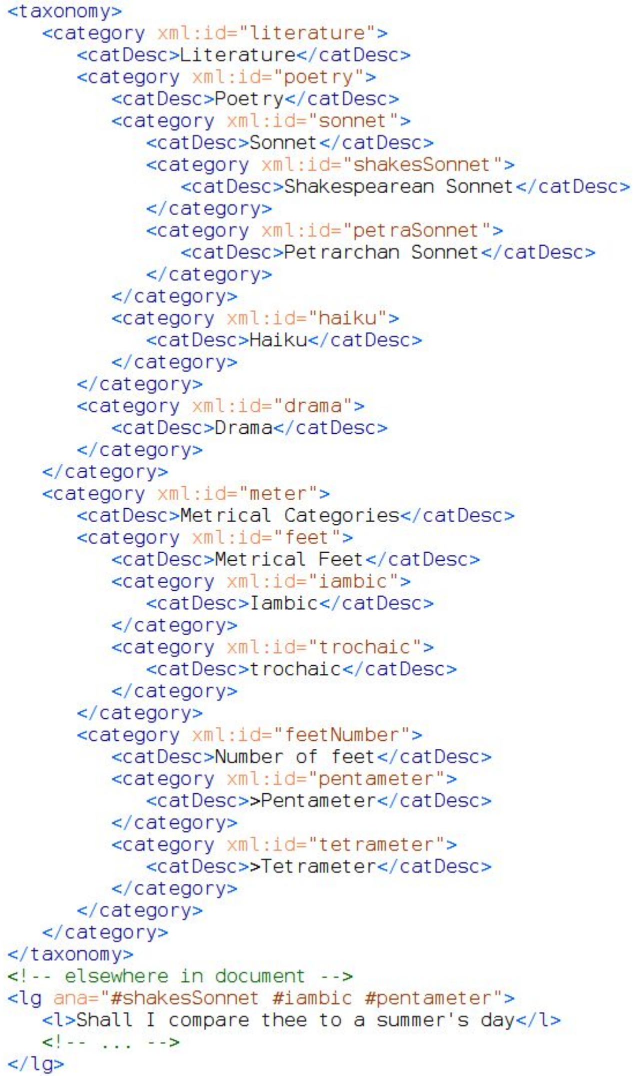
- One can store lists of <person> and <place> elements in <particDesc> and <settingDesc> in <profileDesc>
- Or where using as authority lists some projects store these in <sourceDesc>
- One can also store bibliographic lists of related works, detailed manuscript or object descriptions here



```
<sourceDesc>
   <listPerson>
      <person><!-- list of people --></person>
   </listPerson>
   <listOrg>
      <org><!-- list of orgs --></org>
   </listOrg>
   <listEvent>
      <event><!-- list of events--></event>
   </listEvent>
   <listBibl>
      <bibl><!-- list of works --></bibl>
      <msDesc>
         <msIdentifier>
            <!-- including manuscripts -->
         </msIdentifier>
      </msDesc>
   </listBibl>
</sourceDesc>
```

Some common <teiHeader> tasks: classification

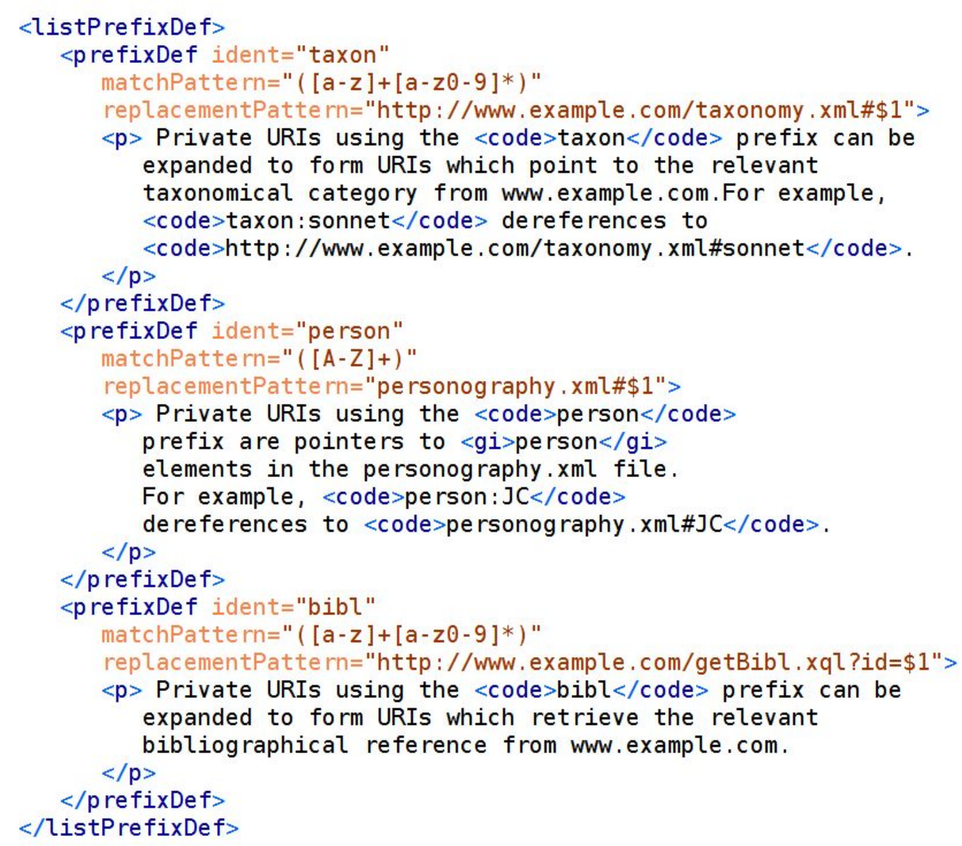
- Many elements have a built-in @type and @subtype attributes
- For more fine-grained classification use <category> elements in: encodingDesc/classDecl/taxonomy
- These <taxonomy> elements may have any number of depth of <category> elements each with a <catDesc> describing it
- You point to categories using @ana from any element in the document



```
<taxonomy>
   <category xml:id="literature">
      <catDesc>Literature</catDesc>
      <category xml:id="poetry">
         <catDesc>Poetry</catDesc>
         <category xml:id="sonnet">
            <catDesc>Sonnet</catDesc>
            <category xml:id="shakesSonnet">
               <catDesc>Shakespearean Sonnet</catDesc>
            </category>
            <category xml:id="petraSonnet">
               <catDesc>Petrarchan Sonnet</catDesc>
            </category>
         </category>
         <category xml:id="haiku">
            <catDesc>Haiku</catDesc>
         </category>
      </category>
      <category xml:id="drama">
         <catDesc>Drama</catDesc>
      </category>
   </category>
   <category xml:id="meter">
      <catDesc>Metrical Categories</catDesc>
      <category xml:id="feet">
         <catDesc>Metrical Feet</catDesc>
         <category xml:id="iambic">
            <catDesc>Iambic</catDesc>
         </category>
         <category xml:id="trochaic">
            <catDesc>trochaic</catDesc>
         </category>
      </category>
      <category xml:id="feetNumber">
         <catDesc>Number of feet</catDesc>
         <category xml:id="pentameter">
            <catDesc>>Pentameter</catDesc>
         </category>
         <category xml:id="tetrameter">
            <catDesc>>Tetrameter</catDesc>
         </category>
      </category>
   </category>
</taxonomy>
<!-- elsewhere in document -->
<lg ana="#shakesSonnet #iambic #pentameter">
   <l>Shall I compare thee to a summer's day</l>
   <!-- ... -->
</lg>
```

Some common <teiHeader> tasks: prefix definitions

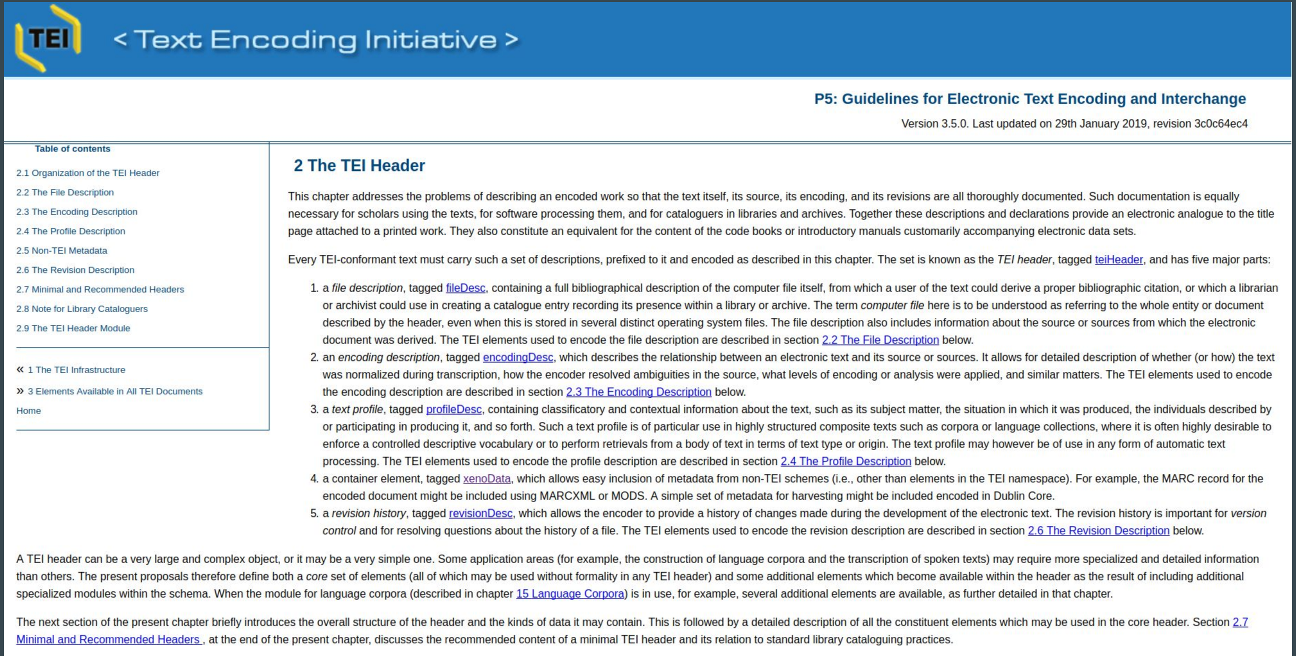
- If pointing out of the document all the time, you can use <listPrefixDef> to define a ‘private URI syntax’, basically a shortcode to make the URL easier to type
- So instead of having to type http://www.example.com/taxonomy.xml#sonnet to point to a sonnet, one could type ‘taxon:sonnet’ or similar



```
<listPrefixDef>
   <prefixDef ident="taxon"
      matchPattern="([a-z]+[a-z0-9]*)"
      replacementPattern="http://www.example.com/taxonomy.xml#$1">
      <p> Private URIs using the <code>taxon</code> prefix can be
         expanded to form URIs which point to the relevant
         taxonomical category from www.example.com.For example,
         <code>taxon:sonnet</code> dereferences to
         <code>http://www.example.com/taxonomy.xml#sonnet</code>.
      </p>
   </prefixDef>
   <prefixDef ident="person"
      matchPattern="([A-Z]+)"
      replacementPattern="personography.xml#$1">
      <p> Private URIs using the <code>person</code>
         prefix are pointers to <gi>person</gi>
         elements in the personography.xml file.
         For example, <code>person:JC</code>
         dereferences to <code>personography.xml#JC</code>.
      </p>
   </prefixDef>
   <prefixDef ident="bibl"
      matchPattern="([a-z]+[a-z0-9]*)"
      replacementPattern="http://www.example.com/getBibl.xql?id=$1">
      <p> Private URIs using the <code>bibl</code> prefix can be
         expanded to form URIs which retrieve the relevant
         bibliographical reference from www.example.com.
      </p>
   </prefixDef>
</listPrefixDef>
```

I want to know more!

- Chapter 2: The TEI Header



P5: Guidelines for Electronic Text Encoding and Interchange

Version 3.5.0. Last updated on 29th January 2019, revision 3c0c64ec4

Table of contents

« 1 The TEI Infrastructure
» 3 Elements Available in All TEI Documents
Home

2 The TEI Header

This chapter addresses the problems of describing an encoded work so that the text itself, its source, its encoding, and its revisions are all thoroughly documented. Such documentation is equally necessary for scholars using the texts, for software processing them, and for cataloguers in libraries and archives. Together these descriptions and declarations provide an electronic analogue to the title page attached to a printed work. They also constitute an equivalent for the content of the code books or introductory manuals customarily accompanying electronic data sets.

Every TEI-conformant text must carry such a set of descriptions, prefixed to it and encoded as described in this chapter. The set is known as the *TEI header*, tagged teiHeader, and has five major parts:

1. a *file description*, tagged fileDesc, containing a full bibliographical description of the computer file itself, from which a user of the text could derive a proper bibliographic citation, or which a librarian or archivist could use in creating a catalogue entry recording its presence within a library or archive. The term *computer file* here is to be understood as referring to the whole entity or document described by the header, even when this is stored in several distinct operating system files. The file description also includes information about the source or sources from which the electronic document was derived. The TEI elements used to encode the file description are described in section 2.2 The File Description below.
2. an *encoding description*, tagged encodingDesc, which 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. The TEI elements used to encode the encoding description are described in section 2.3 The Encoding Description below.
3. a *text profile*, tagged profileDesc, containing 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. The TEI elements used to encode the profile description are described in section 2.4 The Profile Description below.
4. a container element, tagged xenoData, which allows easy inclusion of metadata from non-TEI schemes (i.e., other than elements in the TEI namespace). For example, the MARC record for the encoded document might be included using MARCXML or MODS. A simple set of metadata for harvesting might be included encoded in Dublin Core.
5. a *revision history*, tagged revisionDesc, which 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. The TEI elements used to encode the revision description are described in section 2.6 The Revision Description below.

A TEI header can be a very large and complex object, or it may be a very simple one. Some application areas (for example, the construction of language corpora and the transcription of spoken texts) may require more specialized and detailed information than others. The present proposals therefore define both a *core* set of elements (all of which may be used without formality in any TEI header) and some additional elements which become available within the header as the result of including additional specialized modules within the schema. When the module for language corpora (described in chapter 15 Language Corpora) is in use, for example, several additional elements are available, as further detailed in that chapter.

The next section of the present chapter briefly introduces the overall structure of the header and the kinds of data it may contain. This is followed by a detailed description of all the constituent elements which may be used in the core header. Section 2.7 Minimal and Recommended Headers , at the end of the present chapter, discusses the recommended content of a minimal TEI header and its relation to standard library cataloguing practices.