Text Encoding and Semantic Representation

XML / TEI (1)

marilena.daquino2@unibo.it | https://github.com/marilenadaquino/tesr_dhdk

Outline

Exercise review

Introduction to TEI

Validation

Hands-on

Work at home

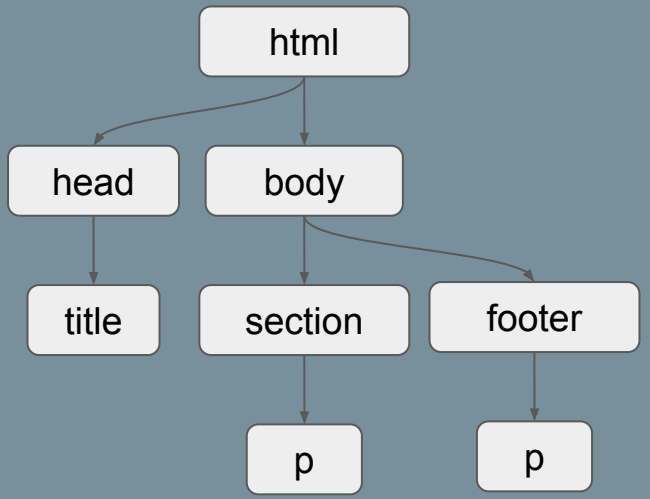
NOW… Change the file extension to .html and open the new file in a browser

Practice markup

Given the following tree, create a XML document. This time, do not add the prologue!

Fill elements with some text when appropriate.

Tip: open the xml file of both the exercise and the homework in a browser (e.g. Chrome)



TEI

In pills

Stands for **Text Encoding Initiative**.

It is both a schema to annotate literary texts and a community of scholars (mainly philologists). The website includes a set of **guidelines** on how to use the XSD by thematic areas and examples.

The first version of the guidelines was released in 1990 (currently v4.6). The reference guidelines are called P5 Guidelines.

TEI

Modules

The schema is divided in (thematic) modules, i.e. groupings of XML elements and attributes (~500 elements in total).

Many modules can be used in the same document.

The **core module** includes those elements that are likely to be used to describe any XML document.

Module name	Formal public identifier
analysis	Analysis and Interpretation
certainty	Certainty and Uncertainty
core	Common Core
corpus	Metadata for Language Corpora
dictionaries	Print Dictionaries
drama	Performance Texts
figures	Tables, Formulae, Figures
gaiji	Character and Glyph Documentation
header	Common Metadata
iso-fs	Feature Structures
linking	Linking, Segmentation, and Alignment
msdescription	Manuscript Description
namesdates	Names, Dates, People, and Places
nets	Graphs, Networks, and Trees
spoken	Transcribed Speech
tagdocs	Documentation Elements
tei	TEI Infrastructure
textcrit	Text Criticism
textstructure	Default Text Structure
transcr	Transcription of Primary Sources
verse	Verse

TEI

Global attributes

Likewise, there are a number of **global attributes**, meaning they can be used with any element (while non-global attributes can be used only with specific sets of attributes)

@xml:id	(identifier) provides a unique identifier for the element bearing the attribute.
@n	(number) gives a number (or other label) for an element, which is not necessarily unique within the document.
@xml:lang	(language) indicates the language of the element content using a 'tag' generated according to BCP 47.
@rend [att.global.rendition]	(rendition) indicates how the element in question was rendered or presented in the source text.
@style [att.global.rendition]	contains an expression in some formal style definition language which defines the rendering or presentation used for this element in the source text
@rendition [att.global.rendition]	points to a description of the rendering or presentation used for this element in the source text.
@xml:base	provides a base URI reference with which applications can resolve relative URI references into absolute URI references.
@xml:space	signals an intention about how white space should be managed by applications.
@source [att.global.source]	specifies the source from which some aspect of this element is drawn.
@cert [att.global.responsibility]	(certainty) signifies the degree of certainty associated with the intervention or interpretation.
@resp [att.global.responsibility]	(responsible party) indicates the agency responsible for the intervention or interpretation, for example an editor or transcriber.

TEI

Mandatory elements

The root element **TEI** includes the namespace declaration

The element **teiHeader** includes all the metadata about the digital edition of the text.

The element **text** includes the actual text.

TEI

Mandatory elements

When annotating multiple texts, the root element is **teiCorpus**, with a bespoke **teiHeader**, and there are as many **TEI** elements as the number of documents to be annotated.

```
<teiCorpus xmlns="http://www.tei-c.org/ns/1.0">
 <teiHeader>
 <!-- corpus-level metadata here -->
 </teiHeader>
 <TEI>
  <teiHeader>
  <!-- metadata specific to this text here -->
  </teiHeader>
  <text><!-- ... --></text>
 </TEI>
 <TEI>
  <teiHeader>
<!-- metadata specific to this text here -->
  </teiHeader>
  <text><!-- ... --></text>
 </TEI>
</teiCorpus>
```

Disclaimer

Many solutions to the same problem create a new problem

Other than mandatory elements, in most cases there is not just one way to annotate the text, but several variants are possible (e.g. you may wrap sentences in an element or you can annotate substrings in multiple elements). Unfortunately, such a variety is not that good for exchange purposes :(

Disclaimer

500 elements are a lot…

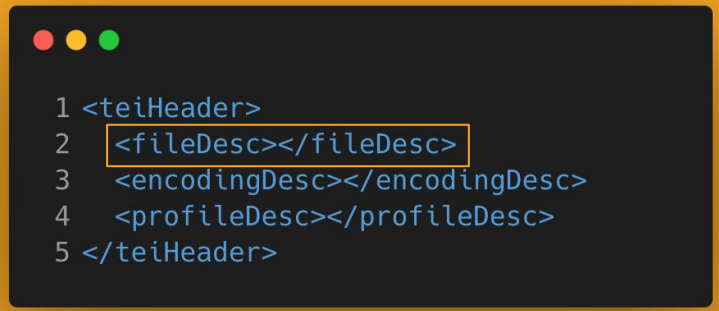
We are not going to see the full elementset, just some of the most important ones. You are free to study more elements according to your (encoding) needs.

TEI

teiHeader

teiHeader includes the metadata of the digital edition, such as people's responsibilities, the bibliographic citation, and the strategies used to encode the text (e.g. naming conventions, critical choices in the transcription).

Only **fileDesc** is mandatory.



TEI

fileDesc

fileDesc (mandatory) describes the current electronic file, with the exception of sourceDesc, which describes the original source.

titleStmt, **publicationStmt** and **sourceDesc** are **mandatory**.

Guidelines

```
<teiHeader>
 <fileDesc>
  <titleStmt>
   <title><!-- title of the resource --></title>
  </titleStmt>
  <editionStmt>
   <p><!-- the edition of the resource  --></p>
  </editionStmt>
  <extent><!-- the size of the resource --></extent>
  <publicationStmt>
   <p><!-- the distribution of the resource --></p>
  </publicationStmt>
  <seriesStmt>
   <p><!-- any series to which the resource belongs  --></p>
  </seriesStmt>
  <notesStmt>
   <note><!-- other aspects of the resource --></note>
  </notesStmt>
  <sourceDesc>
   <p><!-- the source from which the resource was derived  --></p>
  </sourceDesc>
 </fileDesc>
</teiHeader>
```

TEI

titleStmt

title includes the title of the digital edition (derived from the original title)

It can include also the **author** of the original work and all the people that contributed to the digital edition - each recorded in **respStmt**, including the role (**resp**) and the **name** of the person.

Guidelines

```
<titleStmt>
 <title>Resistance: digital edition</title>
 <author>Muse</author>
 <respStmt>
  <resp>compiled by</resp>
  <name>Marilena Daquino</name>
 </respStmt>
</titleStmt>
```

TEI

editionStmt

editionStmt includes information about the current edition of the digital text (not the edition of the original text).

edition includes a string and/or children elements. **respStmt** can be used to record people and roles.

Guidelines

```
<editionStmt>
 <edition n="1">Private edition, <date>November 2023</date></edition>
 <respStmt>
  <resp>Annotations by</resp>
  <name>Marilena Daquino</name>
 </respStmt>
</editionStmt>
```

TEI

extent

Describes the extent of the digital file, e.g. in terms of MB, GB or other units of measures.

Notice the usage of the element measure and its attributes.

Guidelines

```
<!-- a string -->
<extent>between 1 and 2 Mb</extent>

<!-- or one or more measurements -->
<extent>
 <measure unit="MiB" quantity="2">About 2 megabytes</measure>
 <measure unit="pages" quantity="1">1 page of source material</measure>
</extent>
```

TEI

publicationStmt

publicationStmt includes information about the publisher of the digital edition and the rights/licenses under which the resource is available.

NB. The digital edition of a text that is under copyright cannot be open access!

Guidelines

```xml
<publicationStmt>
 <publisher>University of Bologna</publisher>
 <pubPlace>Bologna</pubPlace>
 <date>2023</date>
 <idno type="DOI">10.xxxx</idno>
 <availability>
  <p>Open access</p>
 </availability>
  <licence target="https://creativecommons.org/licenses/by/4.0/">CC-BY</licence>
</publicationStmt>
```

TEI

sourceDesc

sourceDesc includes the bibliographic reference of the original source.

Details can include authors, title, dates, publisher, etc.

Guidelines

```
<sourceDesc>
  <bibl>
    <title level="a">Resistance</title>. In
    <author>Muse</author>,
    <title level="b">Resistance</title>.
    <date>2009</date>.
  </bibl>
</sourceDesc>
```

TEI

encodingDesc

It specifies the methods and **editorial principles** which guided the transcription or encoding of the text, such as corrections (if any error was corrected), normalisation strategies (e.g. en-us spelling), or interpretations (e.g. if you add elements of linguistic analysis).

```
<encodingDesc>
  <projectDesc>
    <p><!-- The purpose of the digital resource --></p>
  </projectDesc>
  <editorialDecl>
    <correction>
      <p><!-- If the text includes corrections--></p>
    </correction>
    <normalization>
      <p><!-- If any changes happened to uniform ecitorial choices --></p>
    </normalization>
    <interpretation><!-- If any analytical feature was added to the text--></interpretation>
  </editorialDecl>
</encodingDesc>
```

Exercise

Create a XML/TEI file

Create a XML/TEI file to describe the following song:

Bohemian Rhapsody, Queen, 1975. https://genius.com/Queen-bohemian-rhapsody-lyrics

Include the element **teiHeader** and all the mandatory children we have seen.

TEI

The transcription of the text

The transcribed text is included in the element <text>. It may include a front matter <front>, a text body <body>, and a back matter <back>.

Guidelines

```xml
<TEI xmlns="http://www.tei-c.org/ns/1.0">
 <teiHeader>
    <!-- ... -->
 </teiHeader>
 <text>
  <front>
    <!-- front matter of copy text, if any, goes here -->
  </front>
  <body>
    <!-- body of copy text goes here -->
  </body>
  <back>
    <!-- back matter of copy text, if any, goes here -->
  </back>
 </text>
</TEI>
```

TEI

Elements in all TEI documents

Some elements can be used in any TEI document regardless of the type/genre of literary text. These include: paragraphs <p>, quotations <q>, graphically highlighted terms: <hi>, foreign terms: <foreign>, emphasis <emph>, etc…

Guidelines

```
<p>This is a paragraph mentioning
  a <foreign xml:lang="fr">croissant</foreign>,
  a <high rend="italic">highlight</high>,
  and some <emph>important concept</emph>.
</p>

<q>It is followed by a quotation from one of
your favourite authors</q>
```

TEI

Divisions of the text

A prose text can be divided in several structures, e.g. parts, chapters, sections. In this case divisions are represented with the element and the attributes @type and @num further specify their peculiarities and differentiate them. Values are free from naming conventions.

```
<body>
 <div type="part" n="1">
    <div type="chapter" n="1">
        <!-- text of part 1, chapter 1 -->
    </div>
    <div type="chapter" n="2">
        <!-- text of part 1, chapter 2 -->
    </div>
 </div>
 <div type="part" n="2">
  ...
 </div>
</body>
```

TEI

Headings

Inside a division, titles of all levels can be represented with <head> and further specified via the attributes @rend, @type and more

```
<div n="19" type="chap">
 <head rend="bold" type="main">Chapter 19</head>
 <p>To say that Deronda was romantic would be to
   misrepresent him: but under his calm and
   somewhat self-repressed exterior ...</p>
</div>
```

TEI

Lists

Bullet lists are represented via the element <list> and its children <item>. Lists can be ordered or unordered. Items can include other lists.

```
<list rend="numbered">
 <item>a butcher</item>
 <item>a baker</item>
 <item>a candlestick maker, with
 <list rend="bulleted">
   <item>rings on his
fingers</item><item>bells on his toes</item>
  </list>
 </item>
</list>
```

Exercise

Modify a XML/TEI file

- Modify your XML, include the element **text** and add the transcription of the lyrics.

TEI

Deletions, corrections

Omissions, changes (deletions and additions), as well as unclear texts are recorded in elements like <gap> (words omitted), <del> and <add>

Guidelines

```
<gap reason="illegible" unit="word" quantity="2"
resp="#editor04"/>
<!-- notice the attribute resp -->

<l>
 <del rend="overstrike">Inviolable</del>
 <add place="below">Inexplicable</add>
 splendour of Corinthian white and gold
</l>
```

TEI

Letters

Letters include peculiar logical elements such as <opener>, <closer> and <postscript>.
Line breaks in a paragraph can be recorded via <lb>, a milestone element written in the short version.
In the example, also notice the element <unclear> to mark words difficult to read.

Guidelines

```xml
<div type="letter">
 <opener>
  <dateline>
   <placeName>Newport</placeName>
   <date when="1761-05-27">May ye 27th 1761</date>
  </dateline>
  <salute>Gentlemen</salute>
 </opener>
 <p>Capt Stoddard's Business
 <lb/>calling him to Providence, have
 <lb/>got him to look at Hopkins brigantine
 <lb/>&amp;amp; if can agree to Purchase her, shall
 <lb/>be much oblig'd for your further
 <lb/>assistance herein, &amp;amp; will acquiesce with
 <lb/>whatever you &amp;amp; he shall Contract
 <lb/>for — I Thank you for your
 <lb/>
  <unclear>Line</unclear> respecting the brigantine &amp;amp;
Beg<lb/>leave to Recommend the Bearer
 <lb/>to you for your advice &amp;amp; Friendship
 <lb/>in this matter</p>
 <closer>
  <salute>I am your most humble servant</salute>
  <signed>Joseph Wanton Jr</signed>
 </closer>
 <postscript>
  <label>P.S.</label>
  <p>I have Mollases, Sugar,
  <lb/>Coffee &amp;amp; Rum, which
  <lb/>will Exchange with you
  <lb/>for Candles or Oyl</p>
 </postscript>
</div>
```

TEI

Verses

Poems can be characterised by ungrouped lines (element <l>). When grouped, lines are included in the element <lg>. The attribute @type can specify the structure or unit, e.g. a sonnet or a stanza.

Guidelines

```
<text>
 <body>
   <lg type="quatrain">
    <l>My Mistres eyes are nothing like the Sunne,</l>
    <l>Currall is farre more red, then her lips red</l>
    <l>If snow be white, why then her brests are dun:</l>
    <l>If haires be wiers, black wiers grown on her head:</l>
   </lg>
    ...
   <lg type="couplet">
    <l>And yet by heaven I think my love as rare,</l>
    <l>As any she beli'd with false compare.</l>
   </lg>
 </body>
</text>
```

TEI

Performances

Performance texts may include descriptions of the stage and the sequence of speakers' dialogues.

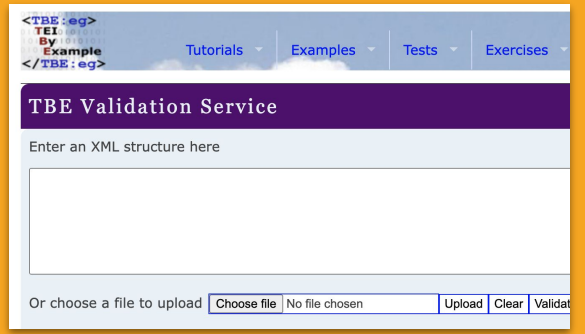
Here like in other situations, the hierarchy of text divisions can be further specified by using the elements <div1> and <div2> (not mandatory).

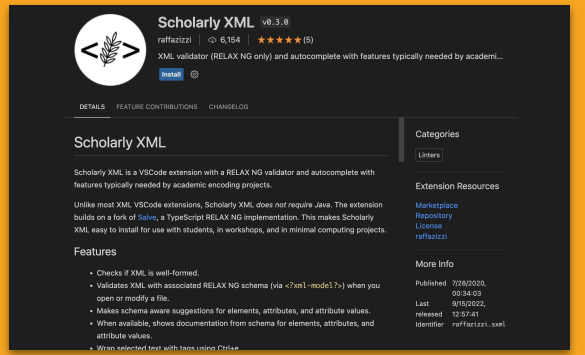
Guidelines

```
<body>
 <div1 type="act" n="1">
  <head>Act One</head>
  <div2 type="scene" n="1">
   <stage>Pa Ubu, Ma Ubu</stage>
   <sp>
    <speaker>Pa Ubu</speaker>
    <p>Pschitt!</p>
   </sp>
  </div2>
  <div2 type="scene" n="2">
   <stage>A room in Pa Ubu's house, where a magnificent
       collation is set out</stage>
  </div2>
 </div1>
 <div1 type="act" n="2">
  <head>Act Two</head>
  <div2 type="scene" n="1">
   <head>Scene One</head>
  </div2>
  <div2 type="scene" n="2">
   <head>Scene Two</head>
  </div2>
 </div1>
</body>
```

TEI

Validation





There are several ways to validate a XML/TEI document

- Online validator: tells you whether the schema is respected (valid) and the XML is well-formed
- XML/TEI plugin for VS Code: tells you whether the XML is valid, well-formed, and suggests TEI elements
- Python library: a shell script to validate XML/TEI docs

TEI

Validation w/ VSCode plugin

- Download and install the plugin for VS Code
- **Read the instructions**
- Create a XML file in VS Code
- Include the minimal structure of TEI doc
- Include this XML/TEI sample
- Modify the elements and check the validity

NB. Relax ng (.rng) is yet another way to encode a XML schema, like XSD and DTD.

```
<?xml version="1.0" encoding="UTF-8"?>
<?xml-model
  href="https://vault.tei-c.org/P5/current/xml/tei/custom/schema/relaxng/tei_all.rng"
  schematypens="http://relaxng.org/ns/structure/1.0"
  type="application/xml"?>
<TEI xmlns="http://www.tei-c.org/ns/1.0">
  ...
</TEI>
```

Homework

Quiz time!

Answer the questions and get your results immediately.
https://forms.gle/DNt9cPCxuUZzwxue8