Digital Humanities

Digital Humanities for medieval philosophical sources

4. Principles of TEI-XML

conf. dr. Mihai Maga

Babeş–Bolyai University, Cluj–Napoca
Master in Ancient and Medieval Philosophy

2nd semester, 2020–2021
HME2415/04

https://www.mihaimaga.ro/dh/

Course outline

1. The XML format

XML (eXtended Markup Language) = digital format text with metatextual markup
Only 5 types of content:

1. **tags** (contain metatextual elements)
   `<tag></tag>`
   syntax: between the signs < > and closed with /
2. **attributes** (specify parameters for tags)
   `<tag attr="val"/>`
   syntax: name (without spaces) followed by = followed by the value between quotes " "
3. **text** (text as is)
   `text`
   syntax: any characters (except <>) and which are not tags
4. **declarations** (processing commands)
   `<? decl ?>`
5. **comments** (content which is ignored on processing)
   `<!-- comm -->`

XML Example

XML

```
<text>
    This is XML text. Mark something <bold>important</bold>.
    It was written at <city country="Romania">Cluj</city>.
    <!-- here is a comment -->
    <tag>A tag may contain
        <subtag>a subtag which can also contain
            <subsubtag>sub-subtags</subsubtag>
        </subtag>
    </tag>
    A tag may have zero, <separator/> one or more attributes
        <city country="RO" department="CJ" prefix="0264">Cluj-Napoca</city>
</text>
```

Specifications

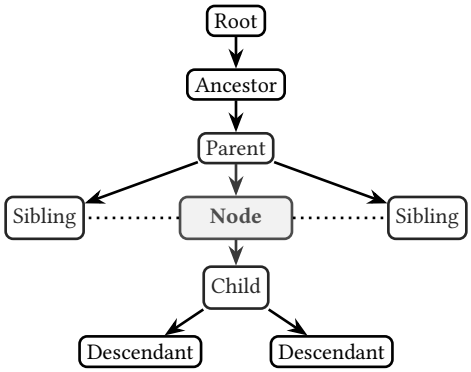
- the tags have two forms:
  - **pair**: delimit a portion of text (`<tag>text</tag>`)
    - must always be closed in reverse order of the opening (`<a><b><c>...</c></b></a>`)
    - only the opening tag may have attributes (`<tag attr="val">...</tag>`)
  - **single**: as standalone element, without text (`<tag/>`)
    - may have attributes (`<tag attr="val"/>`)
- the contents between tags may have tags inside, creating a tree
  `<root><branch><leaf/></branch><branch><leaf/></branch></root>`
- in a well-formed XML, there must be a single root tag which contains the whole document
- extra spaces and line ends are usually ignored, but they are used for easier code editing

```
<document>
    Format:
        <b>bold</b>,
        <i>italic</i>.
</document>
```

→ Format: **bold**, *italic*.

Basic concepts in tree type structures

node	any element which is part of the structure
root	element which subordinates all the other elements of the document
parent	relation between elements in which the target element has sub-elements
children	relation in which the target elements are immediately subordinated to the parent element
siblings	relation in which the target elements are on the same level and have a common parent
descendants	relation in which the target elements are inferior to a superior element
ancestors	relation in which the target elements are hierarchically superiors and connected to a descendant element



5

2. About TEI

- The Text Encoding Initiative Consortium Guidelines (TEI) establish the annotation system for the documents from the humanities
- TEI uses XML as file format
  - TEI is a subset of XML instructions to which a semantic is assigned
- in TEI are specified the XML elements used for digital editions
- the root element for a TEI-XML document is <TEI> </TEI>
- a TEI document usually has two mandatory parts:
  - a preamble <teiHeader> </teiHeader>
    - in the preamble the document properties are described: title, author, version, sources etc.
  - the body of the text <text> </text>
    - contains the text with TEI tags; the main text is contained between <body> </body>; the text paragraphs are comprised within
- all the TEI specifications are on the website: http://www.tei-c.org/

6

Example of basic TEI structure

TEI-XML

```
<?xml version="1.0"?>
<TEI xmlns="http://www.tei-c.org/ns/1.0">
    <teiHeader>
        <fileDesc>
            <titleStmt>
                <title>Titulus operae</title>
                <author>Nomen auctoris</author>
            </titleStmt>
            <sourceDesc>
                <p>Textus fictivus</p>
            </sourceDesc>
        </fileDesc>
    </teiHeader>
    <text>
        <body>
            <p xml:id="par01" lang="la">Hic est textus editionis.
            Transcriptus est in linguam programmandi qui
            nomen <ref target="http://www.tei-c.org/">TEI</ref> habet.</p>
        </body>
    </text>
</TEI>
```

7

Example of documentation from TEI Guidelines

<titleStmt>

<titleStmt> (title statement) groups information about the title of a work and those responsible for its content. [2.2.1 The Title Statement 2.2 The File Description]	
Module	header — The TEI Header
Attributes	att.global (@xml:id, @n, @xml:lang, @xml:base, @xml:space) (att.global.rendition (@rend, @style, @rendition)) (att.global.linking (@corresp, @synch, @sameAs, @copyOf, @next, @prev, @exclude, @select)) (att.global.analytic (@ana)) (att.global.facs (@facs)) (att.global.change (@change)) (att.global.responsibility (@cert, @resp)) (att.global.source (@source))
Contained by	**header:** biblFull fileDesc
May contain	**core:** author editor meeting respStmt title **header:** funder principal sponsor

Example	`<titleStmt>`
	`<title>Capgrave's Life of St. John Norbert: a machine-readable transcription</title>`
	`<respStmt>`
	`<resp>compiled by</resp>`
	`<name>P.J. Lucas</name>`
	`</respStmt>`
	`</titleStmt>`
	Show all

Content model	`<content>`
	`<sequence>`
	`<elementRef key="title" minOccurs="1" maxOccurs="unbounded"/>`
	`<classRef key="model.respLike" minOccurs="0" maxOccurs="unbounded"/>`
	`</sequence>`
	`</content>`

Schema Declaration	XML syntax
	element titleStmt
	{
	att.global.attributes,
	att.global.rendition.attributes,
	att.global.linking.attributes,
	att.global.analytic.attributes,
	att.global.facs.attributes,
	att.global.change.attributes,
	att.global.responsibility.attributes,
	att.global.source.attributes,
	(title+, model.respLike*)
	}

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

8

2.1. Software for TEI-XML editing

- Advanced editors

they offer TEI validation, autocomplete, documentation

  - **oXygen XML Editor** (paid) https://www.oxygenxml.com/
  - **VS Code** (open source) https://code.visualstudio.com/
  - **Atom** (open source) https://atom.io/
  - **XML Copy Editor** (open source) http://xml-copy-editor.sourceforge.net/
  - **jEdit** (open source) http://www.jedit.org/
- Simple editors

they offer syntax colorization, sometimes XML validation

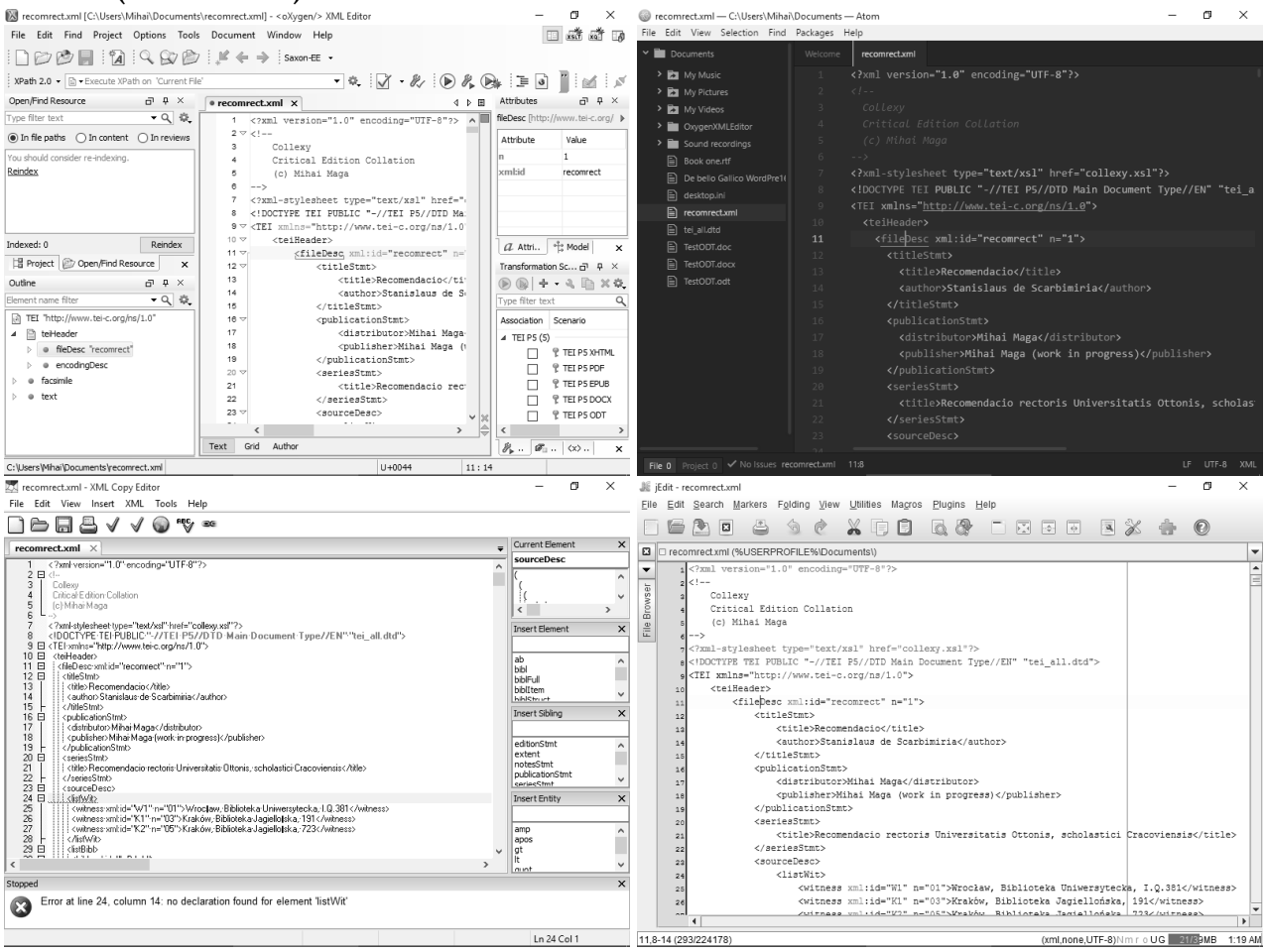
  - **Notepad++** (open source) https://notepad-plus-plus.org/
  - **Notepad2** (free) http://www.flos-freeware.ch/notepad2.html
  - **Eclipse** (open source) https://eclipse.org/
- Web editors

work in browser, they offer various functionalities

  - **CodeMirror** (open source) https://codemirror.net/
  - **eXide** (open source) http://exist-db.org/exist/apps/eXide/index.html

9

Editors (screenshots)



Homework

Using the following tags, encode the text below:

<text>	root element for the document
<p>	paragraph
<title>	title of a work
<name>	name of an author
<quote>	quoted text

> Quia intellectus habet duas operationes: scilicet unam qua format quiditates, in qua non est falsum, ut dicit Aristoteles in III *De anima*; aliam qua componit et dividit; et in hac etiam non est falsum, ut patet per Augustinum in libro *De vera religione*, qui dicit sic: “nec quisquam intelligit falsa”. Ergo falsitas non est in intellectu.
>
> Praeterea, Augustinus in libro *LXXXIII quaestionum*, quaestio 32: “omnis qui fallitur, id in quo fallitur, non intelligit”. Ergo in intellectu non potest esse falsitas.
>
> Item Algazel dicit: “aut intelligimus aliquid sicut est, aut non intelligimus”. Sed quicumque intelligit rem sicut est, vere intelligit. Ergo intellectus semper est verus; ergo non est in eo falsitas.

Thomas de Aquino, *Quaestiones disputatae de veritate*, Q. 1, art. 12