Digital Humanities

Digital Humanities for Medieval Philosophical Sources

9. Integration and digital processing

conf. dr. Mihai Maga

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

2nd semester, 2023–2024
HME2415/09

https://www.dhcluj.ro/dhm/

1. Indices and concordance tables
2. Query languages: XQuery, XPath
3. Search in text
4. Lemmatization, normalization, dictionaries
5. Digital corpora
6. Data-mining

Homework9

The extraction of significant elements (e.g. titles, names, quotes, terms) can be done automatically and immediately to the extant that:

- elements are marked with tags and/or attributes
- a procedure of query, extraction and display of elements is used

- `XQuery` (XML Query) is a functional language for querying collections of data in `XML` format
- `XPath` (XML Path Language) is subset of `XQuery`, defined as a query language for selecting nodes from `XML`

XPath Example (1)

If the following XML code is given:

```
<text>
   <body>
     <p>...</p>
   </body>
</text>
```

the expression to select all <p> nodes from <body> is:

```
/tei:text/tei:body/tei:p
```

XPath Example (2)

To extract all the elements with the attribute `class="sic"` from the next XML:

```
<p><w class="sic">unus</w> <w>duo</w> <w class="non">tres</w>
<w class="sic">quattuor</w></p>
```

the XPath expression can be:

```
//tei:w[@class='sic']
```

Text search can be one of many types:

- simple search: the given character string is searched

sum ⇒ ego sum, summa, assumptio, sensum

- wildcard search: the words which match the given character string and any characters in place of wildcard are searched

sum* ⇒ ego sum, summa, ~~assumptio~~, ~~sensum~~

- structured search: the given character string is searched in some given structure elements

e.g. `XPath: //tei:title[contains(., 'Sum')]`
⇒ `<title>Summa</title>`, ~~`<place>Sumer</place>`~~

The search by character string is not always satisfying, since the terms may appear in different forms because of the grammar inflexion and of the changing orthographic rules.

res, rei ⇒ res; rei; rem; re; rerum; rebus; ~~rex~~; ~~ires~~; ~~reddo~~; ~~resto~~; ~~recitares~~

Disambiguation techniques (in order of complexity):

- normalization: bringing the text to a standard form
- stemmatisation: extraction of the word stem (root)

re-s; ir-es; rest-o; recit-ares ...

- lemmatisation: attaching the standard form of the word to each occurence

rerum ~ res,rei; ires ~ eo,ire; recitares ~ recito,-are ...

TEI Example

```
<p><w lemma="omnis,-is,-e">Omnes</w> <w lemma="homo,-inis">homines</w>
<w lemma="naturaliter">naturaliter</w> <w lemma="scio,-re">scire</w> <w
lemma="desidero,-are">desiderant</w>.</p>
```

- dictionary: linking the words to an internal or external dictionary
- analysis: attaching the whole grammar analysis to a word

- A text corpus is a large structured collection of texts selected by a certain criterion (subject, period, author, intellectual milieu etc.)
- A digital corpus is a big dataset (text and metadata) which comprises a text corpus with annotations and which covers (at least at concept level) a certain principal criterion.
  - due to the intention of completeness, the corpus is tree-like built as an ontology (see *Course 7*)
  - a digital corpus is organized by scientific and functional principles which must ensure the reliability of its exploitation, including:
    - access to standardized versions of every text
    - simple and structured searches in the whole collection
    - exhaustive statistical analysis
    - automatic content analysis (e.g. linguistic analysis, data validation)
    - hypotheses testing, ensuring a high degree of confidence
    - transparency and predictability of the structures

The design of a corpus is a complex and important task, because it must cover all the possible manifestations of the contents, often without complete access to the contents.

- Data mining is the process of computer analyzing big datasets to identify meaningful patterns which help understanding and interpretation of contents.
  - the process is extremely complex, involving database systems, statistics, semantics, artificial intelligence and intellectual interpretation
  - usage scenarios:
    - validation of work hypotheses

    proving or disproving the existence of a pattern (e.g. influence of a certain factor)
    - discovery of significant models

    discovery of previously unknown regularities (e.g. a major but ignored factor)
    - analysis of pattern predictability

    using verified patterns, the content affiliation can be proved (e.g. authorship)

Write a XPath expression which extracts from the following TEI fragment all the occurrences of the verb *sum, esse, fui.*

ARISTOTELES, *Perihermeneias*

```
<p><w lemma="amplus">Amplius</w>, <w lemma="si">si</w> <w
lemma="sum">est</w> <w lemma="albus">album</w> <w lemma="nunc">nunc</w>,
<w lemma="verbum">verum</w> <w lemma="sum">erat</w> <w
lemma="dico">dicere</w> <w lemma="primus">primo</w> <w
lemma="quoniam">quoniam</w> <w lemma="sum">erit</w> <w
lemma="albus">album</w>, <w lemma="quare">quare</w> <w
lemma="semper">semper</w> <w lemma="verbum">verum</w> <w
lemma="sum">fuit</w> <w lemma="dico">dicere</w> <w
lemma="quilibet">quodlibet</w> <w lemma="ille">illud</w> <w
lemma="is">eorum</w> <w lemma="qui">quae</w> <w lemma="factus">facta</w>
<w lemma="sum">sunt</w> <w lemma="quoniam">quoniam</w> <w
lemma="sum">erit</w>; <w lemma="quod">quod</w> <w lemma="si">si</w> <w
lemma="semper">semper</w> <w lemma="verus">verum</w> <w
lemma="sum">est</w> <w lemma="dico">dicere</w> <w
lemma="quoniam">quoniam</w> <w lemma="sum">est</w> <w
lemma="vel">vel</w> <w lemma="sum">erit</w>, <w lemma="non">non</w> <w
lemma="possum">potest</w> <w lemma="hic">hoc</w> <w lemma="non">non</w>
<w lemma="sum">esse</w> <w lemma="nec">nec</w> <w lemma="non">non</w> <w
lemma="futurus">futurum</w> <w lemma="sum">esse</w>.</p>
```