

Digital Humanities

DIGITAL HUMANITIES PENTRU SURSELE FILOSOFICE MEDIEVALE

9. Integrarea și procesarea digitală

conf. dr. Mihai MAGA

Universitatea Babeș-Bolyai, Cluj-Napoca
Masteratul de Filosofie Antică și Medievală

semestrul II, 2020–2021

HME2415/09

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

Planul cursului

1.	Indici și tabele de concordanțe	2
2.	Limbaje de interogare: XQuery, XPath	2
3.	Căutarea în text	2
4.	Lematizare, normalizare, dicționare	2
5.	Digital corpora	3
6.	Data-mining	3
	Temă pentru acasă	3

1. Indici și tabele de concordanțe

Extragerea unor elemente semnificative (ex. titluri, nume, citate, termeni) se poate face automat și pe loc în măsura în care:

- elementele sunt marcate cu tag-uri și/sau atribute
- se folosește un procedeu de interogare, extragere și afișare a elementelor

2. Limbaje de interogare: XQuery, XPath

- XQuery (XML Query) este un limbaj funcțional de interogare pentru colecțiile de date în format XML
- XPath (XML Path Language) este un subset al XQuery, definit ca limbaj de interogare pentru selectarea nodurilor din XML

Exemplu XPath (1)

Dacă este dat următorul cod XML:

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

expresia pentru a selecta toate nodurile <p> din <body> este:

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

Exemplu XPath (2)

Pentru a extrage toate elementele cu atributul class="sic" din următorul XML:

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

expresia XPath poate fi:

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

3. Căutarea în text

Căutarea în text poate fi de mai multe tipuri:

- **căutare simplă:** se caută șirul de caractere dat

☞ `sum` ⇒ ego sum, summa, assumptio, sensum

- **căutare cu metacaractere:** se caută cuvintele care se potrivesc șirului de caractere dat și cu orice caractere în locul metacaracterelor

☞ `sum*` ⇒ ego sum, summa, assumptio, sensum

- **căutare structurată:** se caută șirul de caractere dat în anumite elemente de structură

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

4. Lematizare, normalizare, dicționare

Căutarea după șiruri de caractere nu este întotdeauna satisfăcătoare, întrucât termenii pot apărea în forme diferite din cauza flexiunii gramaticale și a normelor ortografice schimbătoare.

☞ `res, rei` ⇒ res; rei; rem; re; rerum; rebus; rex; ired; reddo; resto; recitares

Tehnici de dezambiguizare (în ordinea complexității):

- **normalizare:** aducerea textului la o formă standard
- **stematizare:** extragerea stemei (rădăcinii) cuvântului

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

- **lematizare:** atașarea formei standard a cuvântului la fiecare ocurență

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

Exemplu TEI

```
<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>
```

- **dictionar:** conectarea cuvintelor la un dictionar intern sau extern
- **analiză:** atașarea întregii analize gramaticale la un cuvânt

5. Digital corpora

- Un **corpus de texte** este o mare colecție structurată de texte selectate după un anumit criteriu (subiect, perioadă, autor, mediu intelectual etc.)
- Un **corpus digital** este un set mare de date (text și metadata) care cuprinde un corpus de texte cu adnotări și care acoperă (cel puțin la nivel de concept) un anumit criteriu principal.
 - datorită intenției de completitudine, corpus-ul este construit arborescent ca o ontologie (vezi *Curs 7*)
 - un corpus digital este organizat după principii științifice și funcționale care trebuie să asigure fiabilitatea exploatării lui, incluzând:
 - acces la versiunile standardizate ale fiecărui text
 - căutări simple și structurate în întreaga colecție
 - analize statistice exhaustive
 - analize automate de conținut (ex. analize lingvistice, validarea datelor)
 - testarea ipotezelor, cu asigurarea unui grad mare de încredere
 - transparența și predictibilitatea structurilor

! Proiectarea unui corpus este o sarcină complexă și importantă, deoarece trebuie să acopere toate manifestările posibile ale conținutului, deseori fără a avea acces complet la conținut.

6. Data-mining

- **Data mining** (mineritul datelor) este procesul de analiză computerizată a seturilor mari de date pentru identificarea de modele semnificative care ajută înțelegerea și interpretarea conținuturilor.
 - procesul este extrem de complex, implicând sisteme de baze de date, statistică, semantică, inteligență artificială și interpretare intelectuală
 - scenarii de utilizare:
 - validarea unor ipoteze de lucru
 - ➡ dovedirea sau respingerea existenței unui model (ex. influența unui anumit factor)
 - descoperirea de modele semnificative
 - ➡ descoperirea de regularități necunoscute anterior (ex. un factor major, dar ignorat)
 - analiza predictibilității modelelor
 - ➡ folosind modele verificate, se poate proba apartenența conținutului (ex. paternitatea)

Temă pentru acasă

Scrieți o expresie XPath care să extragă din următorul fragment TEI toate ocurențele verbului *sum*, *esse*, *fui*.

ARISTOTELES, *Perihermeneias*

```
<p><w lemma="amplus">Amplus</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
```

ARISTOTELES, *Perihermeneias* (cont)

```
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>
```