

Semantic content retrieval in digital cultural heritage objects

Cristina Vertan
cristina.Vertan@uni-hamburg.de



Ştir

Vineri, 28 Octombrie 20

Vizita premierului Dacian Ciolos în județul Sălaj
Dacian Ciolos: Am avut o întâlnire cu conducerea județului Sălaj, cu domnul președinte al Consiliului Județean, primari Municipiului Zalău și ai orașelor din județul Sălaj și cățiva membri ai Parlamentului, parlamentari de aici, din județ și am trecut în revistă câteva proiecte.

Vineri, 28 Octombrie 201

Participarea premierului Dacian Cioloș la ceremonia de comemorare a luptelor din Defileul Jiului
Dacian Cioloș: Eroii bătăliei din Defileul Jiului au avut totul de pierdut și nimic de câștigat pentru personal. Sacrificiul lor constituie cel mai bun exemplu de patriotism.

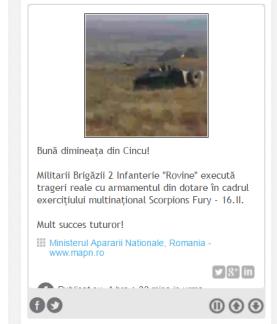
Joi, 27 Octombrie 2016

301, 27 octombrie 2016
Dragos Dipu, propunerea premierului pentru ministrul Fondurilor Europene

Dragoș Dinu, propunere premierului pentru ministrul Fondurilor Europene
Premierul Dacian Cioloș a înaintat în această dimineață domnului președinte al României, Klaus Iohannis, propunerea de numire a domnului Dragos Cristian Dinu în funcția de ministru al Fondurilor Europene.

1

[VEZI TOATE STIE](#)



Bună dimineață din Cincu!

Mult succes tuturor!

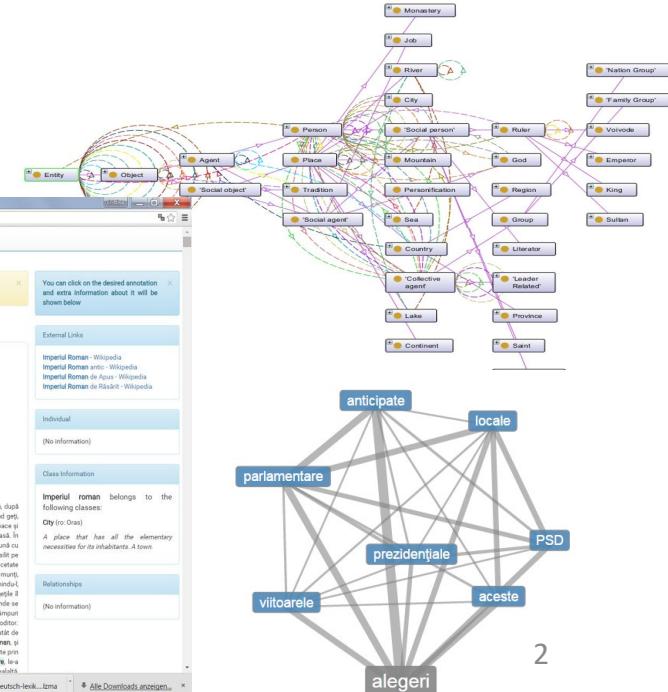
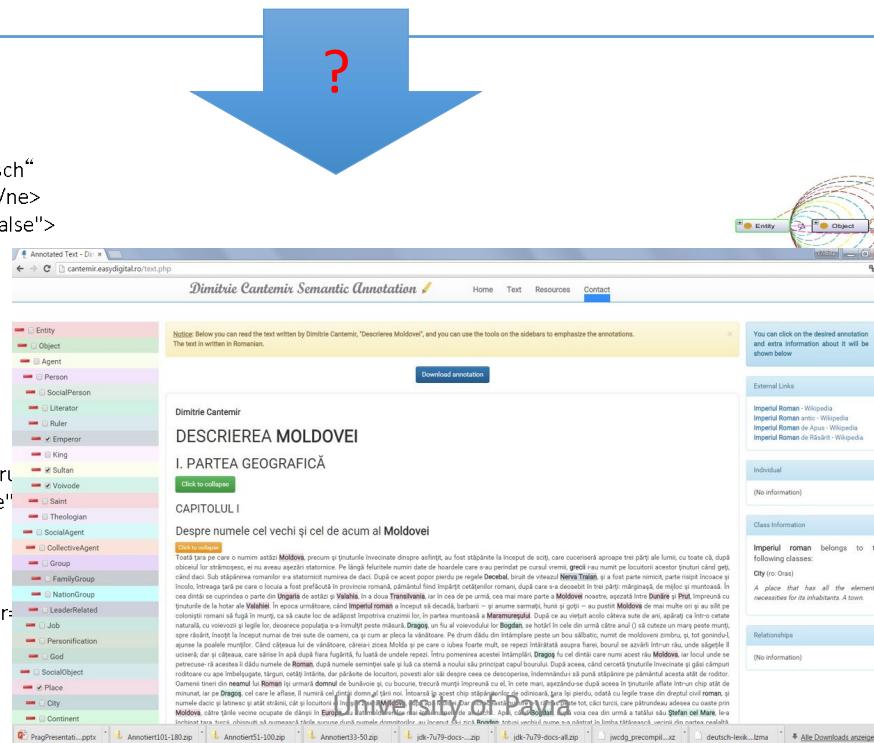

```
<phrase>
    <ne> <person personT="f" first="„Dragosch“
        relationT="„son“ refT=" „w123“" /> </ne>
    <phraseParts constituentT="NP" mult="false">
        <tok idT="" langT="ro" styleTranscr="de">
            <pos posT="N"/>
            <string> Dragosch</string>
        </tok></phraseParts>
    </phrase>
```

, ein Sohn ihres
<phrase>

```
<ne> <person personT="tl"/></ne>
<phraseParts constituentT="NP" mult="tru"
  <tak idT="" langT="de" styleTranscr="de"
  <pos posT="N"/>
  <string> Königs </string>
</tok>
<tak idT="w123" langT="ro" styleTranscr=
  <pos posT="N"/>
  <string> Bogdan</string>
</tok> </phraseParts>

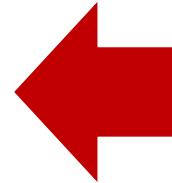
```

19.02.2020



Contents

- Digital Humanities
- Shallow annotation and retrieval
- Light semantic annotation and retrieval
- Ontologies and semantical retrieval



Digital Humanities (DH)

- Alternative labels:
 - Humanities Computing
 - Computerphilology /Computerphilologie (especially in German speaking area)
- Humanities and Social sciences are usually different areas but under „Digital Humanities“ is embedded also the processing of social science data.
- Main principle: Use methods from CS in order to:
 - Make data accessible by computers (browsing, retrieval)
 - Investigate data and discover new insights in own field of research.

Linguistics

History

Philology

Archeology

Sociology

Etnography

Music

Digital Humanities

Computer Science:

- ◆ Data Structures
- ◆ Software engineering
- ◆ Image processing
- ◆ Charcter recognition
- ◆ Language Technology
- ◆ Machine Learning
- ◆ Intelligent Retrieval
- ◆ Networks and Protocols
- ◆ Vizualization
- ◆ Testing and Evaluation

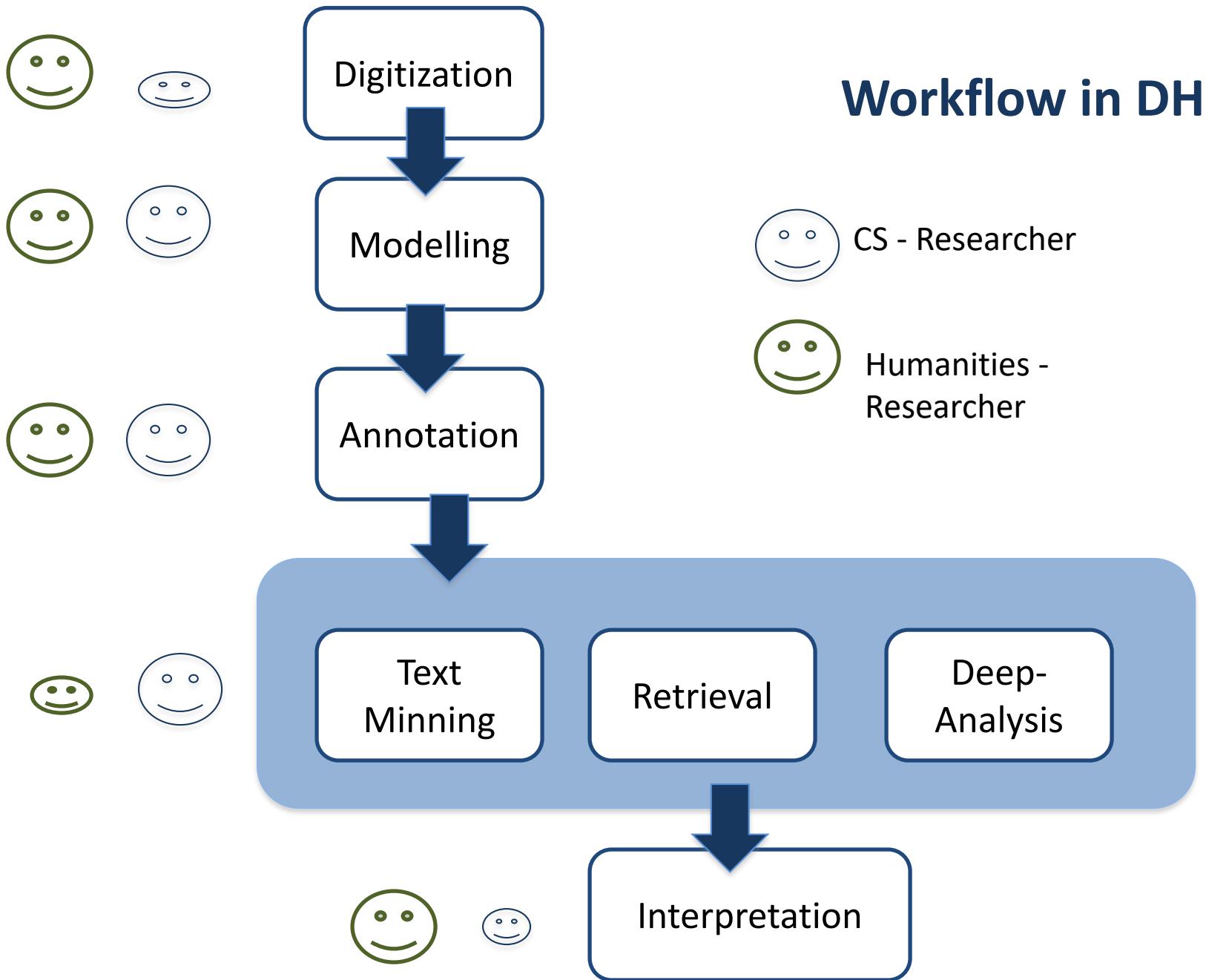
Statistics

- ◆ Support for CS –Methods
- ◆ Automatic Evaluation
- ◆ Quantitative analysis

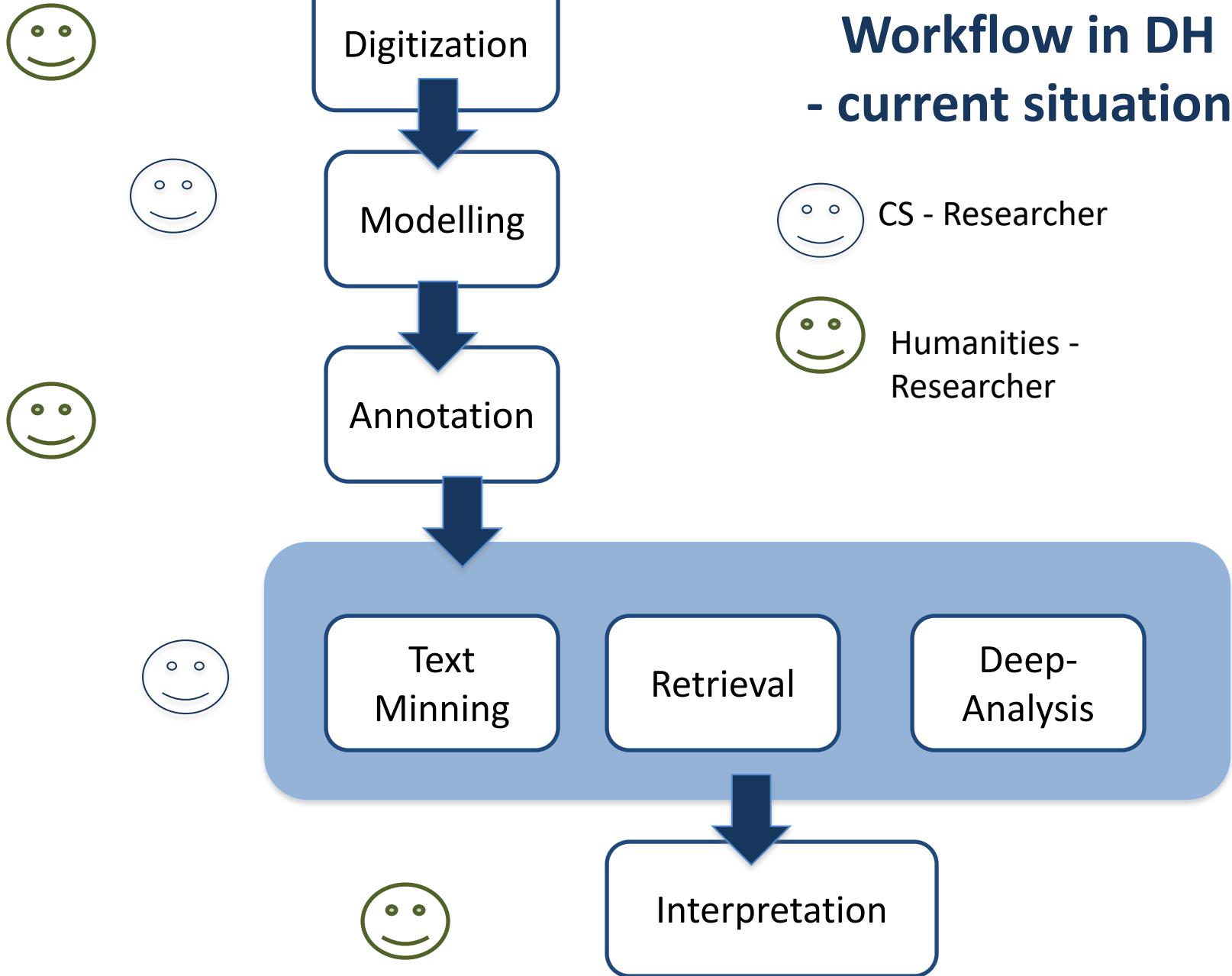
Digital Humanities enable

- Cultural Heritage preservation
- Discovery of new results by means of computer processes
- Analysis of data sets spread in different locations
- Analysis of big data sets
- Make cultural heriatge accessible to broader groups
- Remote interdisciplinary work among different fields of Humanities

Workflow in DH



Workflow in DH - current situation-



Workflow in Digital Humanities

-Desiderata-

- Each step involves both researchers in Humanities and Computer Sciences in order to:

Humanists

- deliver specifications (they are the end-users)
- Become aware of the newest methods in CS and thus decide what seems better for their data
 - Explain particularities of the data
- Give feedback on results, user friendly interface

Computer Scientists

- deliver the newest methods of the field
- adapt these methods to the data and user requirements
- design models and software
- Evaluate performance

In contrast with many other applications data is highly unstructured, underspecified, evaluation is difficult as gold standard is missing

DH means **RESEARCH** in Computer Science and NOT ONLY Programming

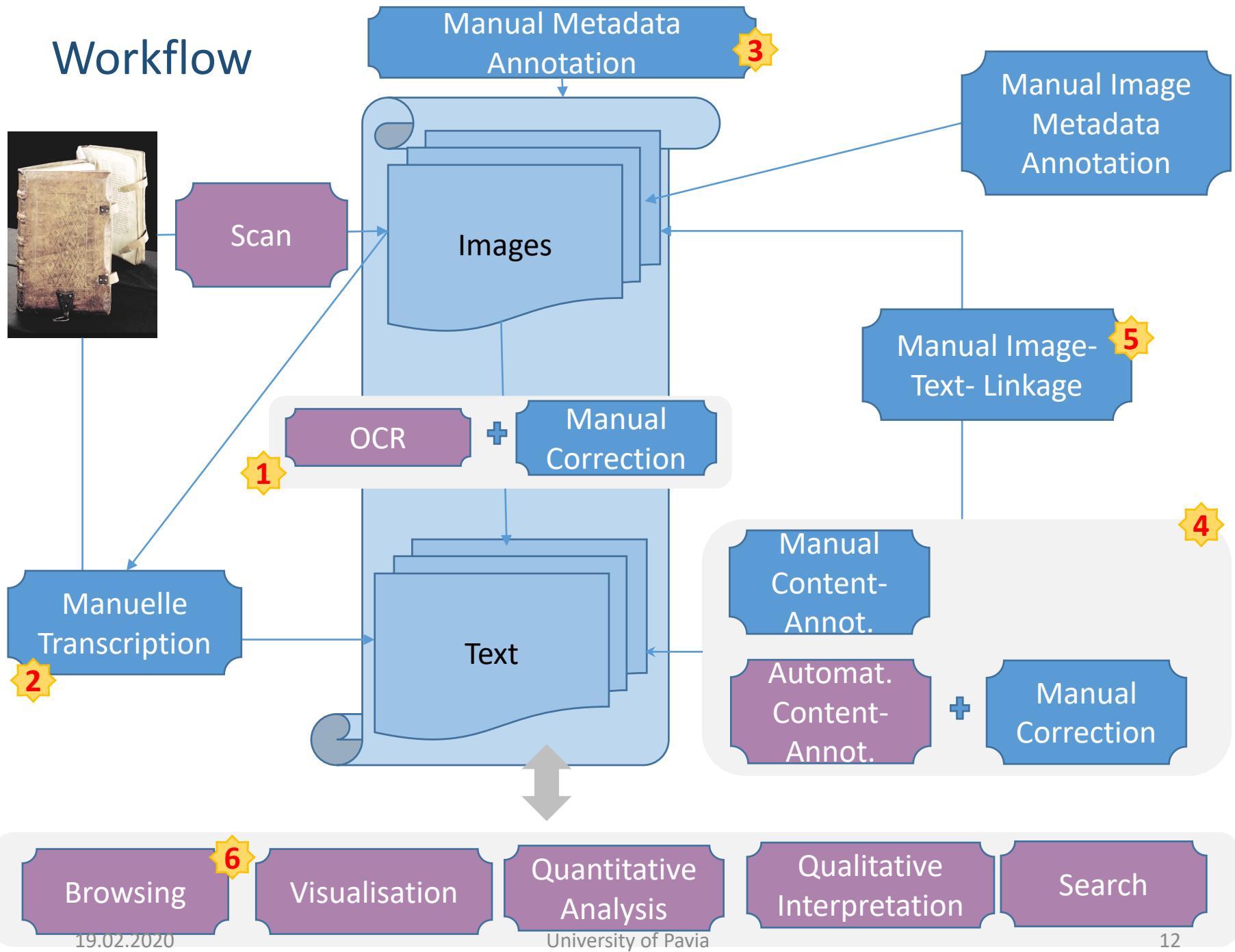
Cutting edge fields in CS of great interest for DH -1-

- Big Data (meanwhile million of digitized books and other material related to DH: "The Billion-Word Library" -<http://asv.informatik.uni-leipzig.de/en/projects/29>). Fast access to heterogenous content, up to word or event character level
- Linked Open Data (access and linking of units within different digitized materials)
- Machine Learning –ML- (detect automaticaly inherent structures within texts, autmatic recognition of certain literary motivs, linguistic features etc). Usually in humanities these structures / features are not known → unsupervised learning (which is still a theoretical probelm of ML)
- Processing of multilingual data (historical texts are usually a mixture of local language, Latin, ancient Greek, at least)
- Processing of multimodal data (Image and text in old books, audio and video in social netwokr analysis, audio and text in musicology)

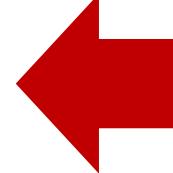
Cutting edge fields in CS of great interest for DH -2-

- Machine Translation (translation of oldtexts is highly requested, to make texts available to the new generations, but the process is even complexer as for modern languages)
- Intelligent Retrieval (semantic access not only to meta-data but to content)
- Multidimensional Visualisation (data is usually annotated in more dimensions which eventually interact one with the other).
- Man-Machine Communication – design of user interfaces which suit the requirements of researchers in humanities (not necessary the current GUI-Desktop solutions)

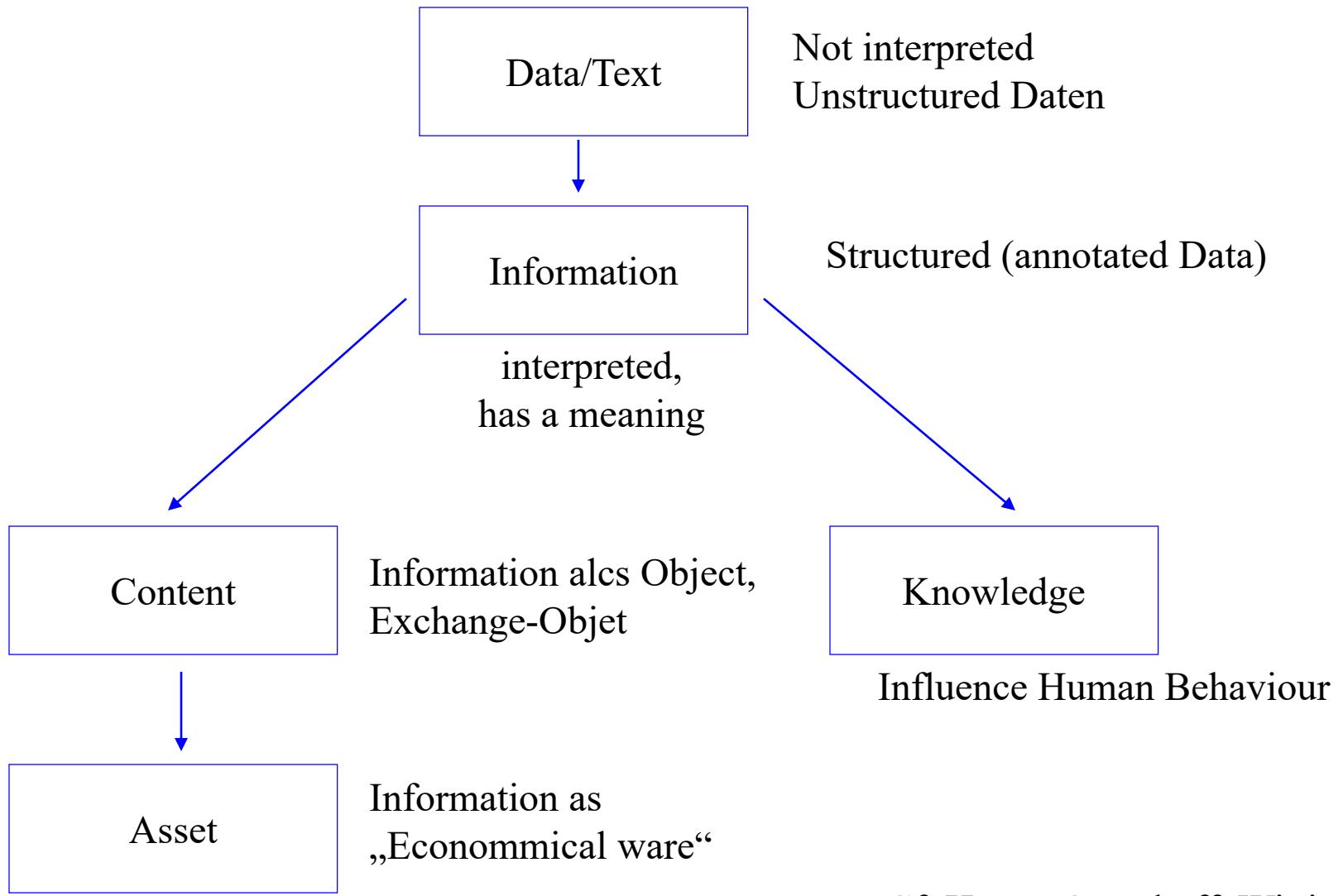
Workflow



Contents

- Digital Humanities
 - Shallow annotation and retrieval
 - Light semantic annotation and retrieval
 - Ontologies and semantical retrieval
- 

Role of Annotations



Annotation Levels

- Metadata

- Textstructure, Paragraphs, Sentences, other text parts
- Pagebreak, Line Break
- Word-Formatting features like Bold, Italic, Underlined, Colours, Size
- Alignment Right, Left, Centered

LAYOUT

- Text-Edition (Quotation, Comments, Text-critical edition Mark-up)
- Musiknotations
- Linkage with domain external sources (Knowledge base)

DOMAIN

- | | | |
|-----------------|---|--------------|
| • Morpho-Syntax | Part of Speech | LINGUISTIC S |
| • Morphologie | Inflected forms, Lemmas | |
| • Syntax | Constituents, Dependencies | |
| • Semantics | semantic Roles, Proper Names | |
| • Pragmatics | co-referrence, Information structure, Discourse structure | |
| • Other | Orthography, Time, Emotion, Gestures, Mimics | |

TEI-Text Encoding Innitiative

TEI is an XML-based mark-up language which tries to provide an unique vocabulary for various humanities fields

"The Text Encoding Initiative (TEI) is a consortium which collectively develops and maintains a standard for the representation of texts in digital form"

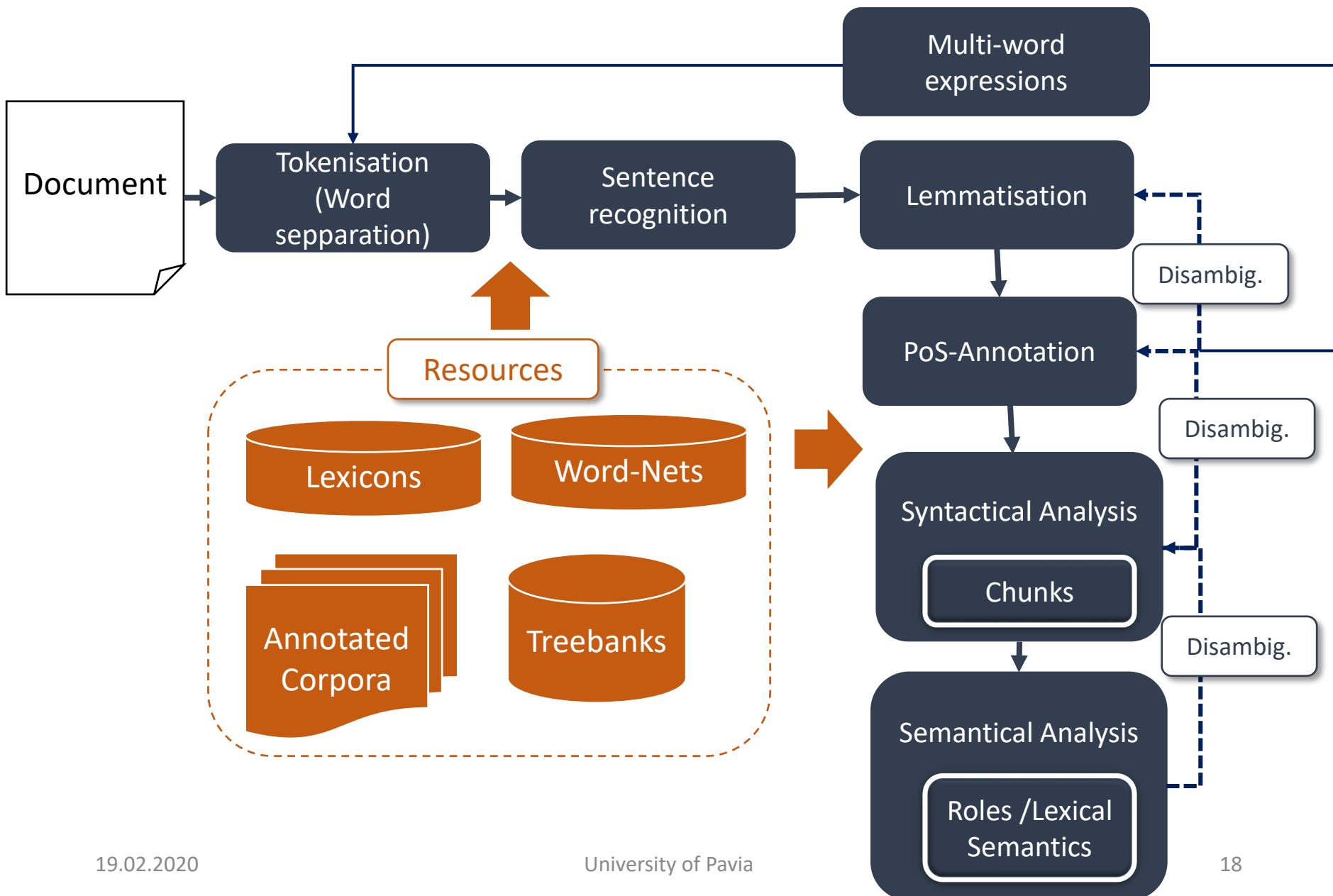
aus „<http://www.tei-c.org/index.xml>“

There are TEI–Modules for different application types and domains.
The TEI-Schema can be extended. A Web-basedTool for TEI-Extesions is available

TEI-Module

Module name	Formal public identifier	Where defined
analysis	Analysis and Interpretation	17 Simple Analytic Mechanisms
certainty	Certainty and Uncertainty	21 Certainty, Precision, and Responsibility
core	Common Core	3 Elements Available in All TEI Documents
corpus	Metadata for Language Corpora	15 Language Corpora
dictionaries	Print Dictionaries	9 Dictionaries
drama	Performance Texts	7 Performance Texts
figures	Tables, Formulae, Figures	14 Tables, Formulae, and Graphics
gaiji	Character and Glyph Documentation	5 Representation of Non-standard Characters and Glyphs
header	Common Metadata	2 The TEI Header
iso-fs	Feature Structures	18 Feature Structures
linking	Linking, Segmentation, and Alignment	16 Linking, Segmentation, and Alignment
msdescription	Manuscript Description	10 Manuscript Description
namesdates	Names, Dates, People, and Places	13 Names, Dates, People, and Places
nets	Graphs, Networks, and Trees	19 Graphs, Networks, and Trees
spoken	Transcribed Speech	8 Transcriptions of Speech
tagdocs	Documentation Elements	22 Documentation Elements
tei	TEI Infrastructure	1 The TEI Infrastructure
textcrit	Text Criticism	12 Critical Apparatus
textstructure	Default Text Structure	4 Default Text Structure
transcr	Transcription of Primary Sources	11 Representation of Primary Sources
verse	Verse	6 Verse

Automatic linguistic annotation



Challenges of Shallow linguistic annotation in DH and limitations of content retrieval an example

Controlled semi-automatic annotation of a diachronic
corpus of classical Ethiopic texts



TraCES - From Translation to Creation: Changes in Ethiopic Style and Lexicon from Late Antiquity to the Middle Ages

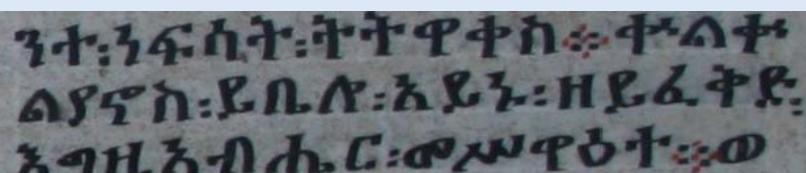
- ERC Advanced Grant 2014–2019



Overview

- Ge'əz language
- Tagset and annotation workflow
- Challenges for annotation software
- Data model behind GeTa
- Integration in ANNIS





a	u	i	ā	e	ə/ø	o
h	ሀ	ሁ	ሃ	ሇ	ህ	ላ
l	ለ	ሉ	ሉ	ሉ	ሉ	ሉ
h	ሐ	ዐ	ዘ	ዘ	ዘ	ዘ
m	መ	መ	ማ	ማ	ማ	ማ
ś	ሮ	ሮ	ሮ	ሮ	ሮ	ሮ
r	ረ	ሩ	ሩ	ሩ	ሩ	ሩ
s	ሰ	ሰ	ሰ	ሰ	ሰ	ሰ
q	ቁ	ቁ	ቁ	ቁ	ቁ	ቁ
b	በ	በ	በ	በ	በ	በ
t	ተ	ተ	ተ	ተ	ተ	ተ
h	ተ	ተ	ተ	ተ	ተ	ተ
n,	ና	ና	ና	ና	ና	ና
k	ከ	ከ	ከ	ከ	ከ	ከ
w	ወ	ወ	ወ	ወ	ወ	ወ
c	ወ	ወ	ወ	ወ	ወ	ወ
z	ዘ	ዘ	ዘ	ዘ	ዘ	ዘ
y	የ	የ	የ	የ	የ	የ
d	ደ	ደ	ደ	ደ	ደ	ደ
g	ገ	ገ	ገ	ገ	ገ	ገ
t	ተ	ተ	ተ	ተ	ተ	ተ
P	ጠ	ጠ	ጠ	ጠ	ጠ	ጠ
s	ሳ	ሳ	ሳ	ሳ	ሳ	ሳ
d	ሳ	ሳ	ሳ	ሳ	ሳ	ሳ
f	፩	፩	፩	፩	፩	፩
p	፪	፪	፪	፪	፪	፪
q ^w	ቁ	ቁ	ቁ	ቁ	ቁ	ቁ
h ^w	ተ	ተ	ተ	ተ	ተ	ተ
k ^w	ከ	ከ	ከ	ከ	ከ	ከ
g ^w	ዘ	ዘ	ዘ	ዘ	ዘ	ዘ

Gə'əz - features

- (Classical) Ethiopic, Semitic language
- Each symbol represents a **syllable (fidal)**
- Vowels are represented
- Left-to-right
- Dedicated symbols for **Digits**
 - Transcription and value are different
 - ጥዣ፫ 1 wa 5, ‘1und5’ (value: 6)
- Dedicated interpunction: (Space) :: (. or ,) :: (,) :: (:) :: (;) :: (?) :: (Paragraph)
- Nonconcatenative Morphology
 - ✓ **qds** × 'asta1a22a3a → 'astaqaddasa



Gə'əz- features

Vowels can be themselves PoS

Letter compression in original script, but not in the transcription

and	House	his
ω	ə	ħ
wa	be	tu
wa Conj	bet N	u Pr

before		the days
ħ	ω	ɸ
'ə	ma	wā
'əm Prep		mawā'əl N



Language Documentation

- **vast array of written sources** (monumental inscriptions, manuscripts manufactured until today)
- literary texts: **early translations from Greek**, later translations from **Arabic**, and **indigenous works compositions**
- ‘understudied’ & underresourced language
- **central aims:**
 - **tool box**
 - **provide linguistic data** for a diachronic analysis of Gə‘əz related to its lexicography, morphology and style



Gə‘əz Texts

- **current corpus:** texts belonging to different historical periods, type of transmission, literary genres
 - Gə‘əz of the Aksumite period: vocalized inscriptions
 - Gə‘əz of the Aksumite translations from Greek: *Gospel of Matthew, Testamentum Domini, Epistle of Eusebius to Carpius*)
 - Gə‘əz of mediaeval translation from Arabic: *Life of Secundus the Silent*
 - Indigenous texts composed in Gə‘əz (partially inspired by translated models): *Glory of the Kings, Book of Mysteries, Chronicle of ‘Amida Ṣayon*
- **genres:** Bible, theology, history and hagiography, and fiction
- based ideally on **text-critical editions**



Tag Set

- **33 PoS tags and features**
- **six main categories**, some have further subdivisions
 - **Nominals**: nouns (2 subdivisions), pronouns (10), numerals (2)
 - **Verbs**
 - **Existentials**: affirmative and negative
 - **Particles**: adverbs (2), prepositions, conjunctions, interjections, further particles (9)
 - **Foreign Material**
 - **Punctuation**
- **most complex PoS**: Common Nouns (numerous features, ambiguous forms)



Research Infrastructure

- main tool for the linguistic annotation:
Gə‘əz Text annotator (GeTa)
- digital dictionary of Gə‘əz
<https://betamasaheft.eu/Dillmann/>
Lexicon linguae Aethiopicae by Dillmann 1865

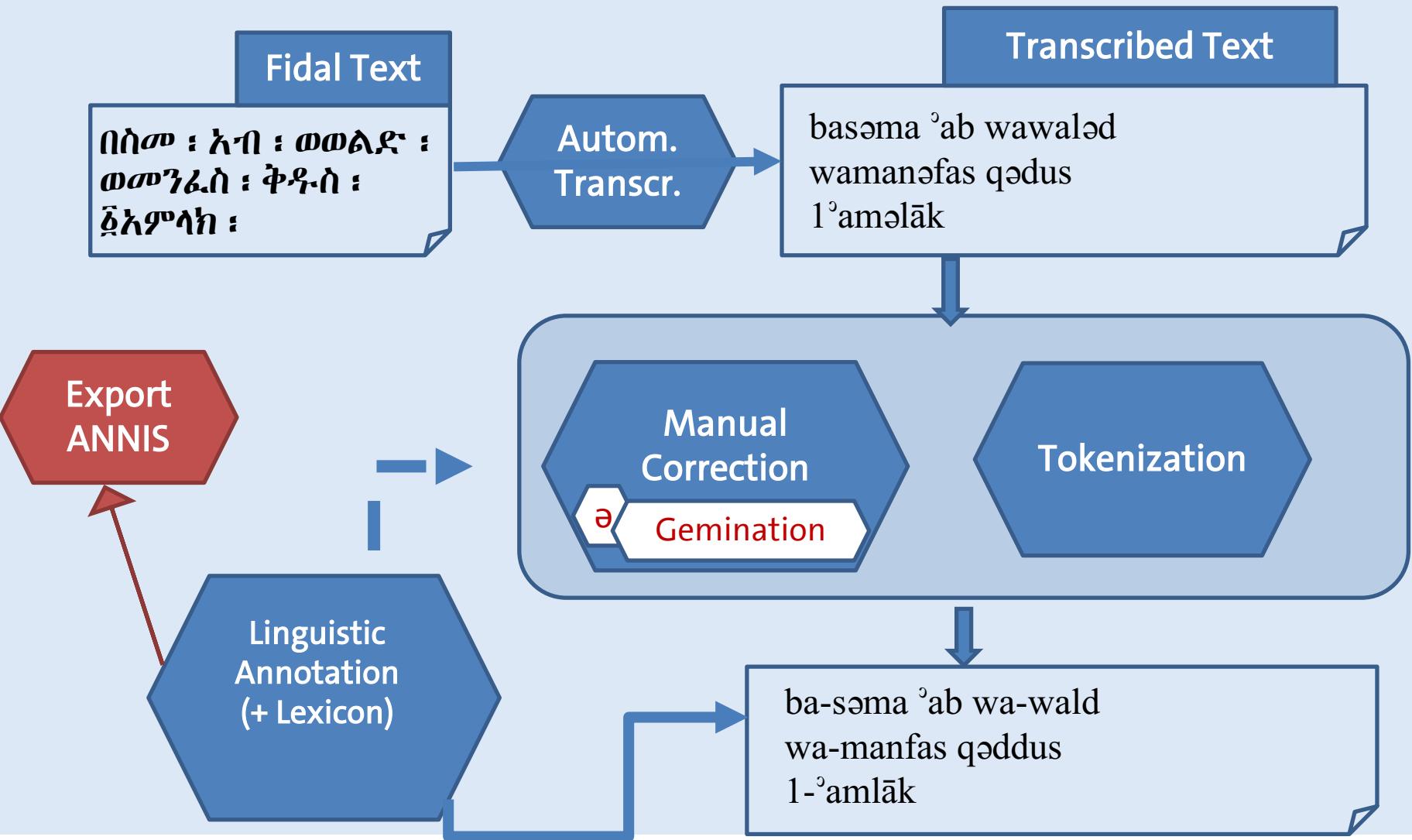


Multilevel annotation

- Each text is (in full or in part) annotated at different levels
- main level: detailed linguistic (part-of-speech) annotation, each word linked to the corresponding dictionary entry
 - morphological criteria, but disambiguation (transliteration, tokenization, assigning the correct PoS tag)
- named entities (persons, places, dates, titles of work, offices)
- text division (parts, chapters, sentences, verses)
- features of the edition (e.g., editorial intervention)



Workflow – Linguistic Annotation



Disambiguation: Gemination of a Consonant

ይተቀተለ :

Gə'əz form for
Imperfect,
Jussive
(3 m.Pl., 'to make war')



Disambiguation of the vowel -ə

ex.:

ያደርን :

Gə'əz form for
Imperfect,
Jussive
(3 m.s., 'to save')

yādəhən → Imperfect 3 m.s.

yādhən → Jussive 3 m.s.



Disambiguation: Tokenization

same Gə'əz forms with
different meanings → different no. tokens

Tokens:

gabru 'they did' or *gabr-u* 'his servant'

1 TOKEN = Verb
(3 m. Sg., Impf.)

2 TOKENS = NCom + PSuff



Annotation Common Noun in GeTa

Annotations for the word "rətu'āna" (highlighted in blue) are shown across three windows:

- Linguistic Annotation rətu'āna** (Left Window):
 - Pos: Common Noun
 - Lemma Id: 2eb31e401f87eb7c55ec9ca5a0-C-kəð
 - Deep Annotation: masculineP PluralExtP UnmarkedS Nominative Construct state
 - Buttons: Complete, Local (radio button), Global, Global-Complete
- Deep Annotation** (Middle Window):
 - Gender: masculine (checked)
 - Number: Pattern (radio button)
 - Singular
 - Plural external (radio button)
 - Plural internal
 - Plural of the plural
 - Unmarked
 - Syntax: Nature, Pattern, Syntax (checkboxes)
 - Case: Nominative (radio button)
 - Accusative
 - 0-Accusative
 - Vocative
 - State: Construct State (radio button)
 - Absolute State
 - 0-Construct State
 - Pronominal State
- Transliterated KN_GeTa Ver5_7May2019.json** (Right Window):
 - Transliteration: la-'əgzi'ābəher
 - Lemma: kʷəll-u
 - Transliteration: 'iyasus
 - Lemma: kʷəll-u
 - Transliteration: 'alb-o
 - Lemma: wa-ba-manfas
 - Transliteration: za-yəwaddə'
 - Lemma: 'əm-wald
 - Transliteration: wa-wald
 - Lemma: na'ammən
 - Transliteration: .
 - Lemma: kəbra
 - Transliteration: fəkkāre
 - Transliteration: rətu'āna
 - Lemma: kəbr
 - Transliteration: za-kama
 - Transliteration: la-daqqa
 - Lemma: wa-fadfāda-ssa
 - Transliteration: wa-kəbr-ā
 - Lemma: həgg-u
 - Transliteration: gabāri-hā
 - Lemma: ba-wəsta
 - Transliteration: 'əm-qədma



Annotation Requirements and Challenges -1-

- Automatic transcription
- Synchronisation of original script and transcription
- Possibility of editing the base text during the annotation with preservation of the annotated features
- Semi-automatic controlled processes for:
 - Tokenisation
 - Edition of the text
 - Linguistic annotation
 - „Sentence“ -recognition
 - NE-Annotation
- Flexible and adaptable annotation of the text structure;
Possibility of annotating overlapping text structures
(different text editions)



Annotation Requirements and Challenges -2-

- Multilevel annotation with possibility of changing the current annotated level dynamically
- **User-friendly annotation especially for the high number of linguistic tags and their features**
- **Support for Texts with non-vocalized transcription or in South-Arabic script (right to left, different script)**

ይና ከሂሳብ የሚከተሉ የሚመለከት ስርዓት በዚህ የሚከተሉ የሚመለከት ስርዓት
ኩል እንደሚከተሉ የሚመለከት ስርዓት የሚመለከት ስርዓት
አሁን የሚከተሉ የሚመለከት ስርዓት የሚመለከት ስርዓት

- Integration with tools on the Beta Maṣāḥēft platform (lexicon, named entities)



Annotation levels

1. Text structure (name of each section can be defined by the user)
 - Part
 - Chapter
 - Section
 - Verse / Sentence
2. Morphology (according with the linguistic tag-set)
3. Entities
 - Persons
 - Places
 - Numbers
 - Titles
 - Ethnical Groups
 - Works
4. Edition
 - Page break
 - Line break
 - Corrections in original text (e.g. according to an edition)
 - Annotator corrections to the original

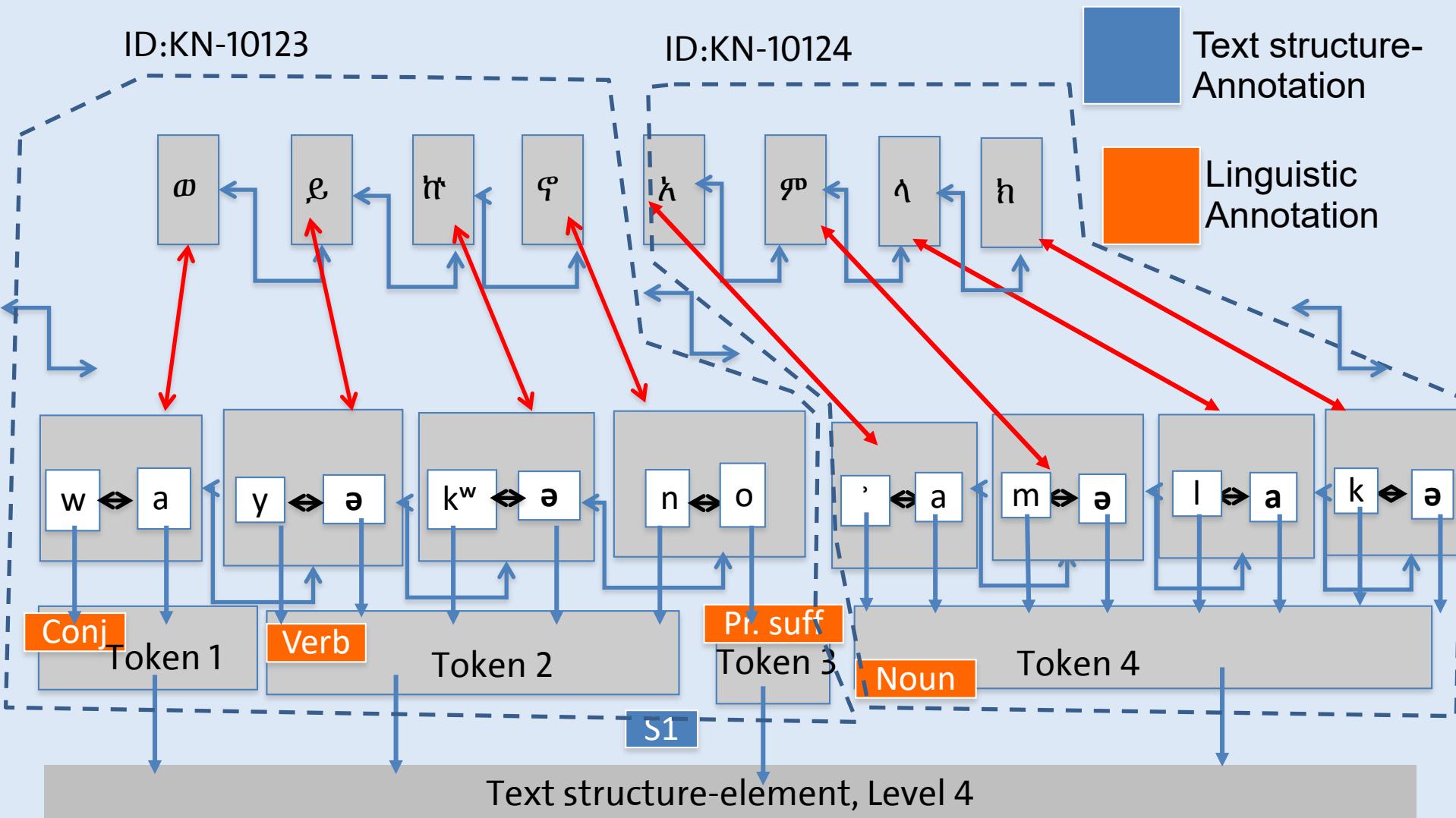
Comments for each:

- Annotation
- Token (transliteration)
- Word in original script (graphical unit)



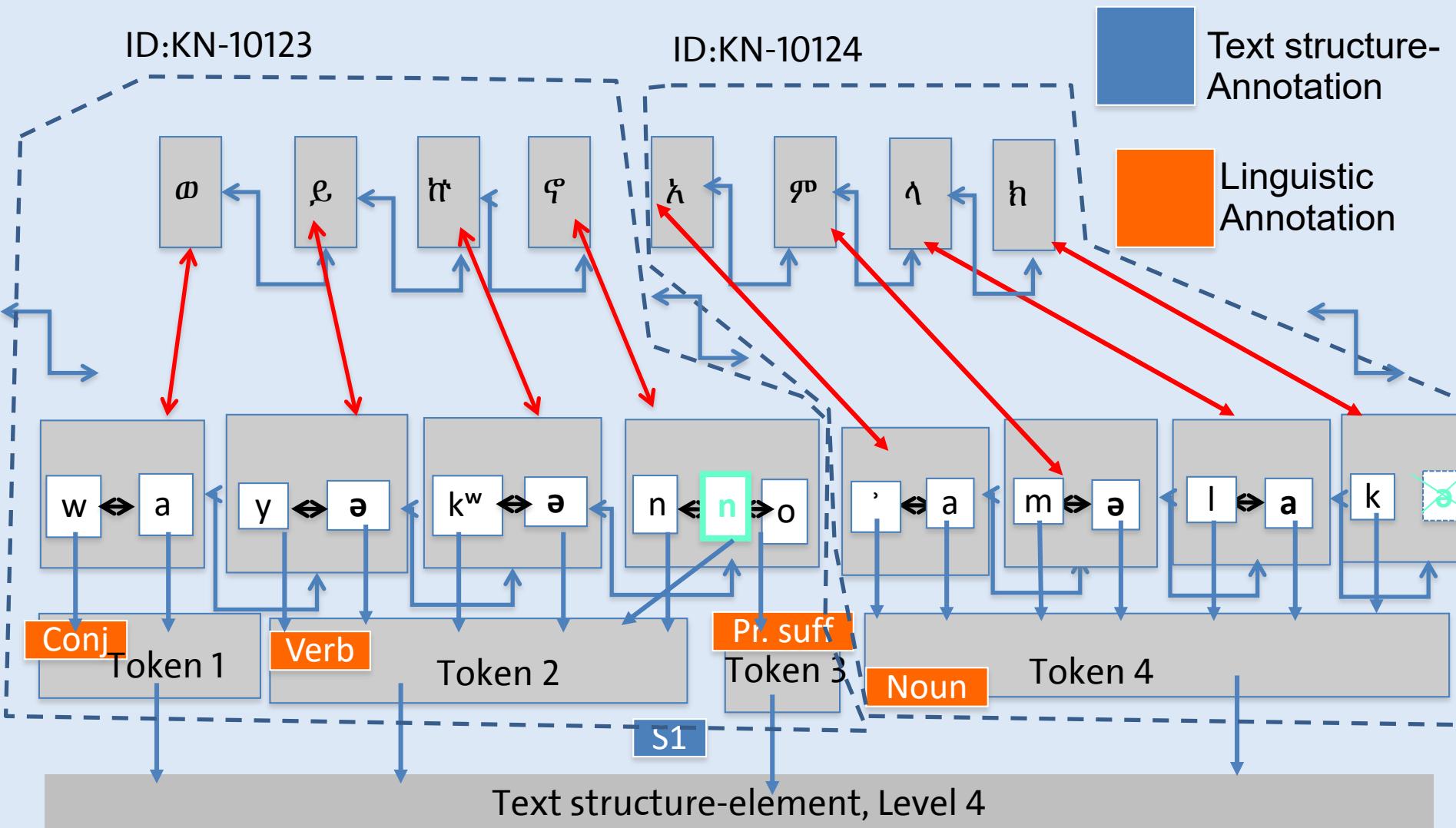
ወይም አጥቃለሁ:

GeTa: Data-Model -1



ወይናል፡ አጥቃለሁ፡

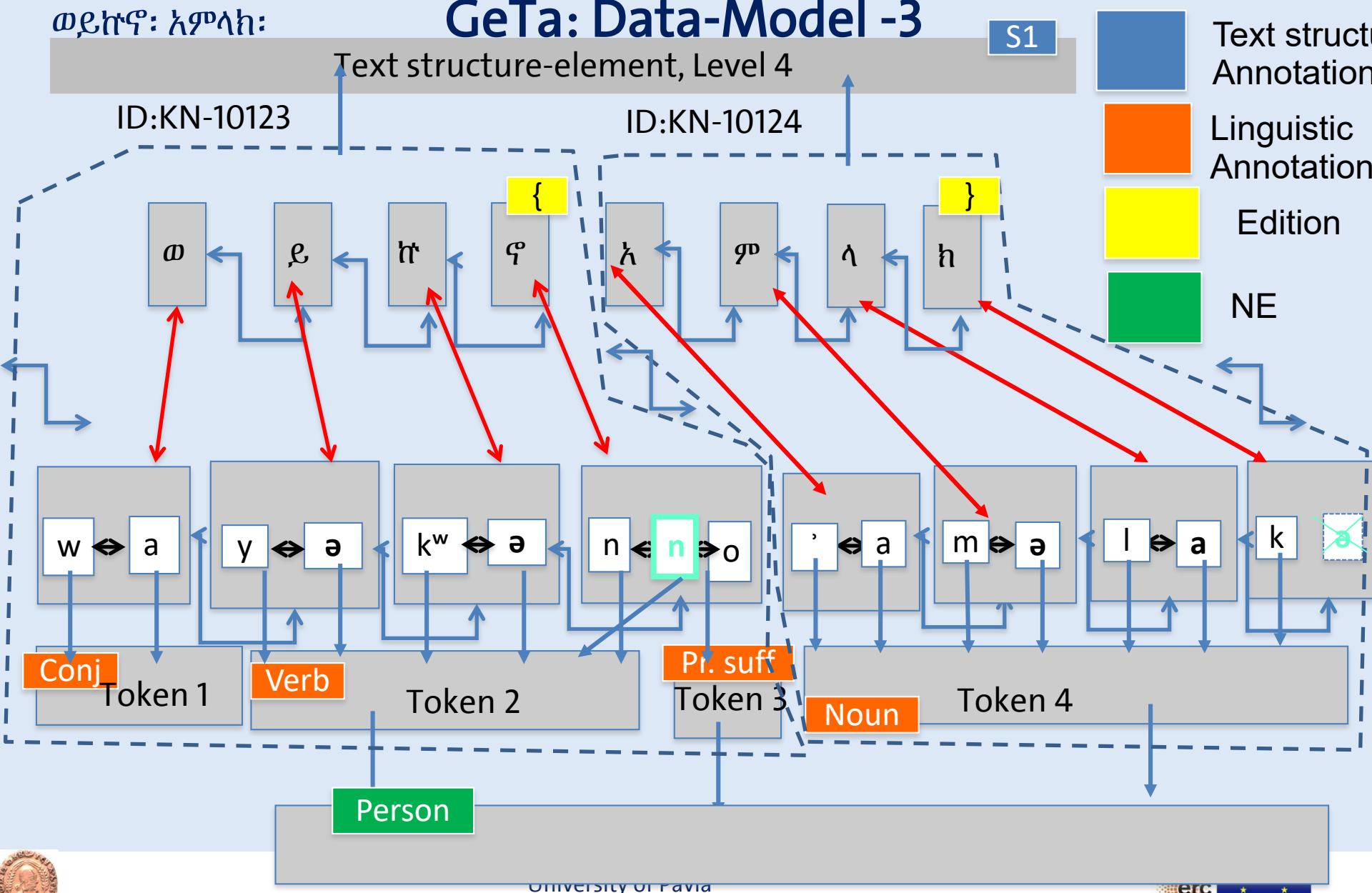
GeTa: Data-Model -2



GeTa: Data-Model -3

ወይኬና፡ አጥልክ፡

S1



GeTa – Semi-automatic controlled processing

GeTa - TrACES Annotation Tool Ver4

File History Show Window Tools View Annotations Keyboards Help Automatic Annotation

Show cursor position Level section: section 60(Internal Nr.: 60)

Original Matthew_20170615_1931.json

በኩ :	ንኩ :	አ
ወቃ :	ወቃ :	ከ
ዘመኑርአም :	በመቁ :	ማ
እስ :	እብጽአም :	፻
መልዕሉ :	አን :	፻
ወኩ :	ርኢም :	፻
ተክ :	ተኩ :	፻
የት :	የት :	፻
ወለማርያም :	አሙ :	፻
አሁ :	ወአጥ :	፻
ወረት :	መቀበት :	፻
አ :	አም :	፻
ወሰ :	ወሰ :	፻
በአለም :	ከ :	፻
ኋ :	ሂደት :	፻
ከፊ :	ጥ :	፻

Transliterated Matthew_20170615_1931.json

hədānāta	za-beta	ləhem
wa-za-kwəll-u	dawal-a	za-kəl'ə
kəramt-u	wa-za-yənə'əs	'əmənne-hu
ba-hassaba	taṭayyaq-omu	la-sab'a
sagal		kama
yəbəsh		nabiyy
za-yəbe		qal
tasam'a		həmām
wa-bəkāy		bəzuḥ
rāhel		
wa-'abayat		
'əsma		
wa-mawito	nagifa	na-hu
malaka		bəher
		la-yosef
		wa-yəbel-
		həqāna
		lə-
		wa-tanṣe-a
		wa-'əm-o
		'əsrā'el

Automatic linguistic annotation

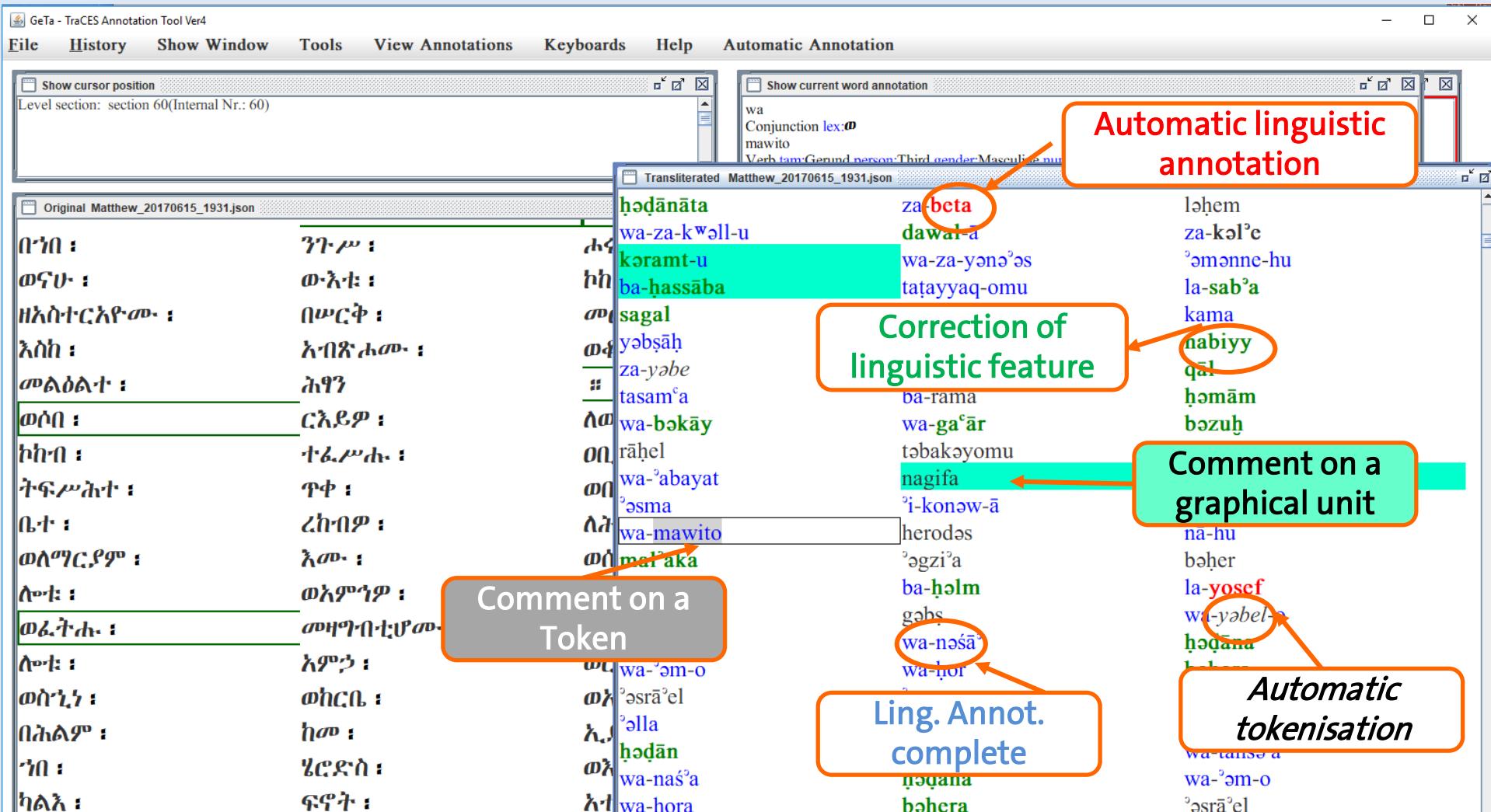
Correction of linguistic feature

Comment on a graphical unit

Comment on a Token

Ling. Annot. complete

Automatic tokenisation




GeTa – Edition of base text

GeTa - TrACES Annotation Tool Ver4

File History Show Window Tools View Annotations Keyboards Help Automatic Annotation

Show cursor position
Level part: part 1(Internal Nr.: 1)
Level chapter: Buch 1.1.0(Internal Nr.: 1.0)
Level sentence: S 1.1.0.2(Internal Nr.: 1.0.2)

Original AmdaSeyon.json

ሰሰመ :	አብ :	ውወልድ :
ወመንፈስ :	ቍኩለ :	ዶአምግለሁ :
ንግድና :	በረድኬት :	አማካኬ :
አየሰሳ :	ከርስተ :	ነይ :
ወመዋ :	ዘንበ :	አማካብኬ :
በ[አ]ደዋሁ :	ለጥም :	ገም :
ንግ-ው :	አትሞክ :	ወሰመ :
መንግሥት :	ገብ :	ሙስቀ :
በዚቃዎች :	መንግሥት :	አም :
ኋጋ :	ወቅመ :	ምስረት :
እኩዎች		ንግድና :
ዘን :		
ንትአመ :	መክር :	አን :
ወመላድ :		ንግ :
ቍኩለ :		
ኋነሥሥ :		
ሠነ :		
ሪቡ :		
ኋመቦ :		
ለይስኬ :		
አማካብኬ :	ዘይሁ :	አከለ :
በሰራ :	ወአደተቋሚ :	ወንበኬ :
ንግሥሥ :	ረዳኬት :	አምኑ :
ወመላድ :	ወመንፈስ :	ቍኩለ :
ከመ :	የጊዜ :	ለዓለመ :

Word Structure

```

graph TD
    Neg(neg) --> Not1[አብ]
    Neg --> NegP1[ይ]
    Not1 --> NotA[አ]
    Not1 --> NotB[ብ]
    NegP1 --> NegKw[ኩ]
    NegP1 --> NegLe[ለ]
  
```

Show current word annotation
la
Preposition state:Nominal state lex:**አ**
kwell
Pronoun of Totality Base case:Nominative lex:**ተ**
u
Pronominal Suffix person:Third gender:Masculine number:Singular lex:**ሁ**

Transliterated AmdaSeyon.json

ba-səma	'ab	wa-wald
wa-manfas	qəddus	1- ² amlāk
nəşəhəf	ba-rad' ² eta	'əgzi' ² ə-na
'iyasus	krəstos	hayla
wa-mawi'a	za-gabra	'əgzi' ² abəher
ba-'ədawi-hu	la-'āmda	səyon
nəguśa	'ityopyā	wa-səma
mangəst-u	gabra	masqlal
ba-10-wa-8-'āmata	mangəst-u	'əm-za
nagśa	wa-'āmata	məhratə-hi
5-100-10-wa-7	.	nəşəhəf-ke
zanta	maşhafa	'ənza
nət'amman	'aba	radā' ² c
wa-walda	nāzāze	wa-manfas
qəddusa	nənza	'ənza
naħħaśšəs	'əm-qəddəst	ba-kama
šəllāse	ba-	hawārəyā
yəbe	ħabba	la-ṭebab
'əmma-b-o	za-yəhub	wahābi
la-yəs'al	wa-i-yət' ² eyyar	la-k <small>w</small> ell-u
'əgzi' ² abəher	ba-səfuḥ	wa-nəħna-ni
haħba	nəħħaś	'əm-' ² ab
ħabba	wa-wald	qəddus
za-yəhub	wa-manfas	qəħħa
wa-i-yət' ² eyyar	kama	la-'ālāma

Modify Token

Token	kw	o	1	1
.	Insert e	Delete e		
Y	Geminate	Remove Gemination		
Z	<input checked="" type="radio"/> Global	<input type="radio"/> Local	Modify	Cancel

Implementation and Sustainability -1-

- Data -model

- Transparent for the user
- New Tags or annotation levels can be easier added
- Adaptable to other scripts (South-Arabian)
- Adaptable to other languages (e.g. Maya, European languages)

Original RIE185_OhneEnter.txt		Transliterated RIE185_OhneEnter.txt	
ń1š	ń1š	czn	mlk
Ń)ŃΨŃ	Ń)ŃΨŃ	Ńksmm	wđhmrm
ŃXŃΠŃ	ŃXŃΠŃ	wrydnm	whbštm
ŃΨŃħŃ	ŃΨŃħŃ	wsb'ṁ	wslḥm
ŃħŃ	ŃħŃ	wŞymm	wksm
ń1š	ń1š	wbgm	mlk
ńŃ	ńŃ	mlkn	bn
ńŃŃXŃħŃ	ńŃŃXŃħŃ	mħrmm	d'ytmw'
ŃŃŃ	ŃŃŃ	Ńdrrm	ħdbm
ńŃ	ńŃ	bgm	fnn
ŃŃŃ	ŃŃŃ		



Implementation and Sustainability -2-

Data encoding

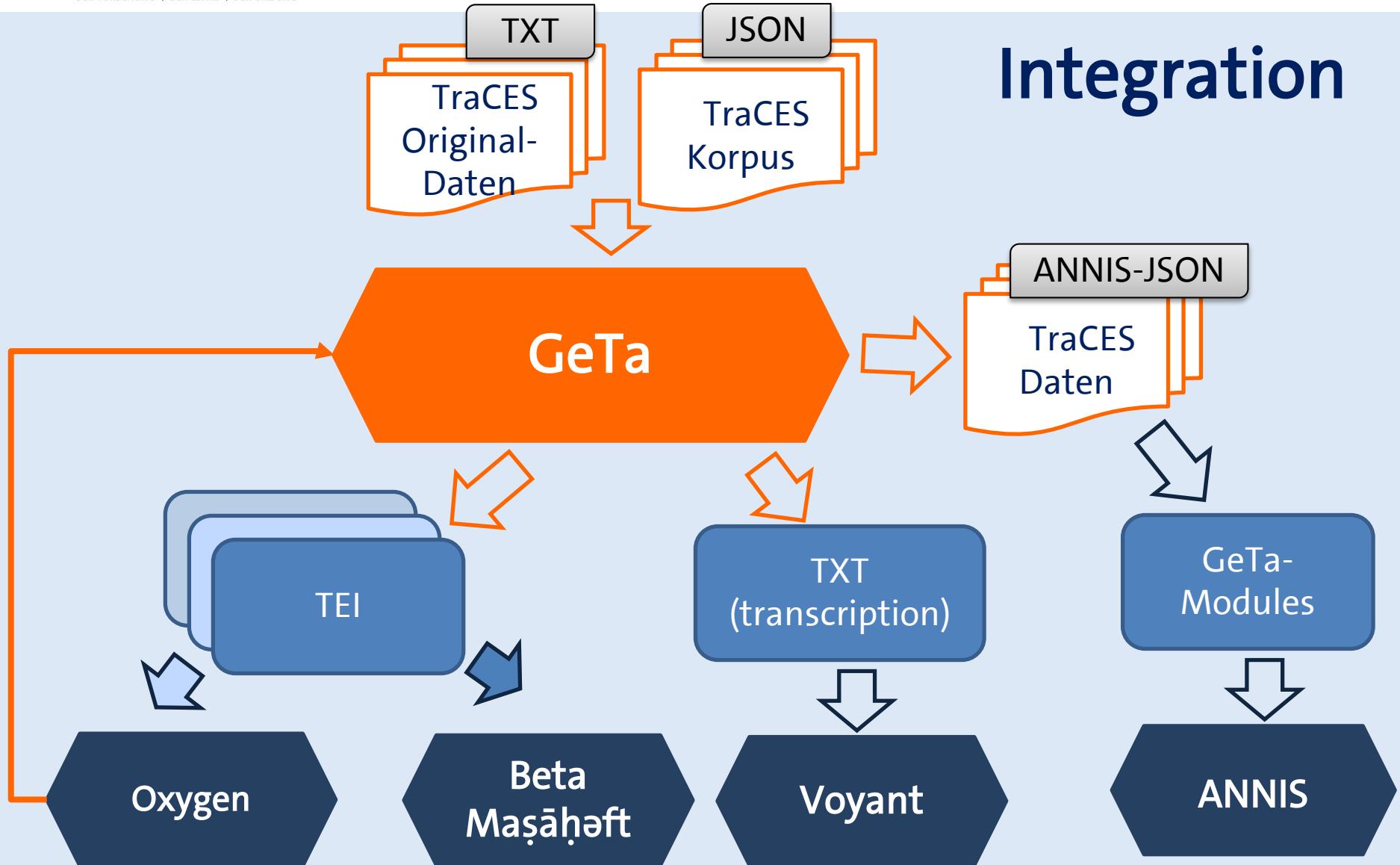
- JSON-Format
- Export to TEI (based on feature structures), ANNIS, TXT

Software

- Open source (CC BY 2.0); requires Java 1.8
- Client application (most part of the functions can be used off-line)
- Tested on Windows7, Windows 10 and Ubuntu platforms (not yet on Apple)
- No third-party software required



Integration



Integration with ANNIS -1-

Help us make ANNIS better!

Base text ▾

1 / 692 ▶▶ Displaying Results 1 - 10 of 6911

Result for: TOKL=/(haba|kama|za)/ & NT=/(
left context: 5

9 ⓘ Path: EpEusebGeTa 18_04_20_17_17_32_>EpEusebGeTa%2018_04_20_17_17_32_EA (tokens 900 - 913)

Həbə: Həmə: Əhə:

FIDED	Həbə:	Həmə:	Əhə:
TR	zi'ā-hu	za-kama	sahafa

grid (anno)

FID	Həbə		Həmə		Əhə		
FIDEDh	Həbə:		Həmə:		Əhə:		
TR	zi'ā-hu		za-kama		sahafa		
FIDLET	h	U:	H	h	ə:	ə	ə:
FIDLETED	h	U	H	h	ə	ə	ə
TRFID	'ā	hu	za	ka	ma	sa	ha
TOKL	zi'ā	hu	za	kama		sahafa	
NT	Possessive Pronoun Base	Pronominal Suffix	Relative Pronoun	Conjunction		Verb	
gender	Masculine	Masculine	Masculine			Masculine	
lex	Həbə	U:	H	həmə		əhə	
lexh	Həbə	U	H	həmə		əhə	
number	Singular	Singular	Singular			Singular	
person		Third				Third	
tam						Perfect	
ne	false	false	false	false		false	
NRI	1						
NR	1						
LE	1						
NA	sentence						
CR	Annotator						
tok	ā	h	u	z	a	k	a

analytic view (FIDED+NT)

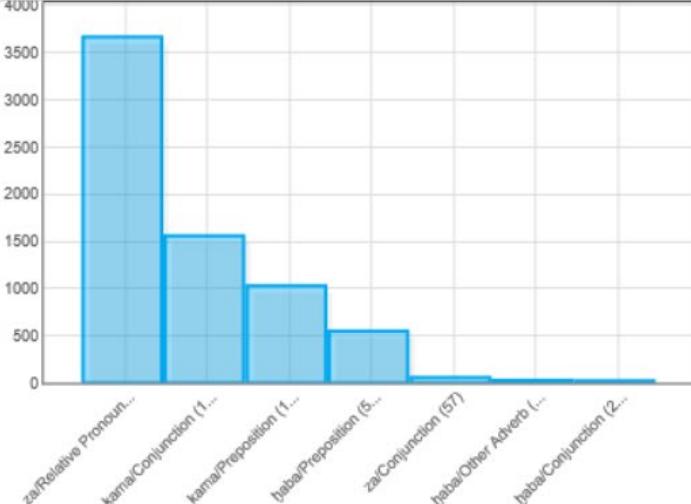


Integration with ANNIS -2-

Help/Examples
EpEusebGeTa 18_04_20_17_17_32_
Frequency Analysis
Query Result

linear scale
 log₁₀ scale

[New Analysis](#)



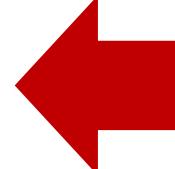
Category	Frequency
za/Relative Pronoun...	~3600
kama/Conjunction (1...	~1550
kama/Preposition (1...	~1050
haba/Preposition (5...	~550
za/Conjunction (57)	~50
haba/Other Adverb (...	~10
haba/Conjunction (2...	~5

7 items with a total sum of 6911 (query on EpEusebGeTa 18_04_20_17_17_32_, ExportAkalKrestosGeTaVer18_04_18_23_32_08, ExportAmdaSeyonVer18_04_19_01)

rank	#1 TOKL	#2 NT
4	haba	Preposition
6	haba	Other Adverb
7	haba	Conjunction
2	kama	Conjunction
3	kama	Preposition
1	za	Relative Pronoun
5	za	Conjunction



Contents

- Digital Humanities
 - Shallow annotation and retrieval
 - Light semantic annotation and retrieval
 - Ontologies and semantical retrieval
- 



Languages for Semantic Annotation

- XML – “shallow”-Language for structuring the data. The information about the structure gives almost no information about the meaning.
- XML Schema –describes the rules for XML-tags, (before XML-DTDs).
- RDF – is a Data model for objects and relations between objects.
- RDFS – describes features and relations between abstract RDF-Objects.
- OWL is a specific and complete language for the representation of ontologies.

Ontologies vs. simple XML-Annotation

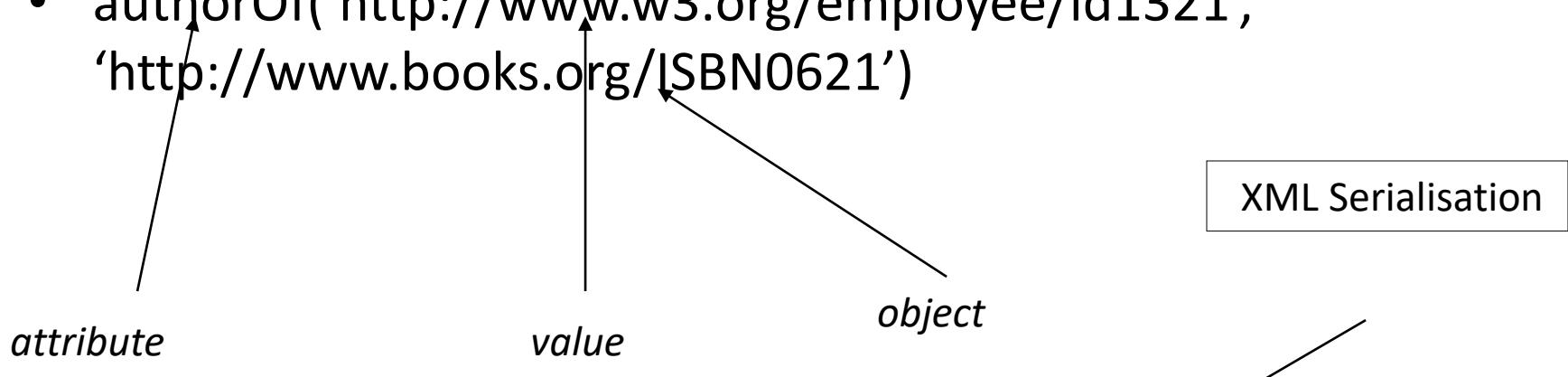
Why XML is not enough:

- XML can be used for
 - Standard encoding and data-annotation,
 - Syntax-definition for Mark-up-Languages),
 - Separation of Layout and Contents of a Web site
- XML cannot describe
 - The Interpretation (semantic Evaluation) of texts,
 - Ontological features (class/subclass, instance, respectively relations between classes).



The RDF Model (Resource Description Framework)

- Is a W3C -Recommendation for standard representation of metadata and content
- Each block in RDF is a triple <object-attribute-value> („subject“, „predicate“, „object“):
- authorOf('http://www.w3.org/employee/id1321',
'http://www.books.org/ISBN0621')



```
<rdf:Description rdf:about="http://www.w3.org/employeeee/id1321">  
  <authorOf rdf: ressource="http://www.books.org/ISBN0621"/>  
</rdf:Description>
```

XML Serialisation

RDFS (RDF Schema)

- Describes the main elements which are needed to represent basic ontological knowledge
- Allows classes, subclasses, properties und sub-properties class, subPropertyOf, subClassOf, etc.

```
<rdfs: Class rdf: about="HardCover">  
<rdfs: subClassOf rdf:resource="#Book"/>  
</rdfs: Class>
```

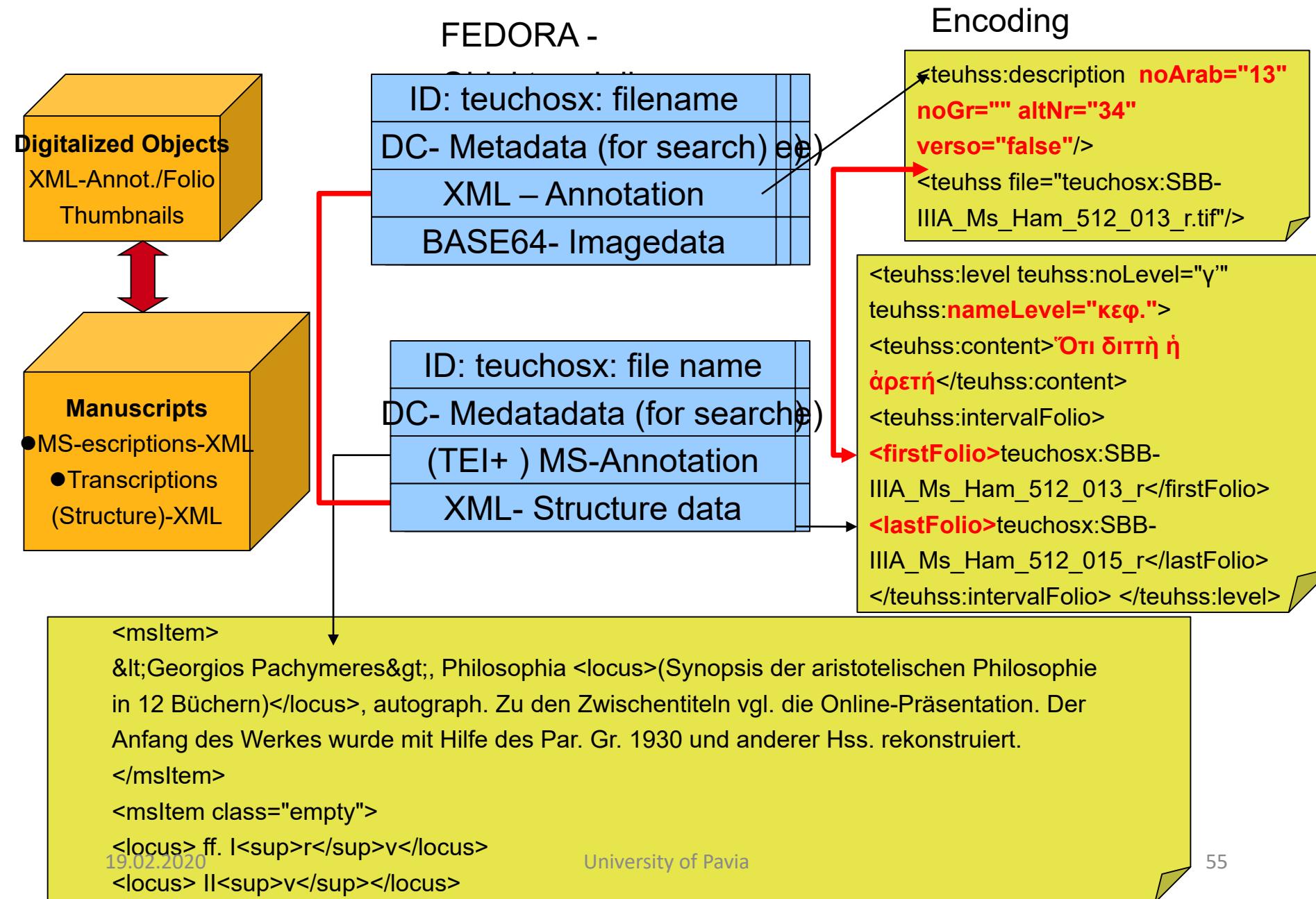
- RDFS cannot be used for modelling complex ontologies as it does not support:
 - Multiple inheritance
 - Specific data types
 - One to many relationships

Use of RDF in CMS

Example Fedora-Objekt

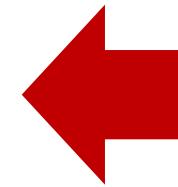
Fedora Object ID
DC - Data stream (informations about Author, date of publication, encoding type)
RDF - Data stream (informations about relation with other objects a well as semantic description of the current object)
XML - Data stream (codicological information)
XML - Data stream (linguistic information)
XML -Data Stream (layoutinformation)
TEXT -Data Stream (transcription text)
TIFF -Dta Stream (image)

Use of Fedora –CMS to index objects



Contents

- Digital Humanities
- Shallow annotation and retrieval
- Light semantic annotation and retrieval
- Ontologies and semantical retrieval



Ontologies and semantical annotation of texts

- Linking an historical dokument content with an ontology.
- Example: CIDOC-CRM Ontology
- Developped by ICS-FORTH- Heraklion-Kreta
 - (http://cidoc.ics.forth.gr/comprehensive_intro.html)
- Goal of CIDOC-CRM: Formal description of historical and/or art objects in museums

Protocol of Proceedings of Yalta Conference



Allied Leaders at Yalta

“The following declaration has been approved:

The Premier of the Union of Soviet Socialist Republics, the Prime Minister of the United Kingdom and the President of the United States of America have consulted with each other in the common interests of the people of their countries and those of liberated Europe. They jointly declare their mutual agreement to concert... and to ensure ...”



E39 Actor



E52 Time-Span

February 1945

E53 Place

Yalta

participated in

within

took place at

E39 Actor



E7 Activity

Crimea Conference

refers to

E38 Image



E39 Actor



carried out

falls within

E65 Creation Event

*

within

has created

E52 Time-Span

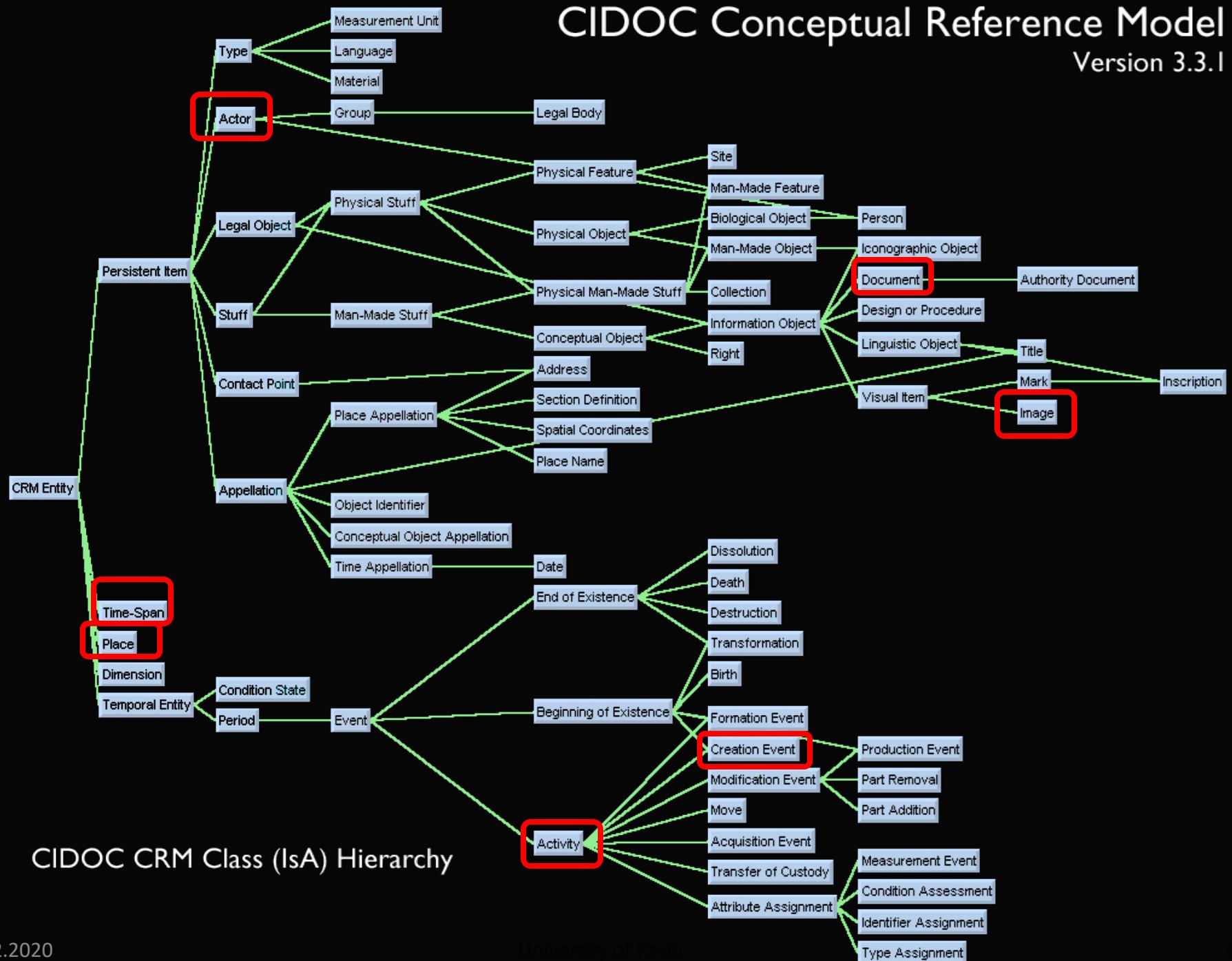
1945-02-11

E31 Document

“Yalta Agreement”

CIDOC Conceptual Reference Model

Version 3.3.1





OWL - 1 -

- Web Ontology Language – a language for representation publication and exchange for ontologies
- A serialisation is done often in RDF/XML
- Elements
 - Taxonomical Relations between classes
 - Data types
 - Data type properties
 - Object properties (relations between objects)
 - Individuals: Instances of Klassen and properties

OWL - 2 -

- Knowledge base (KB)= Set of OWL-propositions& inferences
- Types of OWL:
 - OWL Lite
 - OWL DL
 - OWL Full

OWL-Header

- Header contains
 - Name spaces(rdf, rdfs, owl, dc, ...)
 - Comments
 - Version control (lite, ...)
 - Links to other ontologies

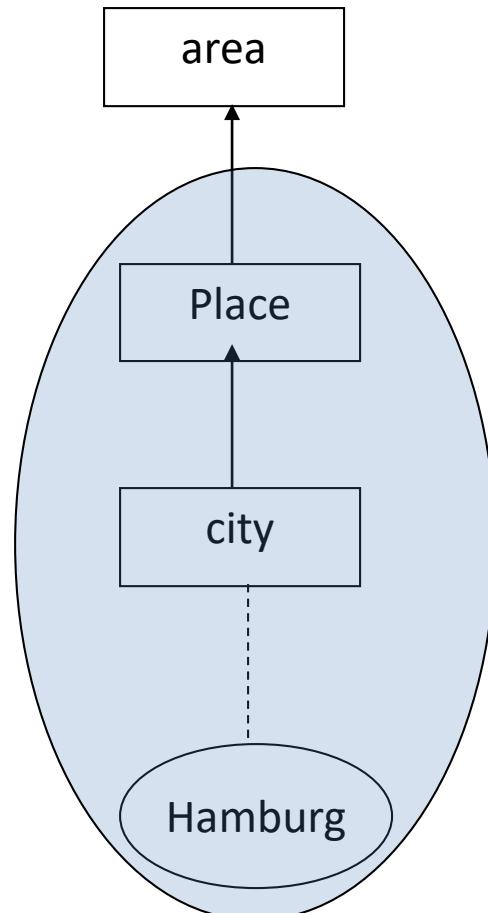
Very important for parsing, and inferences (e.g. map correctly the name spaces)



OWL Basic Definitions -1-

- Each element has an unique identifier rdf:ID
- Class hierarchy
 - Class, subClassOf (rdfs)
`<owl:Class rdf:ID="area"> ...`
 - Implicit root class owl: Thing
- Individual : class instances (concrete assertion of values)

Example of Classes and Individuals



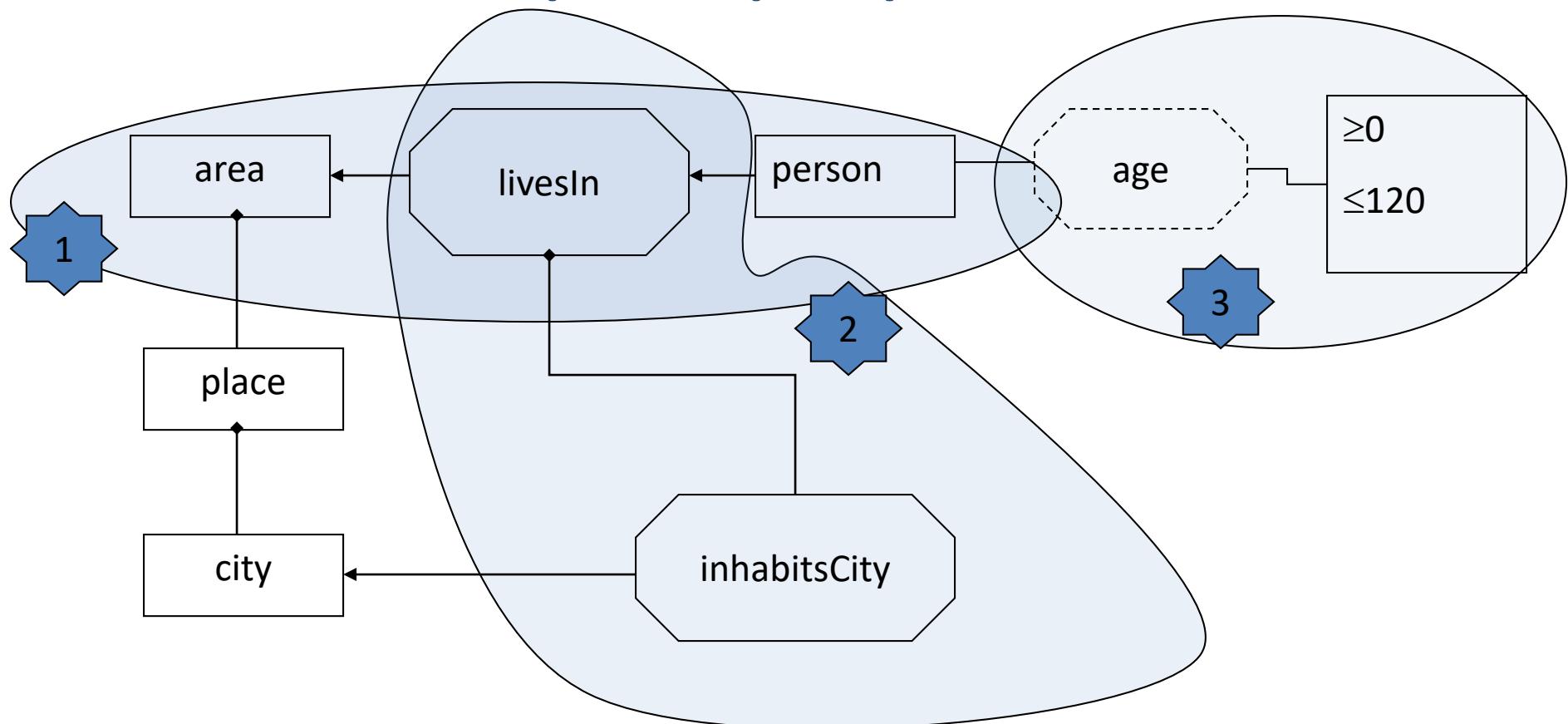
```
<owl:Class rdf:Id="Place"/>
```

```
<owl:Class rdf:ID="city">
  <rdfs:subClassOf rdf:resource="#Place"/>
</owl:class>
<city rdf:ID="Hamburg"/>
```

OWL Basic – Definitions-2-

- Definitions of Properties
 - Object properties
 - **ObjectProperty** (hierarchical)
 - Domain (class restriction)
 - Range (values , objects)
 - **subPropertyOf** (rdfs)
 - Rdf:resource (restricted object property)
 - Range
 - Data type properties
 - **DataTypeProperty**
 - Domain (instance restrictions)
 - Range (values, strings or simpleXML-Data types)

Example- Graph representation



OWL-Formalisation

```
<owl:ObjectProperty rdf="wohntIn">  
  <rdfs:domain rdf:resource="#Person"/>  
  <rdfs:range rdf:resource="#Gebiet"/>  
</owl:ObjectProperty>
```

1

```
<owl:ObjectProperty rdf:ID="bewohntStadt">  
  <rdfs:subPropertyOf rdf:resource="#wohntIn"/>  
  <rdfs:range rdf:resource="Stadt"/>  
</owl:ObjectProperty>
```

2

```
<owl:DatatypeProperty rdf:ID="Alter">  
  <rdfs:domain rdf:resource="#Person"/>  
  <rdfs:range rdf:resource="&dt;Alter">  
</owl:DatatypeProperty>
```

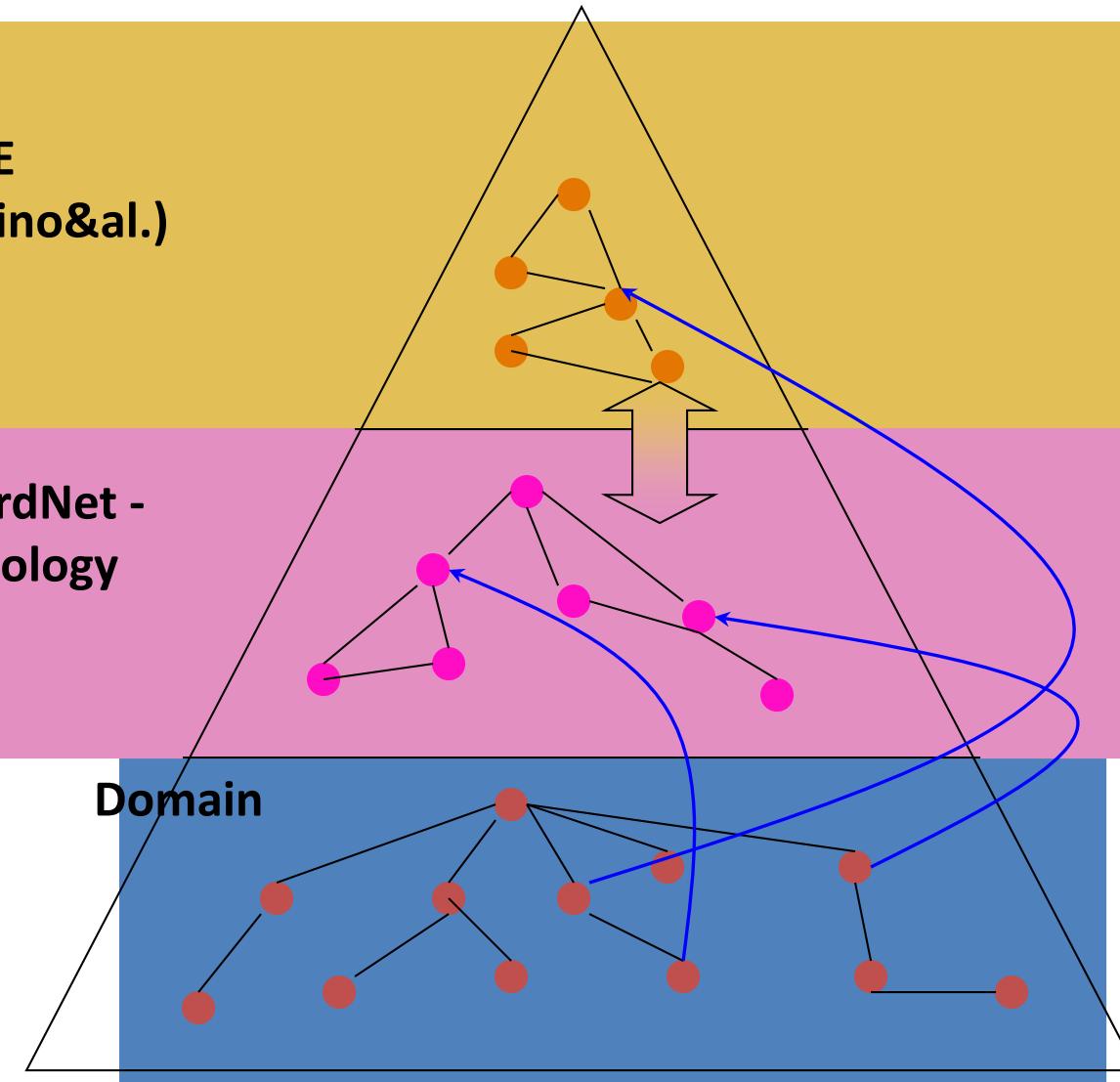
3

Domain & upper ontologies

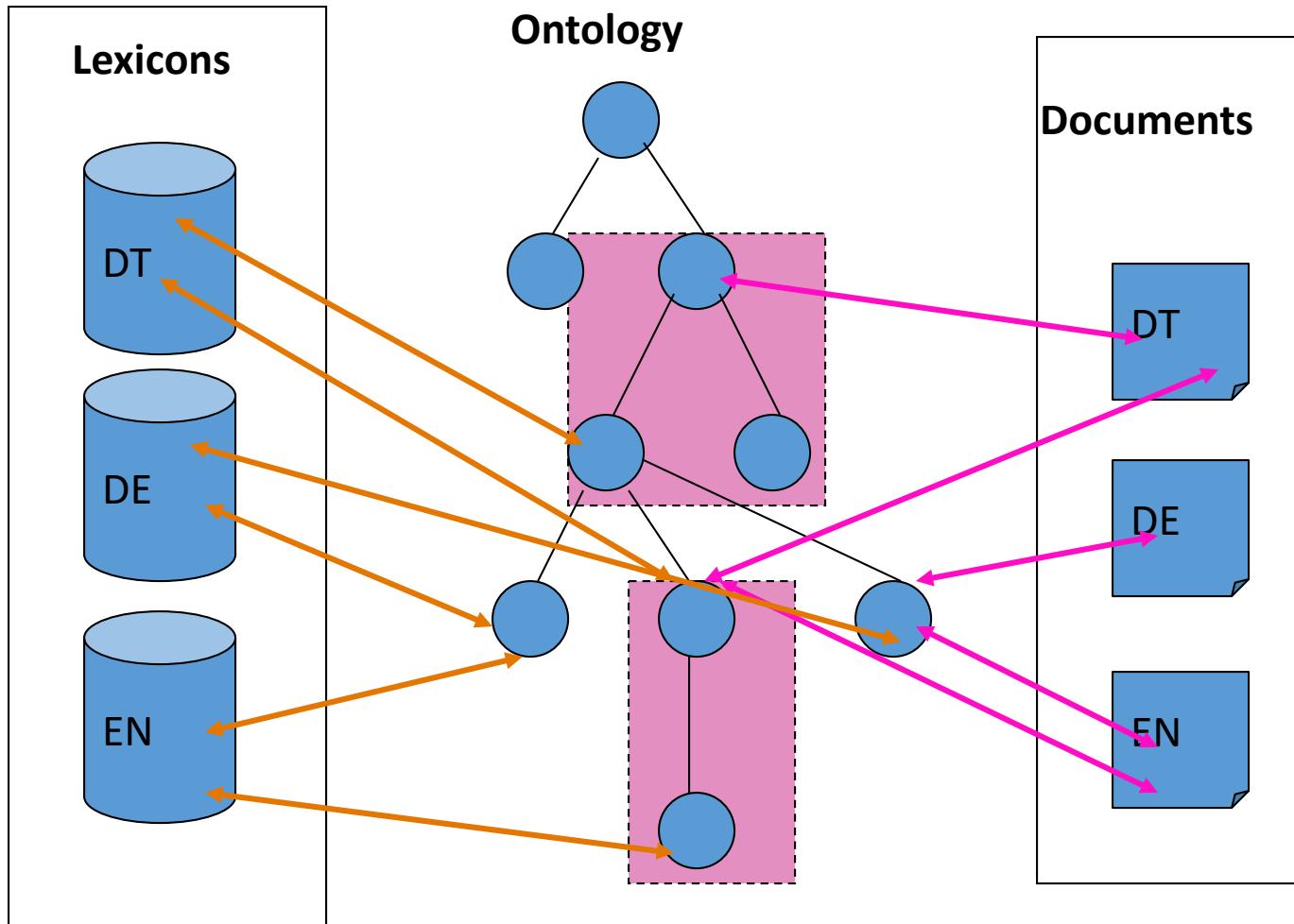
DOLCE
(Guarino&al.)

WordNet -
Ontology

Domain



Ontologies and Lexicons



Ontologies und Lexicons -Example

```
<entry id="id543">
  <owl:Class rdf:about="http://www.lt4el.eu/CSnCS#Presentation">
    <rdfs:comment>Presentation is the process of presenting the content of a topic to an audience.</rdfs:comment>
    <rdfs:comment>Equal OWN: http://www.loa-cnr.it/ontologies/WordNet/OWN#PRESENTATION__PRESENTMENT__DEMONSTRATION</rdfs:comment>
    <rdfs:comment>Equal WN20: ENG20-00496521-n</rdfs:comment>
    <rdfs:comment>ID: id1307</rdfs:comment>
    <rdfs:subClassOf>
      <owl:Class rdf:about="http://www.loa-cnr.it/ontologies/WordNet/OWN#SHOW_1"></owl:Class></rdfs:subClassOf></owl:Class>
    <def>Presentation is the process of presenting the content of a topic to an audience.</def>
    <termg lang="de">
      <term shead="1">Darstellung</term>
      <term>Präsentation</term>
    </termg>
  </entry>
```

HerCoRe – Hermeneutic and Computer based Analysis of Reliability, Consistency and Vagueness in historical texts

- Illustrated through two main works of Dimitrie Cantemir-



Funded by



VolkswagenStiftung

April 2017 – March 2020

„Mixed Methods in Humanities“

Combine hermeneutic approaches and methods from computer science for investigating reliability and consistency of original text from 18th century as well as their translations

H

Compare for the first time “original” with translations made in the 18th- 19th century

(In)Validate assumptions about source quotations in original texts

CS

Demonstrate how to include vagueness and imprecision in annotations and interpretation engines

Extend research in automatic recognition of vague natural language expressions



Dr. Cristina Vertan, UHH
Project coordination,
DH, CL, CS



Prof. Dr. Walther v. Hahn,
UHH
Vagueness, CL, DH
German Linguistics,



Alptug Güney, UHH
Turcology

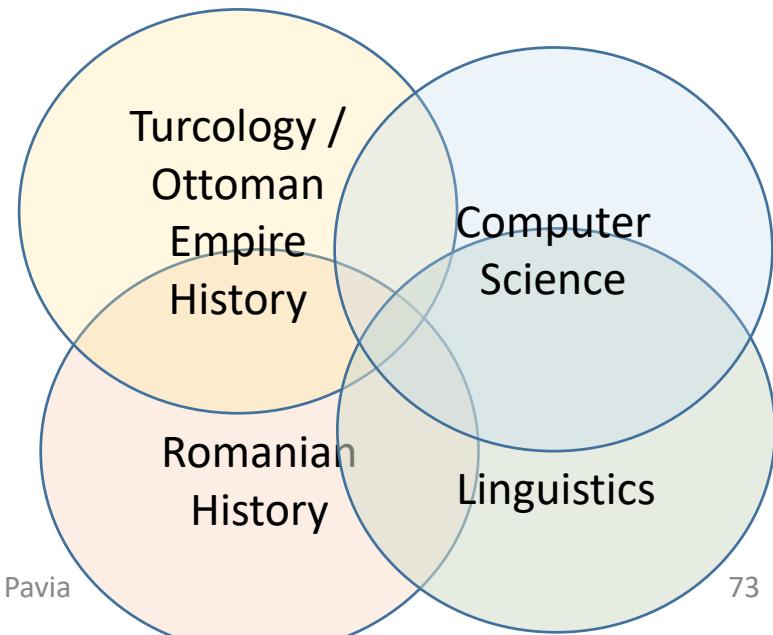
Rafael Quiros Marin,
Miguel Pedegrosa, UHH
Web application



Prof. Dr. Ioana Costa, UB
Cantemir Translations,
Classical philology



Prof. Dr. Yavuz Köse, Univ.
Wien
Turcology



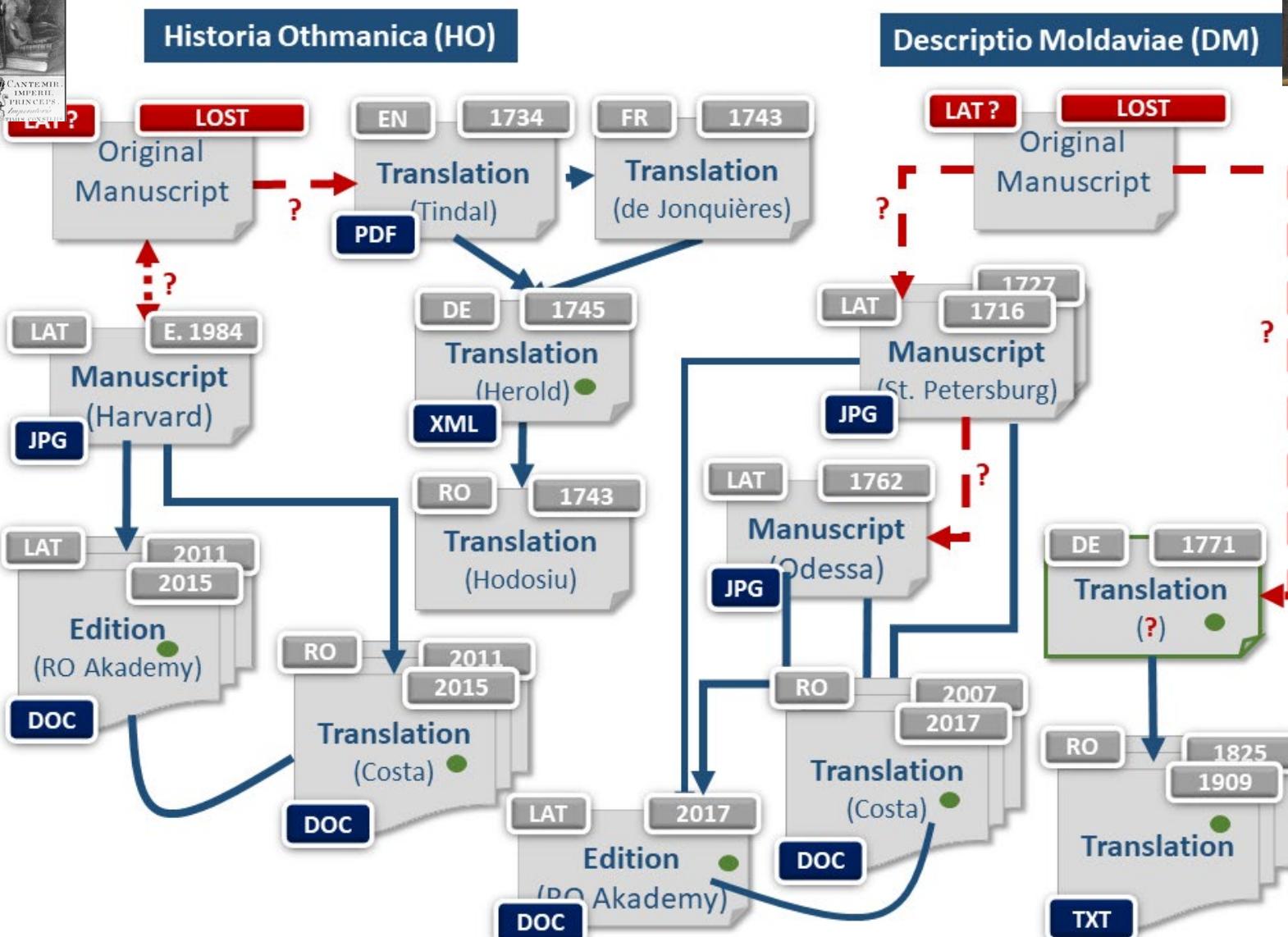
Dimitrie Cantemir (1673 -1723)

- Prince of Moldavia (historical province) as well as „universal“ humanist (linguist, ethnographer, musicologist, historian, writer)
- As member of the Royal Academy in Berlin and at the request of this institution wrote two works :
 - Description of his own country („*Descriptio Moldaviae*“)
 - History of ottoman empire (History of Growth and Decay of Ottoman Empire)
- Original material written in Latin; Both originals were lost already by the end of 18th century
- Several copies were used as basis for translations into German, English (Tindal), French, Russian and later in Romanian
- Sometimes the translation relies on other translation (e.g. first Romanian translation of “*Descriptio Moldaviae*” was done after the German version from 1774.

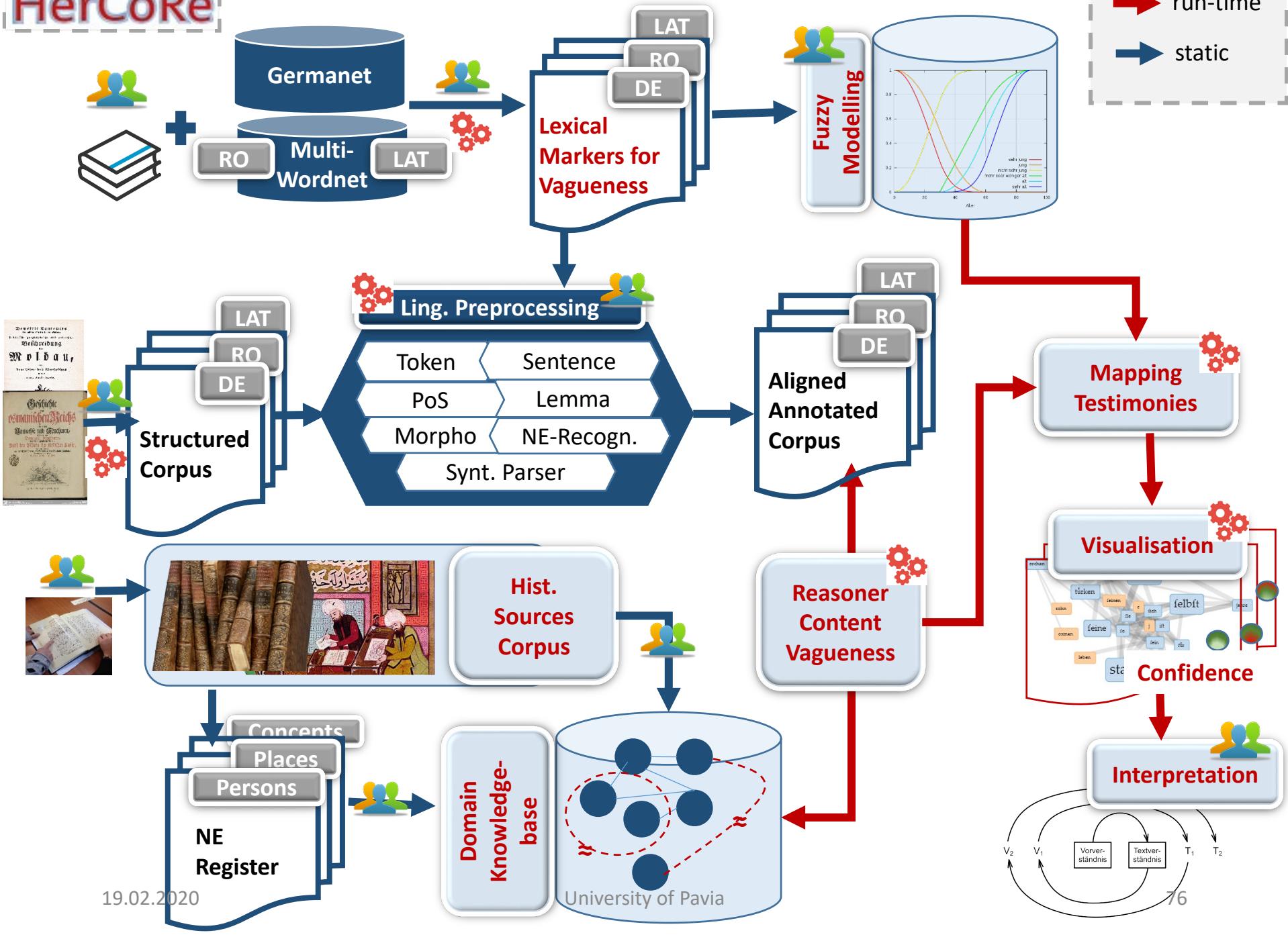


These translations used as reference information about the Ottoman Empire and Romanian provinces until the middle of 19th century, i.e. they give an idea about the reception about this part of the world in Western Europe.

Manuscripts, editions, translations- unclear tradition



→ run-time
→ static



Hermeneutic Approach for Reliability and Content Vagueness -1-

Evaluation of text samples:

Written sources: which references were quoted and how?

*Dumbrava Rofbie or Red Groves, because planted and water'd with Polish Blood. He reduc'd all the Cities between Leopolis and Moldavia. He fought two Battles with Bajazet Ildirim; in the first he was Victor, and in the second he routed Him with a memorable slaughter, which seven vast piles of Turkisb Bodies erected after the Battle, witnessed, by the Confession of **Hezarfenn** himself, the faithful Turkish Historian. He subdued Wallachia as far as Bucurestia, and made*

HO –English
Translation
pp. 47

Hezarfent Hüseyin
(?-1691), *Tenkih üt-Tevarih* [Mid.
C13th-1672/73]

19.02.2020

Explicit oral references: referred persons?

So on the contrary, the most learned Turk, Saadi Effendi (to whom alone I am indebted for my Turkish learning) being one day asked by me, how he, being a great Mathematician and vers'd in the Democratican Philosophy, could believe, that Mabomet broke the Star of the Moon, and caught half of it falling from heaven, in his sleeve. He replied, " that indeed in the course of nature it could not be done, nay was contrary to it, but as this miracle is in the Koran

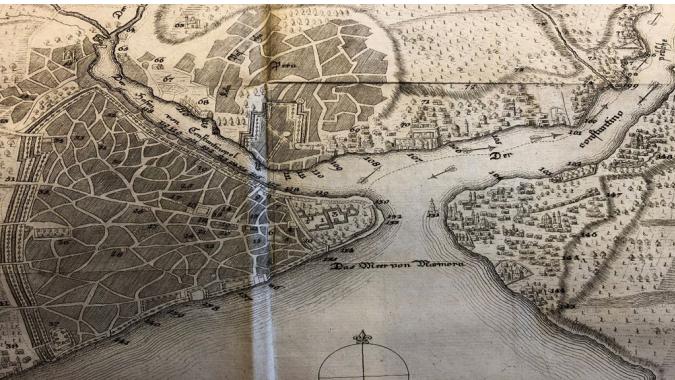
HO –English
Translation
pp. 31

Yanyalı Esad Efendi,
Turkish and Islam
teacher of. D. Cantemir

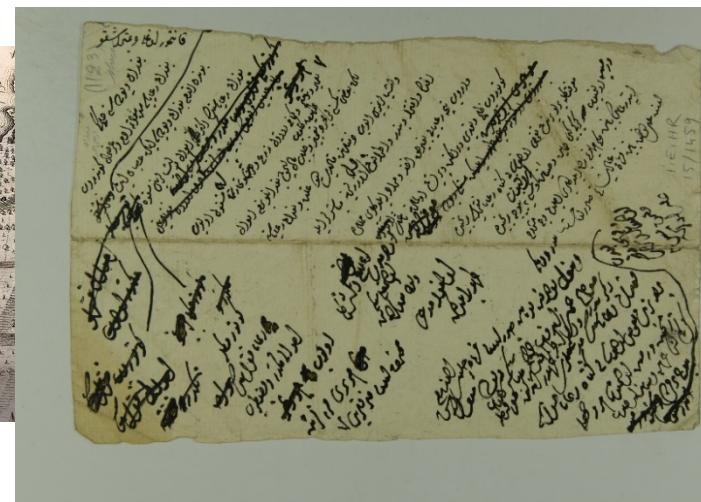
University of Pavia

Hermeneutic Approach for Reliability and Content Vagueness -2-

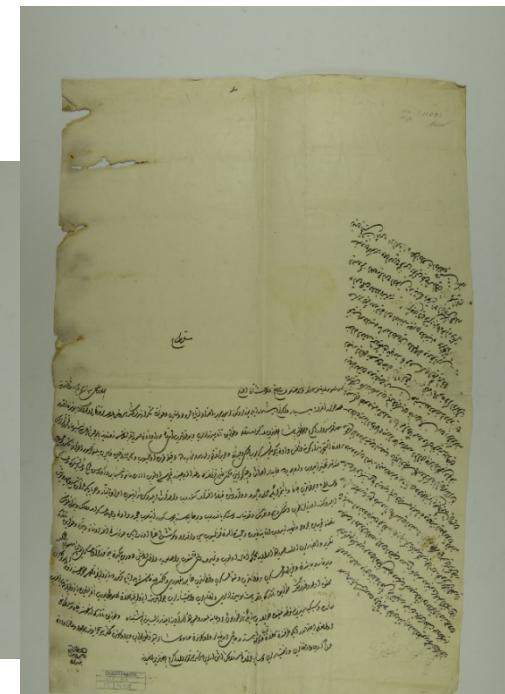
- Research in Ottoman and Patriarchal archives in Istanbul on Cantemir:
 - Additional written sources which might have been available to Cantemir
 - Possible implicit references: Study on Cantemir's life in Istanbul and on his network:
 - Ottoman official documents on Cantemir family and contemporary Wallachian and Moldavian ruler families
 - Neighbourhood of Cantemir's palace in Istanbul



Istanbul map in EN
and DE translations
of HO
19.02.2020



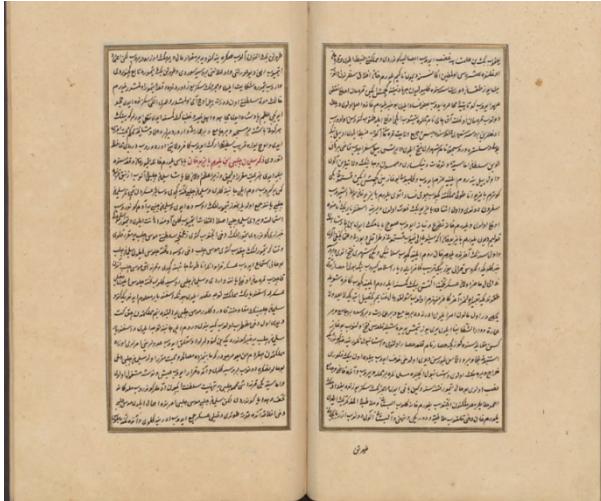
Letter of complaint on D.
Cantemir from Moldavia
University of Pavia



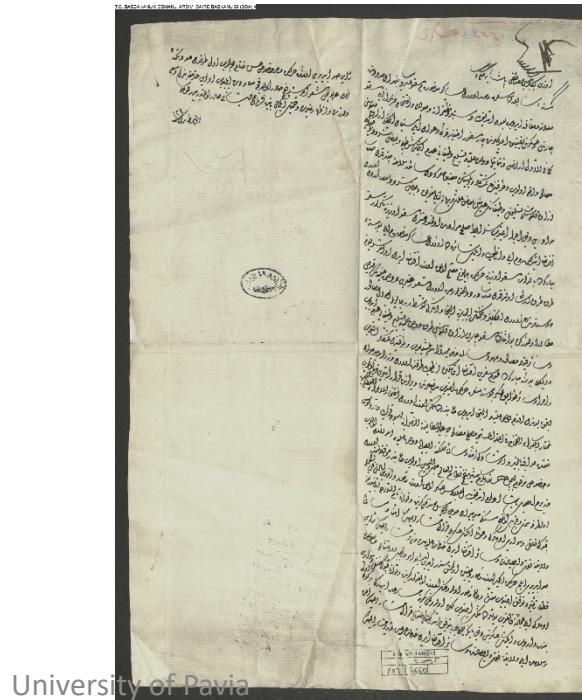
Report of C. Cantemir (father of
D. Cantemir) to Istanbul
78

Corpus of external sources in HerCoRe

- 40 Turkish –Ottoman sources about Ottoman history
- 15 books in other languages (e.g. Latin, Byzantine Greek) quoted by D. Cantemir on diverse topics
- Approx. 50 persons as possible direct network of Cantemir



Hezarfen Hüseyin (?-1691), *Tenkîh üt-Tevârih*
19.02.2020



Official Document
about the place of
D. Cantemir palace

Corpus Challenges

+ - 1:1 fit 62%

596

Ottomanische Geschichte

Prophet Muhammad den smaischen Mönchen gegeben hatte"), und sagte: diese Sertierung von Muhammed sei untergeschoben; oder, wenn sie auf so würden dadurch nur die smaischen Mönche allein, und nicht alle Mönche, vom Tribut freigesprochen. Damit auch außerdem die durch eine ungleiche Erhebung nicht beschworen werden möchten: so er das Chazarisch in den Abtheilungen, und legte den Reichern jenen Löwenthaler zu bezahlen, denen von der mittleren Gottinge sich Armen nur drei Thaler. Alles Gelb, das die Vorfahre aus abe Andacht der Nachkommenchaft hinterlassen und dem Schatz der verschammi vermachte hatten, brachte derselbe in die öffentliche Schatzkammer.

* Man sehe oben, 250 S. 50 Anm.

dieselbe auf drei gewisse Gebungen feststeht, nämlich Aya, Esra und Edra; mit dem Gescheh, daß die Reichen zehn Löwenthaler, die von mittleren Beemigen seien, und die Kriegeren drei dergleichen Thaler² zu bezahlen hätten: bis welcher Einrichtung es auch bis auf den heutigen Tag gehörte ist. Damit aber die Kaiser von ihrem abendländischen Volk nicht das Ansehen haben mögeln, als wenn sie in diesen grauenhaften Erzeugungen dem vorzüglichsten Geiste des Kuruns zusammehandeln: so bestimmen sich die verschamigen Mönchhauser eines geprägten Drei wandes, die Bestie des Dorfs mit den Gedanken ihres Propheten zu vereinigen. Erflich haben sie das Volk überredet: der Preis aller Lebenseinfachheiten sei damals weit geringer, und der Wert des Goldes und Silbers viel höher gewesen, als heutiges Tand; daher müßt das Reich ihrer Vorhersezen, da dasselbe zum Nutzen des omanischen Reichs gegeben werden, also ausgelegert werden, daß jede Person, die sich nicht zu der mohammedanischen Religion bekennet, jährlich soviel Gelb

* das ist, bis oberste, mittlere und niedrigste Ordnung.
** 10 Löwenthaler sind 1 Reichsthaler, 14 Gr. 6 Pf. 6 Löwenthaler sind 5 Reichsthaler, 12 Gr. 7 Löwenthaler sind 3 Reichsthaler, 18 Gr. 6 Pf. nach sächsischem Gelde.
† heißt eigentlich das Böse; Uerfen aber das Billige, Rathfame.

bezahlen folle, als zureiche, dasjenige dafür zu kaufen, was man zur selbigen Zeit um dreyzehn Drachmen lauteres Silber bekommen können. Ferner, damit sie den Leuten alle Gelegenheit abschneiden möchten, sich dagegen zu sperren, haben dieselben erfunden, einen Unterschied zu machen zwischen Scher- ran und Uerfen³: das ist, zwischen denen Sachen, die der Kuron nach dem göttlichen Willen, daß sie geschehen sollten, geboten und den Richtern als eine Regel vorgeschrieben habe; und solchen Sachen, welche der höchsten Gewalt zur Anordnung überlassen worden, darinnen zu beschließen und zu befehlen, was das Beste des Stats erfordere: mit beyge- fügter Lehre, daß man beyden gleichen Gehorftam schuldig sey, und beyderley Geboten mit gleicher Bereitwilligkeit nachleben müsse. Wenn daher ein Christ oder Jud durch ein Fetwa fraget; wie viel Tribut er jährlich zu bezahlen schuldig sey: so wird er zur Antwort bekommen; nach dem Gefetze des Kurons dürfe er nicht mehr, als dreyzehn Drachmen lauteres Silber, bezahlen.

* Man sehe oben, 250 S. 50 Anm.

* das ist, die oberste, mittlere und niedrigste Ordnung.

^{2*} 10 Löwenthaler machen zu Con-

stantinopel 9 Reichsthaler, 5 Gr. 6 $\frac{6}{13}$ Pf. 6 Löwenthaler sind 5 Rei-

$\frac{7}{13}$ Pf.

und 3 Löwenthaler 2 Reichsthaler, 18 Gr. 5 $\frac{7}{13}$ Pf. nach sächsischem

^{3*} Scherran heißt eigentlich das Böse; Uerfen aber das Billige, Rathfame.

Ferner, damit sie den Leuten alle Gelegenheit abschneiden möchten, sich dagegen zu sperren, haben dieselben erfunden, einen Unterschied zu machen zwischen Scher- ran und Uerfen³: das ist, zwischen denen Sachen, die der Kuron nach dem göttlichen Willen, daß sie geschehen sollten, geboten und den Richtern als eine Regel vorgeschrieben habe; und solchen Sachen, welche der höchsten Gewalt zur Anordnung überlassen worden, darinnen zu beschließen und zu befehlen, was das Beste des Stats erfordere: mit beyge- fügter Lehre, daß man beyden gleichen Gehorftam schuldig sey, und beyderley Geboten mit gleicher Bereitwilligkeit nachleben müsse. Wenn daher ein Christ oder Jud durch ein Fetwa fraget; wie viel Tribut er jährlich zu bezahlen schuldig sey: so wird er zur Antwort bekommen; nach dem Gefetze des Kurons dürfe er nicht mehr, als dreyzehn Drachmen lauteres Silber, bezahlen. Wenn * Man sehe oben, 250 S. 50 Anm. * das ist, die oberste, mittlere und niedrigste Ordnung. ^{2*} 10 Löwenthaler machen zu Con- stantinopel 9

Reichsthaler, 5 Gr. 6[FORMEL] Pf. 6 Löwenthaler sind 5 Reichsthaler, 12 Gr. 11[FORMEL] Pf. und 3 Löwenthaler 2 Reichsthaler, 18 Gr. 5[FORMEL] Pf. nach sächsischem Gelde. ^{3*} Scherran heißt eigentlich das Böse; Uerfen aber das Billige, Rathfame.

illa Graeci Imperii periodus fuit, in anno quo acciderit, gravissimi dissentiunt scriptores, aliis illam ad annum 1452., aliis ad 1453. referentibus. Vnde colligi potest, quanti in reliquis Turcicae historiae capitibus, *v(erbi) g(ratia)* Sultanorum diebus natalibus, emortualibus etc. definiendis, commissi fuerint errores. His ut occurramus, haud inconsultum duximus, collatis Christianis Turcisque autoribus, de hoc computo paulo accuratius inquirere, eoque ipso / probare recte nos in ea, quam iam edimus, historia annos Hegirae cum aera Christiana conciliasse. Non vero opus esse existimamus de nomine „Hegirae“ <...> item movere, atque disquirere, utrum ea a fuga Falsoprophetae Mecca Medinam, an ab ipsius obitu (ut n. Muhammedanis placet) ducat initium.

Imp. Oth., p. 6

9 Huius amnis antiquum nomen investigare non potuimus *mg. A, in textu B, C / 10 Ante Suczava, del.* In hunc se exonerant *A / 11 ubi...Metropoli: urbs...Metropolis p. m. A / 12 Moldova] Moldava p. m. A; Post Moldova, in textu cuius nominis (nomen p. m.) rationem supra explicavimus et supra a quo tota feudis regio nomen dicit *del. A, postea infra* rationem supra explicavimus Cap. 1mo *alia manus* scripsit; supra Moldava, *signum inserenda notae, ulterius deletae, exstat. Haec nota ab ipsa manu Cantemirii scripta est*: de cuius nomine, licet incerto authore, talis apud Moldavos fertur fabula: Dragoszum (*-um ulterius add.*), Bogdani filium (filius p. m., *inde de quo* in priori libro fuse dictum est *del.*), cum 300 saltim hominibus, venantis specie, ex Marmarusio (Maramorusio DS), Transylvaniae regione montium transitum tentare constituisse, hoc in itinere casu invenisse bovem sylvestrem, Moldavis ‘Dzimbru’ dictum et, dum eum persequitur, ad montium radices descendisse. Porro, cum catula quaedam venatica, quam prae caeteris diligebat, ‘Molda’ dicta, fortius ferae instaret, aestuans fera (*sic*), in profluentem se proiecisse et telis ibi confectam. Canem vero, quae in ipsis aquis venatum quaeviserat fugientem, rapidis fluvii undis absorptam (*post absorptam, del. fuisse A*) in huius (*post huius, del. me A*) itaque memoriam, fluvium ‘Moldovam’ a Dragosso Principe appellatum fuisse. Loco etiam, ubi haec acciderunt, suae gentis nomen, ‘Roman’ (quae civitas Bonfinio ‘Forum Romanorum’ est), indidisse. / 13 Transylvaniae *p. m. C //**

C20 1 et montibus B, *supra et lineatum, s.m. ex corr.* / 2 Bistriza] Bistri...a *p. m. A / 3 Post* miscentur, *em. (s(ive) miscetur) B / 4 Molnitzal* Molniza *p. m. A / 5 Valeniagra]* Valenagra *p. m. A; Valeneagra B; Valenjagra ex Valeneagra C //*

Descr. Mold. p. 1134
A5, B11, C19

- Abbreviations
- Special editorial marks
- Changes (insertions or removals) done by editors
- Mix of original text from different editions
- Mix of original text and editorial comments

Implications for the workflow

Additional work on preprocessing:

- Separating the texts into:
 - Main content from Cantemir
 - Cantemir's annotations
 - Translator's annotations and
 - Editorial mark-ups

and adding links among them;

- Cleaning unuseful editorial mark-ups (for the purposes of the project);
- Adapting the Latin text (e.g. expansion of abbreviations).

Example: Against whom and where fought Sultan Bayezid -1-

Stephan, Fürst von Moldau, war der größte Kriegesheld seiner Zeit. Er überwand den berühmten König in Ungarn, Matthias, und nahm ihm die siebenbürgischen Alpen ab, die noch heutiges Tages die Grenzen der Moldau an der Westseite ausmachen. Mit Bajeßid Jildirim hielte er zwey Treffen in dem ersten war er Ueberwinder; und in dem andern schlug er denselben nach einer recht merkwürdigen Niederlage aus dem Felde, und ließ nach der Schlacht sieben große Haufen von den Leichnamen der Türken aufwerfen, welches selbst der redliche türkische Geschichtschreiber Heßarfenn*, durch sein Geständniß bezeugt

<Stefan> <rulerOf><Moldavia>
<Matthias><rulerOf><Hungaria>
<Stefan> <fightWith><Matthias>
<Bayeßid Yildirim> ???
<Stefan> <fightWith><Bayeßid Yildirim>



Looks like reliable as it is attested also in the Chronik of Heßarfenn

Example: Against whom and where fought Sultan Bayezid -2-

The screenshot shows the OntoGraf interface with the following details:

- Top Navigation:** Active Ontology, Entities, Object Properties, Individuals by class, DL Query, OntoGraf, SWRLTab, Debugger, SQWRLTab.
- Left Panel:** Class hierarchy under PersonConcept. Nodes include: Yayalar, Yörükler, Zeamet, Şehzade, OttomanState, PersonConcept (highlighted in blue), Ambassador, Archduke, Author, Bey, Cleric, Despot, Diplomat, Duke, Emir, Emperor, Envoy, Historian, Kapikulu, Khan, King, Poet, Prince, Ruler, Archduke, Bey, Despot, Duke, Emir, Emperor, Khan, King, Prince, Shah, Sultan.
- Middle Panel:**
 - DL query:** Ruler and hasName value "Bajeßid Yıldırım"^^xsd:string
 - Buttons:** Asserted (dropdown), Execute, Add to ontology.
 - Query results:** Instances (1 of 1) - Bayezid_I.

Example: Against whom and where fought Sultan Bayezid -3-

The screenshot shows the OntoGraf interface with the following details:

- Top Navigation:** Active Ontology, Entities, Object Properties, Individuals by class, DL Query, OntoGraf, SWRLTab, Debugger, SQWRLTab.
- Left Sidebar (Class hierarchy):** PersonConcept is selected. Other classes listed include Yayalar, Yörükler, Zeamet, Şehzade, OttomanState, PersonConcept, Ambassador, Archduke, Author, Bey, Cleric, Despot, Diplomat, Duke, Emir, Emperor, Envoy, Historian, Kapikulu, Khan, King, Poet, Prince, Ruler, Archduke, Bey, Despot, Duke, Emir, Emperor, Khan, King, Prince, and Shah.
- Middle Panel (Query Area):** The status is "Asserted". The query entered is "Prince and fightWith min 1 Ruler". There are "Execute" and "Add to ontology" buttons.
- Right Panel (Query Results):** The results are titled "Query results" and show "Instances (3 of 3)": Bogdan_III., Mircea_I., and Stephen_III. Each result is preceded by a purple diamond icon.

Example: Against whom and where fought Sultan Bayezid -4-

The screenshot shows the OntoGraf interface with the following details:

- Top Bar:** Active Ontology, Entities, Object Properties, Individuals by class, DL Query, OntoGraf, SWRLTab, Debugger, SQWRLTab.
- Left Sidebar:** Class hierarchy for PersonConcept, listing various roles and titles such as Yavalar, Yörükler, Zeamet, Sehzade, OttomanState, PersonConcept, Ambassador, Archduke, Author, Bey, Cleric, Despot, Diplomat, Duke, Emir, Emperor, Envoy, Historian, Kapikulu, Khan, King, Poet, Prince, and Ruler.
- Middle Panel:** DL query input field containing: Prince and fightWith some (Ruler and hasName value "Matthias Corvinus"). Buttons: Execute (highlighted), Add to ontology.
- Right Panel:** Query results section showing Instances (1 of 1) with the result: Stephen_III.

Example: Against whom and where fought Sultan Bayezid -5-

Which Ruler was contemporary with Bazezid Yildirim ?

The screenshot shows a SPARQL query interface with several tabs at the top: overlapsInt1, overlapsInt2, SQWRL Queries, OWL 2 RL, overlapsInt1, overlapsInt, and overlapsInt2. The overlapsInt2 tab is active.

The results table has two columns: p and her:Ruler(?p). The p column lists: her:Sigismund, her:AحمدCelayir, her:İsfendiyar, her:Tahirten, and her:Mircea_I.. The her:Ruler(?p) column lists: her:Bayezid_I., her:Bayezid_I., her:Bayezid_I., her:Bayezid_I., and her:Bayezid_I..

A red box highlights the query string in the results table:

```
her:Ruler(?p) ^ her:hasTimeInterval(?p, ?TI) ^  
her:refersTimeBegin(?TI, ?TB) ^ her:refersTimeEnd(?TI, ?TE) ^  
her:hasValue(?TB, ?x) ^ her:hasValue(?TE, ?y) ^ her:Ruler(?p1) ^  
her:hasName(?p1, ?s) ^ swrlb:contains(?s, "Yildirim") ^  
her:hasTimeInterval(?p1, ?TI1) ^ her:refersTimeBegin(?TI1, ?TB1)  
^ her:refersTimeEnd(?TI1, ?TE1) ^ her:hasValue(?TB1, ?x1) ^  
her:hasValue(?TE1, ?y1) ^ swrlb:greaterThan(?y, ?y1) ^  
swrlb:greaterThan(?x1, ?x) -> sqwrl:select(?p, ?p1)
```

Red arrows point from the highlighted query string to the corresponding results in the table.

Mircea_I. a Wallachian Prince but not Stephen_III.

Example: Against whom and where fought Sultan Bayezid -6-

Was Stephen_III with Matthias Corvinus and whom else?

Active Ontology Entities Object Properties Individuals by class DL Query OntoGraf SWRLTab Debugger SQWRLTab

Name	Query	Comment
S1	her:AboutModifier(?p) ^ her:TimeUnit(?t) ^ her:hasValue(?t, ?x) -> swrlb:add(?x, 10)	
overlapsInt1	her:Ruler(?p) ^ her:hasTimeInterval(?p, ?TI) ^ her:refersTimeBegin(?TI, ?TB) ^ her:refersTimeEnd(?TI, ?TE) ^ her:hasValue(?TB, ?x) ^ her:hasValue(?TE, ?y) ^ her:Ruler(?p1) ^ her:hasName(?p1, ?s) ^ swrlb:contains(?s, "Mare") ^ her:hasTimeInterval(?p, ?TI1) ^ her:refersTimeBegin(?TI1, ?TB1) ^ her:refersTimeEnd(?TI1, ?TE1) ^ her:hasValue(?TB1, ?x1) ^ her:hasValue(?TE1, ?y1) ^ swrlb:greaterThan(?y1, ?y) ^ swrlb:greaterThan(?x, ?x1) -> sqwrl:select(?p, ?p1)	
overlapsInt2		

SQWRL Queries OWL 2 RL overlapsInt1 overlapsInt overlapsInt2

```

graph TD
    P((P)) --- S1[her:AboutModifier(?p) ^ her:TimeUnit(?t) ^ her:hasValue(?t, ?x) -> swrlb:add(?x, 10)]
    S1 --- S2[her:Ruler(?p) ^ her:hasTimeInterval(?p, ?TI) ^ her:refersTimeBegin(?TI, ?TB) ^ her:refersTimeEnd(?TI, ?TE) ^ her:hasValue(?TB, ?x) ^ her:hasValue(?TE, ?y) ^ her:Ruler(?p1) ^ her:hasName(?p1, ?s) ^ swrlb:contains(?s, "Mare") ^ her:hasTimeInterval(?p, ?TI1) ^ her:refersTimeBegin(?TI1, ?TB1) ^ her:refersTimeEnd(?TI1, ?TE1) ^ her:hasValue(?TB1, ?x1) ^ her:hasValue(?TE1, ?y1) ^ swrlb:greaterThan(?y1, ?y) ^ swrlb:greaterThan(?x, ?x1) -> sqwrl:select(?p, ?p1)]
    S2 --- M[Her: MatthiasCorvinus]
    S2 --- S[her:Stephen_III.  
her:Stephen_III.]

```

her:Ruler(?p) ^ her:hasTimeInterval(?p, ?TI) ^ her:refersTimeBegin(?TI, ?TB) ^ her:refersTimeEnd(?TI, ?TE) ^ her:hasValue(?TB, ?x) ^ her:hasValue(?TE, ?y) ^ her:Ruler(?p1) ^ her:hasName(?p1, ?s) ^ swrlb:contains(?s, "Mare") ^ her:hasTimeInterval(?p1, ?TI1) ^ her:refersTimeBegin(?TI1, ?TB1) ^ her:refersTimeEnd(?TI1, ?TE1) ^ her:hasValue(?TB1, ?x1) ^ her:hasValue(?TE1, ?y1) ^ swrlb:greaterThan(?y1, ?y) ^ swrlb:greaterThan(?x, ?x1) -> sqwrl:select(?p, ?p1)

Her: MatthiasCorvinus her Stephen_III
Her:Mehmed_II. her:Stephen_III.

The information on Matthias Corvinus is correct but in the time of Stephen_III., Mehmed_II. and **NOT** Bayezid Yildirim was sultan

Example: Against whom and where fought Sultan Bayezid -7-

Stephan, Fürst von Moldau, war der größte Kriegesheld seiner Zeit. Er überwand den berühmten König in Ungarn, Matthias, und nahm ihm die siebenbürgischen Alpen ab, die noch heutiges Tages die Grenzen der Moldau an der Westseite ausmachen. Mit Bajeßid Jildirim hielte er zwey Treffen in dem ersten war er Ueberwinder; und in dem andern schlug er denselben nach einer recht merkwürdigen Niederlage aus dem Felde, und ließ nach der Schlacht sieben große Haufen von den Leichnamen der Türken aufwerfen, welches selbst der redliche türkische Geschichtschreiber Heßarfenn*, durch sein Geständniß bezeugt

<Stefan> <rulerOf><Moldavia>
<Matthias><rulerOf><Hungaria>
<Stefan> <fightWith><Matthias>
<Bayeßid Yildirim> ???
<Stefan> <fightWith><Bayeßid Yildirim>



Looks like reliable as attested also in the Chronik of Heßarfenn

<Mircea_I> <rulerOf><Wallachia>
<Mircea_I> <fightWith><Bayeßid Yildirim>



Work performed

- Hermeneutic study
- Corpus cleaning
- Digitisation of *Descriptio Moldaviae* (German Edition)
- Linguistic Annotation (Tokenisation, PoS, Lemma, Morpho Analysis)
- Vagueness Lexical Thesaurus in all 3 languages
- Identification of vagueness lexemes in Text
- Data-model (Document –oriented for the material and Graph-oriented for the annotation)
- Models for representation of vague /uncertain time expressions and geo locations
- Domain Ontology (already 460classes, 80relations and 500 axioms more than 2000 individuals)
- Specification of the multilevel annotation tool
- First attempt to visualise geolocations on OpenStreetMaps

Ongoing work

- Completion of the ontology
- Representation of vague concepts and properties
- Linking with external resources
- Implementation of the reasoner and visualisation components
- Hermeneutic interpretation
- Automatic detection of vague expresions and their focus

Technical issues

- OrientDB for document and annotation storage (document and graph-model can be stored and interrogated with 1 database)
- JSON-LD as internal exchange format
- Java Client application
- OWL 2 and Fuzzy OWL as encoding format for the ontology
- Protege, for the moment with Pellet Reasoner (supports both DL - OWL and SQWRL queries)
- Linguistic encoding following the Universal Dependencies Model in CONLL-U format

Thank you very much for your attention!

<https://www.inf.uni-hamburg.de/de/inst/dmp/hercore>

cristina.vertan@uni-hamburg.de