Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Research Group
„Computerphilology“

UNIVERSITATEA DIN BUCUREȘTI
VIRTUTE ET SAPIENTIA

Research Center „Digital Humanities“

HerCoRe – Hermeneutic and Computer based Analysis of Reliability, Consistencs and Vagueness in histroical texts

- Illustrated through two main works of Dimitrie Cantemir-

Funded by VolkswagenStiftung

2017 -2020

Data : Works of Dimitrie Cantemir

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 XVIIIth century
- Several copies were used as basis for translations into German, English, 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 XIXth century, i.e. they give an idea about the reception about this part of the world in Western Europe.

Challenges for the analysis and interpretation of Cantemir‘s works

- Already in the 1920‘ies, it was demonstrated using selections of texts, that the translations are not respecting the original all the time
  - E.g. Information sources indicated by Cantemir were omitted, because they seemed too unreliable to the translator
- In the XX century researchers claimed that some of the sources, persons and facts quoted by Cantemir were not existing
- BUT given the:
  - Geographic distribution of material (originals in libraries in USA and Russia; translations and copies across Europe; most part of the quoted sources in Turkey),
  - The multilingual character of the materials to be investigated (Latin, German, Romanian, English, Turkish at least) and
  - The volume of data which has to be processed in parallel

no study about the reliability and consistency of the original and the translations could be performed until now

Computational methods could help in performing this study

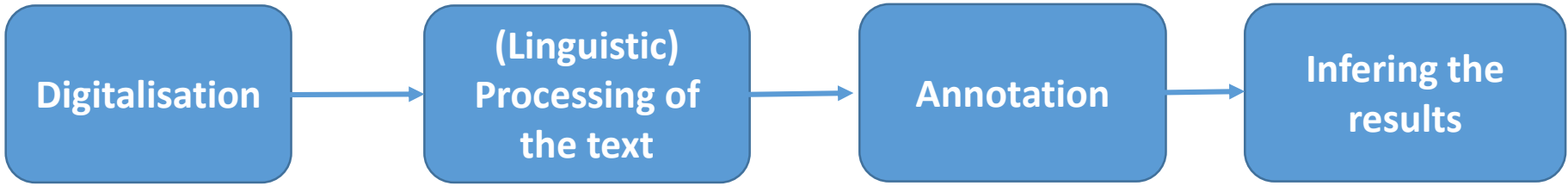
State-of-the-art - Sudies on Cantemir works

- The reception of Cantemir'S works vary significantly depending of epoque and geographical perspective of researchers
- (Lemny 2010) considers that most od the vague or imprecise assertions in Cantemir'S texts were due to lack of previous work about the described regions.
- Franz Babinger in "Aufsätze und Abhandlungen zur Geschichte Südeuropas und der Levante" (1966) argues that some of the sources quoted by Cantemir did nnot even exist. However meanwhile new documents , especially in the Turkey were made available, so Babinger's assertions could be no longer true.
- In "Dimitrie Cantemir's Ottoman History and its Reception in England" (2010) Hugh-Trevor Cooper argues that not the text quality and accuranteness made the work so well-known but the political connections of his son, who represented the Russian empire in London. The argumentation is supported by nomination of other works, claimed to be more accurate than Cantemir's.
- An opposite interpretation is done by M. Leezenberg in "the oriental origins of orientalism: the case of Dimitire Cantemir" (2012) who insists on the Turkish sources to which no other western contemporary of Cantemir had access.

Directions of investigation

- Reliability:
  - Of the original: are the quotations made by Cantemir grounded? Is there a concordance between his degree of trust in these sources and the current knowledge about them (e.g. is there any evidence that a person which Cantemir claims to have spoken to, really lived in that time?)
  - Of the translation against the original; Here an important role have the inserted editorial annotations.
- Consistency:
  - Within the original: keeps Cantemir a constant opinion about persons, events, facts across the text? (see his own annex with annotations vs. the text)
  - Across the 2 originals: Are common persons and events described similarly?
  - Between original and translation: does the translation preserve the degree of vagueness /certainty stated by Cantemir?
- Vagueness
  - Political or tactical reasons for imprecise expressions

Limitations of current computational aproaches used by DH for such study



	Digitalisation	(Linguistic) Processing of the text	Annotation	Infering the results
State of the art	•OCR or •Double keying •Diplomatic transcription (TEI) •Transcription variants are recorded just for digital editions	Existent (blackbox) Pipelines: **Tokenisation->** first **Sentence Recognizer-> Lemmatiser->PoS Annotation** are applied	•Annotation of **crisp** facts (or some uncertainity marked with TEI mechanisms) •If at all, one source of imprecision is marked •Links to ontologies	If at all, **DL crisp Logic**, i.e. even if uncertainity is indicated it is not included in the inference process
Limitations	•OCR with poor performance (Fraktur script, multiling. text.) •Transcription considerend as an interpretation level: Annotation of variants is a must.	•Existent pipelines are not working (too many unknown words) •One needs to decide first which PoS, chunks have to be identified.	Annotation for vague facts / assertions at several levels are required (and beyond TEI standards)	DL crisp logic will reduce to 0 any fuzzy annotatiom -> need of a fuzzy inference mechanism

State-of-the-art - Vagueness representation and annotation

- **Semantic indefiniteness** is a central feature of natural language, can be found in any type of text including specialized ones (vHahn, 1983)
- According to Pinkal semantic indefiniteness can be other vague oder ambiguous.
- Computer linguistic approaches approaches concentrates on handling ambiguity (e.g. use of Wordnet)
- TEI offers (as first mark-up standard) three possibilities: <note>, <certainity> and <precision> but:
  - Overlapping annotations are possible only stand-off which is very difficult in TEI
  - Not all sources of vagueness are allowed to be indicated through <respons> tag
  - <precision> can be specified only for numerical values: problematic for expressions like “some kilometers below the city”
  - No reasoner until now can be applied
- OWL 2 allows representation of fuzzy ontologies
- DELOREAN reasoner (Bobillo et al. 2013) can be used with fuzzy ontologies

Only a combination of mark-up allowing several levels of vagueness and fuzzy reasoner can lead to interpretions about text reliability

State-of-the-art - Computer processing of historical texts

- Most part of applications dealing with processing of historical texts focus on:
  - Digital archiving
  - Diplomatic transcription
  - Automatic collation
  - Text normalisation
  - Shallow annotations of linguistic information including adaptation of existing CL-Tools for modern languages
- Semantic approaches to historical text processing are rare: Averroes and PERSEUS projects use RDF Annotations
- CIDOC-CRM is the most wide-spread ontology to which cultural heritage objects are linked, includes mainly concepts for the meta-data and less for the content
- Thaller (2007) stated that digital texts are not-ambiguous, context-free and contain just the information embedded in the code, whilst historical texts are sequence of symbols, each carrying a meaning, which co-exist in a multidimensional space

No break-through result in exploiting ambiguity and vagueness of historical texts was reported since then

State-of-the art – Digital Resources available

- Cantemir's work about the history of ottoman empire:
  - German translation was digitized and annotated with PoS information by DTA/BBAW, and is available for academic purposes
  - Romanian and Latin Versions are in electronically form available for purposes of the project
- Cantemir's work "Description of Moldavia":
  - German version is partially digitized. Digitzation and annotation will be completed during the project
  - Latin version is available as Wikisource, will be annotated during the project
  - Romanian Version is digitized, and annotated with PoS, and NE associated to a crisp ontology. The digitization and annotation was performed within a previous joint project between the University of Hamburg and University of Bucharest
- Shallow annotation tools for German, Romanian and Latin are as open source available.
- For German we will rely on the CDG-System developed at the university of Hamburg, which is robust to lexical gaps and syntactical variances; thus sentences with untranslated words can be processed easily

Most part of the digital resources are available, or can be acquired with small effort. Thus no important amount of time has to be allocated for digitization /corpus preparation purposes

Der erste demnach, der nach Batia Einfall (*) der Moldau ihren vorigen Glanz wieder verschafft hat, war

1. Dragosch. Obgleich unsre Jahrbücher sein Geschlechtsregister nicht angeben, so ist es doch eine beständige Sage bey uns, daß er aus dem alten königlichen moldauischen Stamme gewesen sey, und den Bogdan zum Vater gehabt habe, welcher ein Sohn des Johannis war, von welchem alle Fürsten den Namen Johannis in ihrem Titel zu führen pflegen; dieser Meinung ist desto mehr Glauben beyzumessen, weil man schwerlich glauben kan, daß einer von gemeiner Herkunft mit einem so großen Gefolge auf die Jagd (welche die Moldau zu entdecken Gelegenheit gegeben,) habe ausgehen, und seine übrigen Landsleute überreden können, ihn zu folgen.

(*) Diese Stelle bestätiget aufs neue, was wir oben schon vermuthet hatten, daß Dragosch erst nach des Tatars Bathy oder Batu Einfall, d.i. ungefähr nach 1250. aus Siebenbürgen ausgewandert ist; vielleicht aber lassen sích beede Meynungen vereinigen, wenn man zwey Auswanderungen annimmt, die eine in der lezten Hälfte des Zwölften, die andere in der ersten Hälfte des dreyzehenten Jahrhunderts (V.)

Dimitrie Cantemir, "Beschreibung der Moldau", Faksimiledruck der Originalausgabe von 1771, Kriterion Verlag Bukarest, 1973, pag. 111

Classical Markup based on crisp ontology

Der erste demnach, der nach Batia Einfall (*) der Moldau ihren vorigen Glanz wieder verschafft hat, war
1. Dragosch. Obgleich unsre Jahrbücher sein Geschlechtsregister nicht angeben, so ist es doch eine beständige Sage bey uns, daß er aus dem alten königlichen moldauischen Stamme gewesen sey, und den Bogdan zum Vater gehabt habe, welcher ein Sohn des Johannis war, von welchem alle Fürsten den Namen Johannis in ihrem Titel zu führen pflegen; dieser Meinung ist desto mehr Glauben beyzumessen, weil man schwerlich glauben kan, daß einer von gemeiner Herkunft mit einem so großen Gefolge auf die Jagd (welche die Moldau zu entdecken Gelegenheit gegeben,) habe ausgehen, und seine übrigen Landsleute überreden können, ihn zu folgen.
(*) Diese Stelle bestätiget aufs neue, was wir oben schon vermuthet hatten, daß Dragosch erst nach des Tatars Bathy oder Batu Einfall, d.i. ungefähr nach 1250. aus Siebenbürgen ausgewandert ist; vielleicht aber lassen sich beede Meynungen vereinigen wenn man zwey Auswanderungen annimmt, die eine in der letzten Hälfte des Zwölften, die andere in der ersten Hälfte des dreyzehenten Jahrhunderts (V.)

Dragosch	belongs	moldavian kings
Dragosch	son_of	Bogdan
Bogdan	son_of	Johannis
Dragosch	has_additional_name	Johannis
Bogdan	has_additional_name	Johannis
Drgaosch	discovered	Moldau
Dragosch	has_acitivity	hunting
Dragosch	has_activity	development
Development	takes_place	after Batia invasion

Dragosch	has_activity	moved
Movement	takes_place	2nd half 12th century
Movement	takes_place	1st half 13th century
Bathy invasion	takes_place	1250
Bathy	has_alternative_name	Batu
Bathy	is_a	Tatar

Enriched Classical Markup based on crisp ontology

Der erste demnach, der nach Batia Einfall (*) der Moldau ihren vorigen Glanz wieder verschafft hat, war
1. Dragosch. Obgleich unsre Jahrbücher sein Geschlechtsregister nicht angeben, so ist es doch eine beständige Sage bey uns, daß er aus dem alten königlichen moldauischen Stamme **gewesen sey,** und den Bogdan zum Vater gehabt habe, welcher ein Sohn des Johannis war, von welchem alle Fürsten den Namen Johannis in ihrem Titel zu führen pflegen; dieser Meinung ist desto mehr Glauben beyzumessen, weil man schwerlich glauben kan, daß einer von gemeiner Herkunft mit einem so großen Gefolge auf die Jagd (welche die Moldau zu entdecken Gelegenheit gegeben,) habe ausgeben, und seine übrigen Landsleute überreden können, ihn zu folgen.

(*) Diese Stelle bestätiget aufs neue, was wir oben schon vermuthet hatten, daß Dragosch erst nach des Tatars Bathy oder Batu Einfall, d.i. **ungefähr** nach 1250. aus Siebenbürgen ausgewandert ist; vielleicht aber lassen sich beede Meynungen vereinigen wenn man zwey Auswanderungen annimmt, die eine in **der letzten Hälfte des Zwölften**, die andere in der **ersten Hälfte des dreyzehenten Jahrhundert** (V.)

Dragosch	≈ belongs	moldavian kings
Dragosch	≈ son_of	Bogdan
Bogdan	≈ son_of	Johannis
Dragosch	≈ has_additional_name	Johannis
Bogdan	≈ has_additional_name	Johannis
Drgaosch	discovered	Moldau
Dragosch	has_acitivity	hunting
Dragosch	has_activity	development
Development	takes_place after	Batia invasion

Dragosch	has_activity	moved
Movement	takes_place	>=1150; <=1200
Movement	takes_place	>=1200; <=1250
Bathy invasion	takes_place	≈ 1250
Bathy	has_alternative_name	Batu
Bathy	is_a	Tatar

HerCoRe - Multiple level fuzzy annotation at document level

Der erste demnach, der nach Batia Einfall (*) der Moldau ihren vorigen Glanz wieder verschafft hat, war
1. Dragosch. Obgleich unsre Jahrbücher sein Geschle es
doch ein r aus
dem
gew
hab
wel
ihre
des
schwerlich glauben kann, daß einer von gemeiner Herkunft mit einem so großen Ge
auf die Jagd (welche die Moldau zu entdeck
Gelegenheit gegeben,) habe ausgehen, und
seine übrigen Landsleute überreden können
zu folge

Legend

Dragosch	most_probable belongs	moldavian kings
Dragosch	discovered	Moldau
Dragosch	has_acitivity	hunting
Dragosch	has_activity	development
Development	takes_place	after Batia invasion

Cantemir

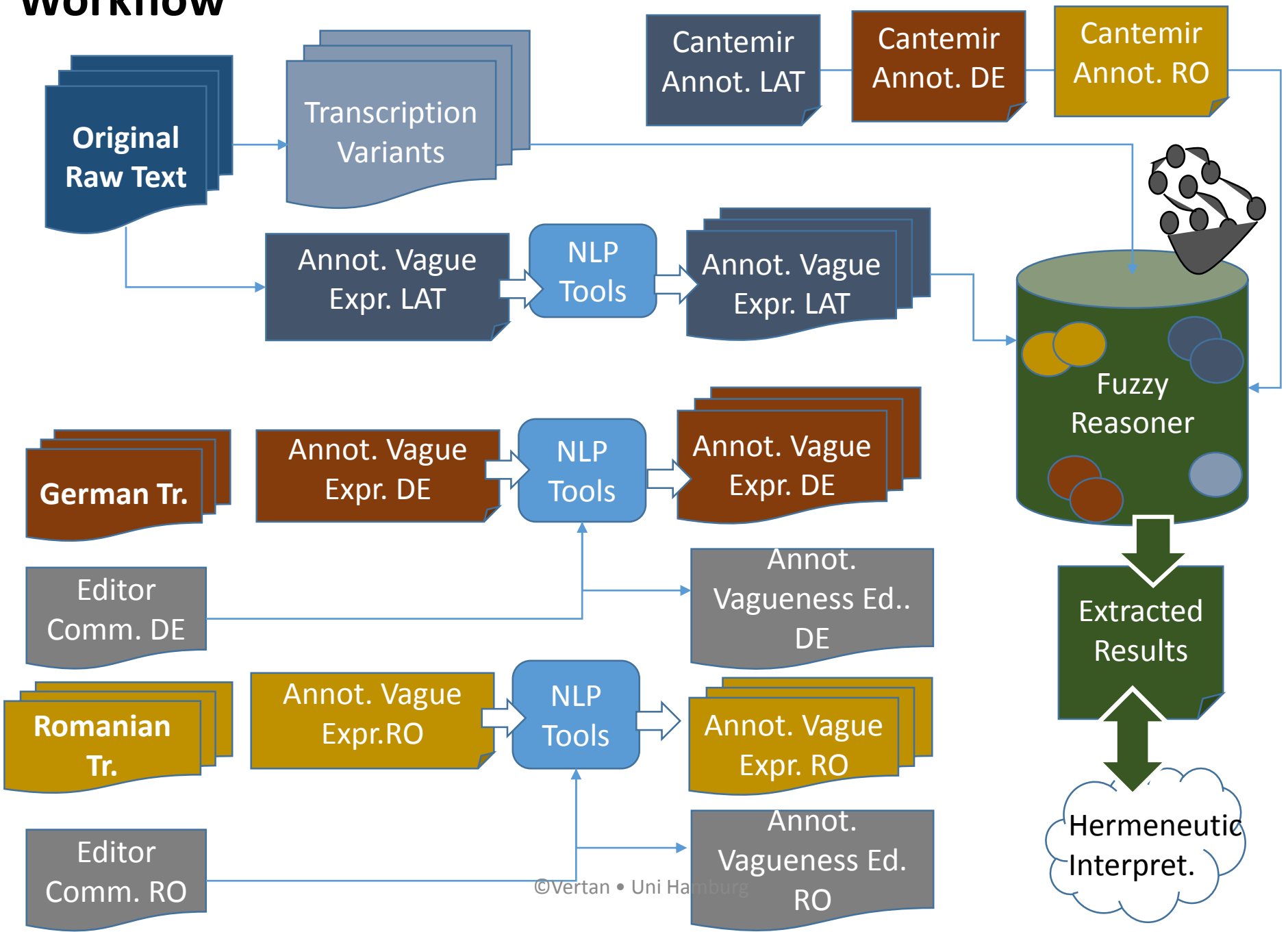
Dragosch	≈ son_of	Bogdan
Bogdan	≈ son_of	Johannis
Dragosch	≈ has_additional_name	Johannis
Bogdan	≈ has_additional_name	Johannis

Translator

Dragosch	has_activity	moved
Movement	takes_place	> ≈ 1150; < ≈ 1200
Movement	takes_place	> ≈ 1200; < ≈ 1250
Bathy invasion	takes_place	≈ 1250
Bathy	has_alternative_name	Batu
Bathy	is_a	Tatar

uthet hatten, daß Dragosch erst nach
athy oder Batu Einfall, D.i. ungefähr
us Siebenbürgen ausgewandert ist;
er lassen sich beide Rechnungen
enn man zwey Auswanderungen
e eine in der letzten Hälfte des
e andere in der ersten Hälfte des
Jahrhundert (B.)

Workflow



Innovative Research

- Humanities : possibility of compare easily multiple multilingual sources; respecification of hermeneutic circle
- CS: reasoning on data annotated at different vagueness layers. This type of data does not occur in traditional CS applications
- Innovative joint work by CS and Humanities Scientists

Traditional Approach

- HR: List of requirements (natural language)
- CSR: System specification
- CSR: System implementation
- HSR: Use of the system;
- CSR: error correction; light improvements of the system

HerCorRE Approach

- STEP I:
  - HR- text introspection and extraction of indicators for vague expressions
  - CSR+HR: system specification
  - CSR: shallow annotation of text
- STEP II:
  - CSR + HR : Fuzzy Ontology and Formalization
  - HR: annotation and first evaluation
  - CSR: ontology and system improvement

Generalization / Development for other research questions, languages ?

- The pilot software application will be developed modular and parts, like extraction of vague expressions, annotation interface will be reusable for other scenarios
- Ontology and vagueness layer are domain dependent, and have to be re-specified for each application
- Best practice documents containing description considered vagueness levels, ontology development will be provided.
- During the project contacts with other disciplines will be realized and explicit adaptation strategies will be discussed.

HerCorRe aims to (re)-integrate part of hermeneutic research into DH. By means of Computer Methods we will model vague facts and assertions, present them to the user, and let the researcher to draw conclusions and lift one or another veil of history

¡Muchas gracias por su atención!

¿Alguna pregunta?