Mining Citations, Linking Texts

Matteo Romanello (EPFL / DAI) @mr56k

L'édition critique à l'ère numérique – Genève 2 March 2016

Overview

My PhD

- **Dissertation**:
  - *From Index Locorum to Citation Network: an Approach to the Automatic Extraction of Canonical References and its Applications to the Study of Classical Texts*.
  - `http://hdl.handle.net/11858/00-1780-0000-002A-4537-A`
- Department of Digital Humanities, King's College London
- **supervisors:**
  - Willard McCarty
  - Shalom Lappin

References in Classics

Canonical References

- Vergil, *Aen.* 12, 101-109
- Thuc. I 89, 1
- Hom. *Il.* 7.180

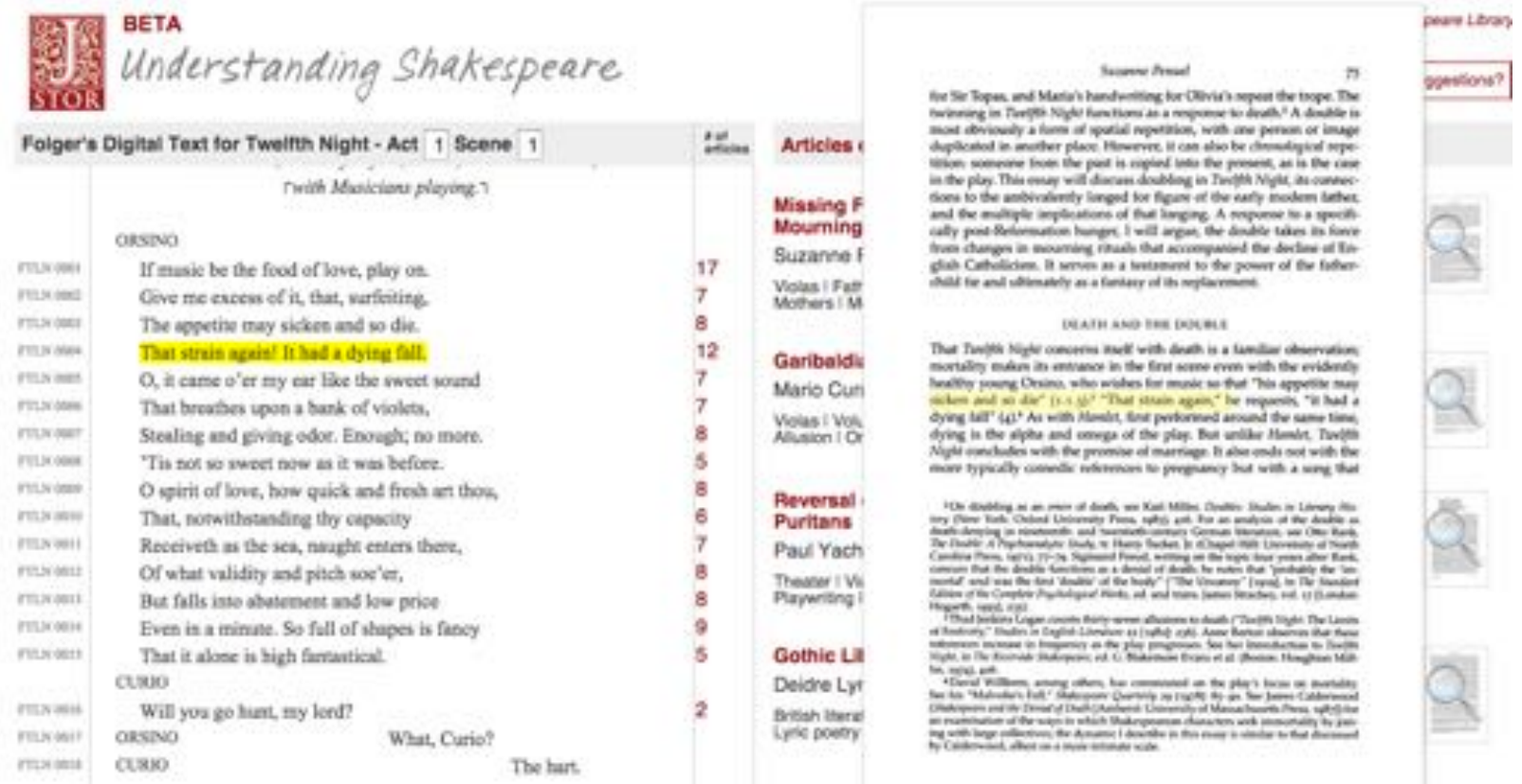
But also references to:

- inscriptions (e.g. CIL 3, 6174; AE 1991, 1405)
- papyri (e.g. PCair. inv. 10750)
- manuscripts (e.g. Vendôme, Bibl. mun. 31)
- fragmentary texts, coins, etc.

Yandle's Line-by-line Bibliography of the *Parzival*

Line-by-line Bibliographical Database of Wolfram von Eschenbach's Parzival, `http://wolfram.lexcoll.com/txts/index.htm`

JSTOR Labs: Understanding Shakespeare



http://labs.jstor.org/shakespeare/

Research questions

1. Is it possible to extract canonical references automatically from modern publications such as journal articles?

Research questions

1. Is it possible to extract canonical references automatically from modern publications such as journal articles?
2. With what level of accuracy can this extraction be performed?

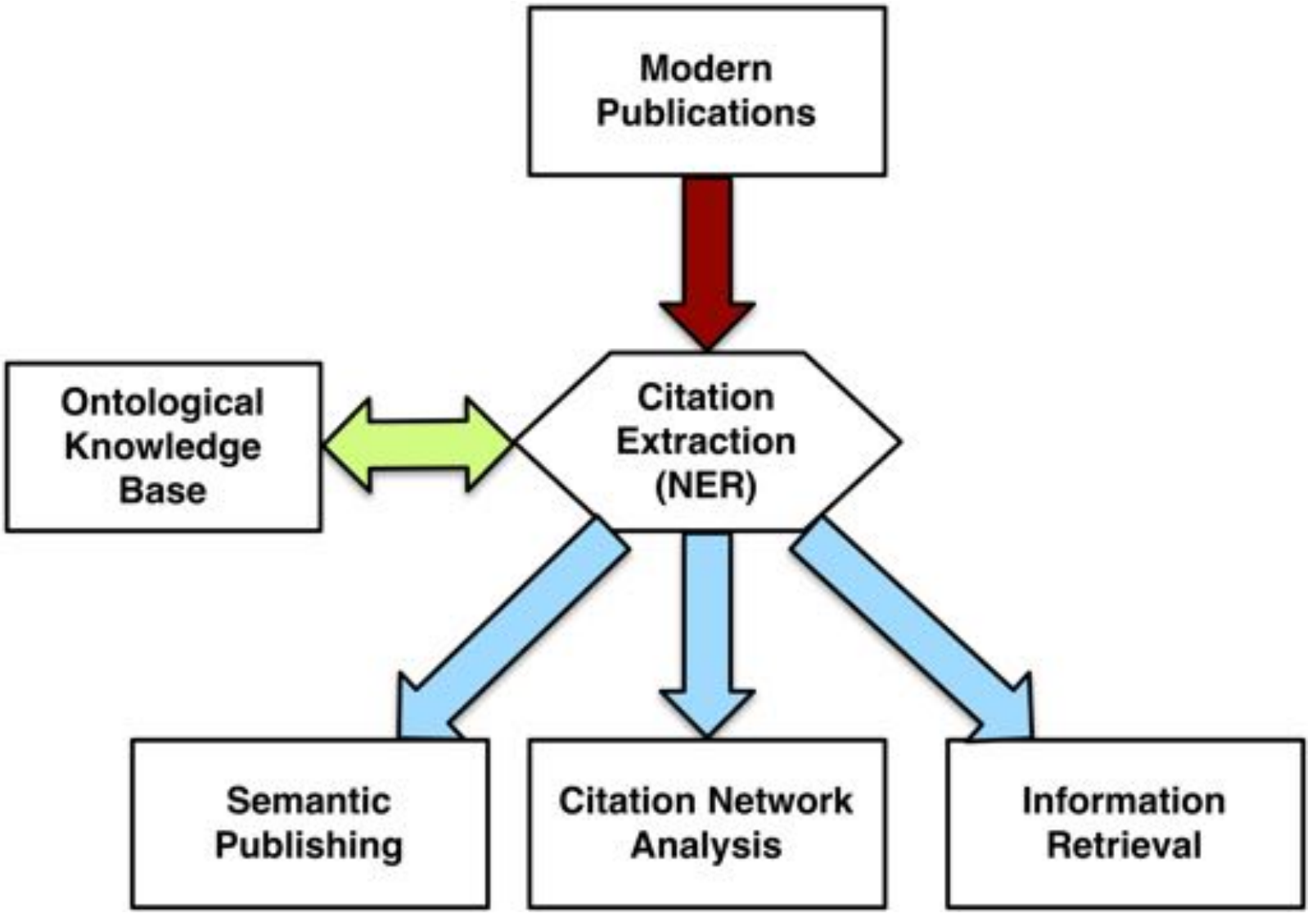
Research questions

1. Is it possible to extract canonical references automatically from modern publications such as journal articles?
2. With what level of accuracy can this extraction be performed?
3. How can a system be implemented to perform this task?

Research questions

1. Is it possible to extract canonical references automatically from modern publications such as journal articles?
2. With what level of accuracy can this extraction be performed?
3. How can a system be implemented to perform this task?
4. How might the applications enabled by the automatic extraction of canonical references change the way we study classical texts?

The Big Picture



Contributions to Research

1. **a system** for the automatic extraction of canonical references;

Contributions to Research

1. **a system** for the automatic extraction of canonical references;
2. **an ontology** of canonical references that allows us to publish the extracted citation data online by means of semantic technologies;

Contributions to Research

1. **a system** for the automatic extraction of canonical references;
2. **an ontology** of canonical references that allows us to publish the extracted citation data online by means of semantic technologies;
3. **a three-level citation network** laying the foundations for

Contributions to Research

1. **a system** for the automatic extraction of canonical references;
2. **an ontology** of canonical references that allows us to publish the extracted citation data online by means of semantic technologies;
3. **a three-level citation network** laying the foundations for
   - the development of new search tools

Contributions to Research

1. **a system** for the automatic extraction of canonical references;
2. **an ontology** of canonical references that allows us to publish the extracted citation data online by means of semantic technologies;
3. **a three-level citation network** laying the foundations for
   - the development of new search tools
   - the analysis of these networks

From *Index Locorum* to Citation Network

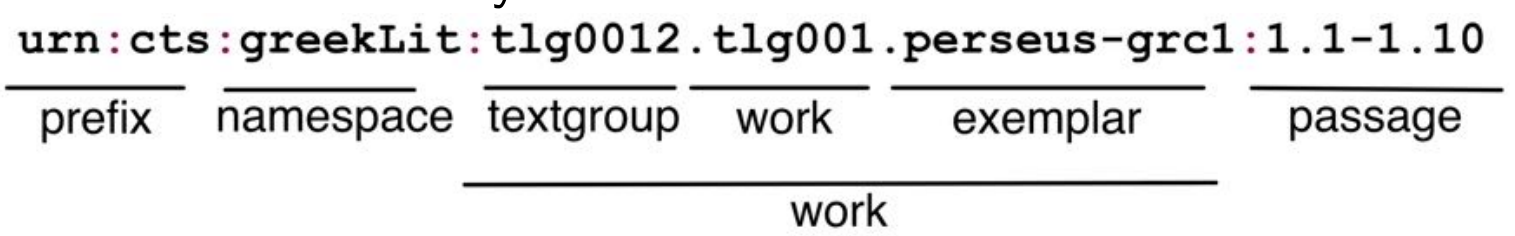
Index locorum.

From Footnotes/References to Links

- *footnotes*: only backwards
- *references*: only outwards
- *links*:
  - future-proof footnotes
  - reverse references
- Pelagios:
  - focus on place names/geographical data
  - common unique identifiers to expresss place references
  - resources referring to same place connected together
- linking texts by mining references/citations
  - but common identifiers are needed...

Canonical Text Services Unique Resource Names (CTS URNs)

A machine-readable syntax for canonical references

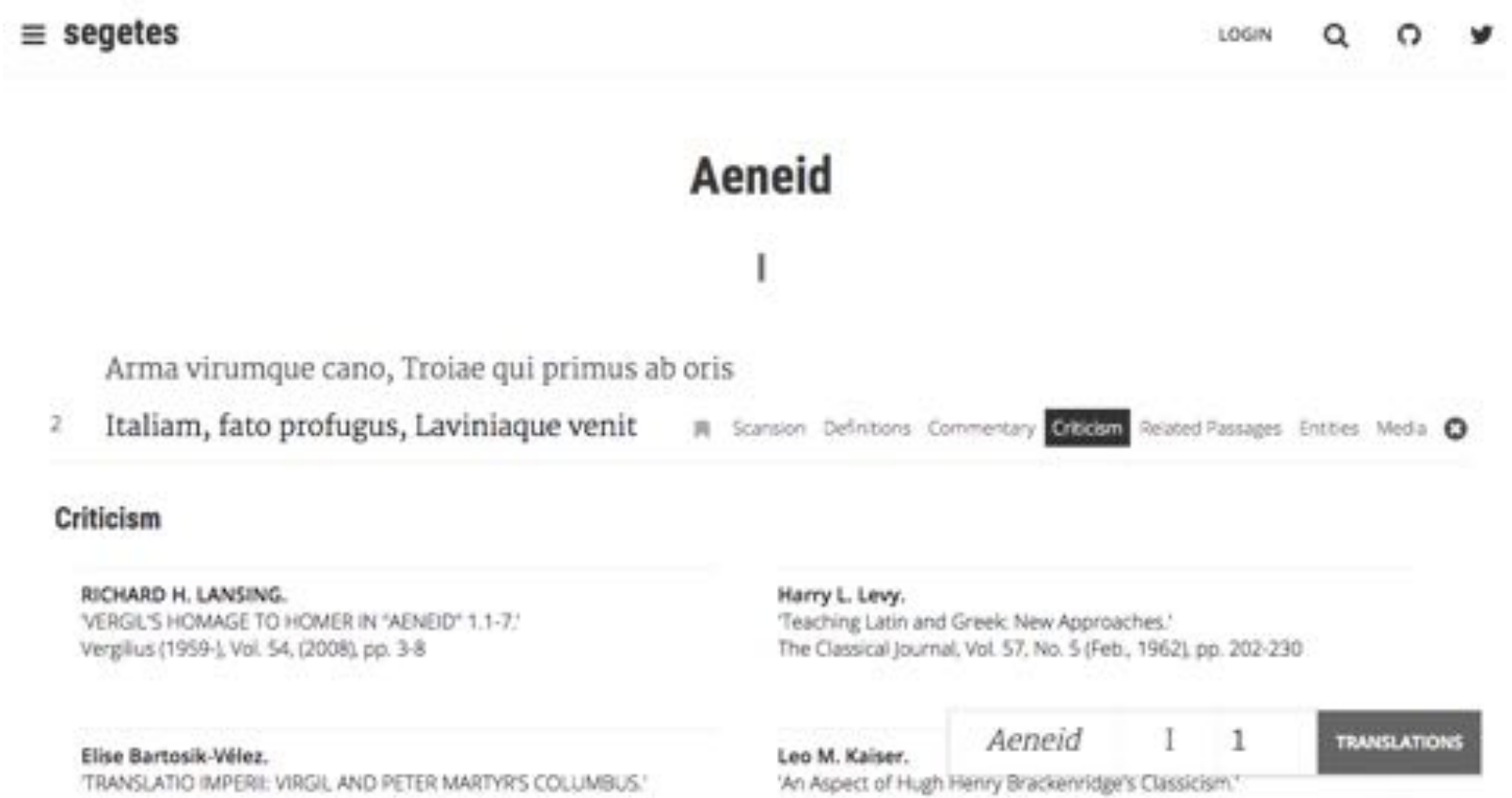


- Homer
  - `urn:cts:greekLit:tlg0012`
- Homer's *Iliad*
  - `urn:cts:greekLit:tlg0012.tlg001`
- Hom., *Il.* 1.1-10
  - `urn:cts:greekLit:tlg0012.tlg001:1.1-1.10`

Enhanced Digital Editions (1)

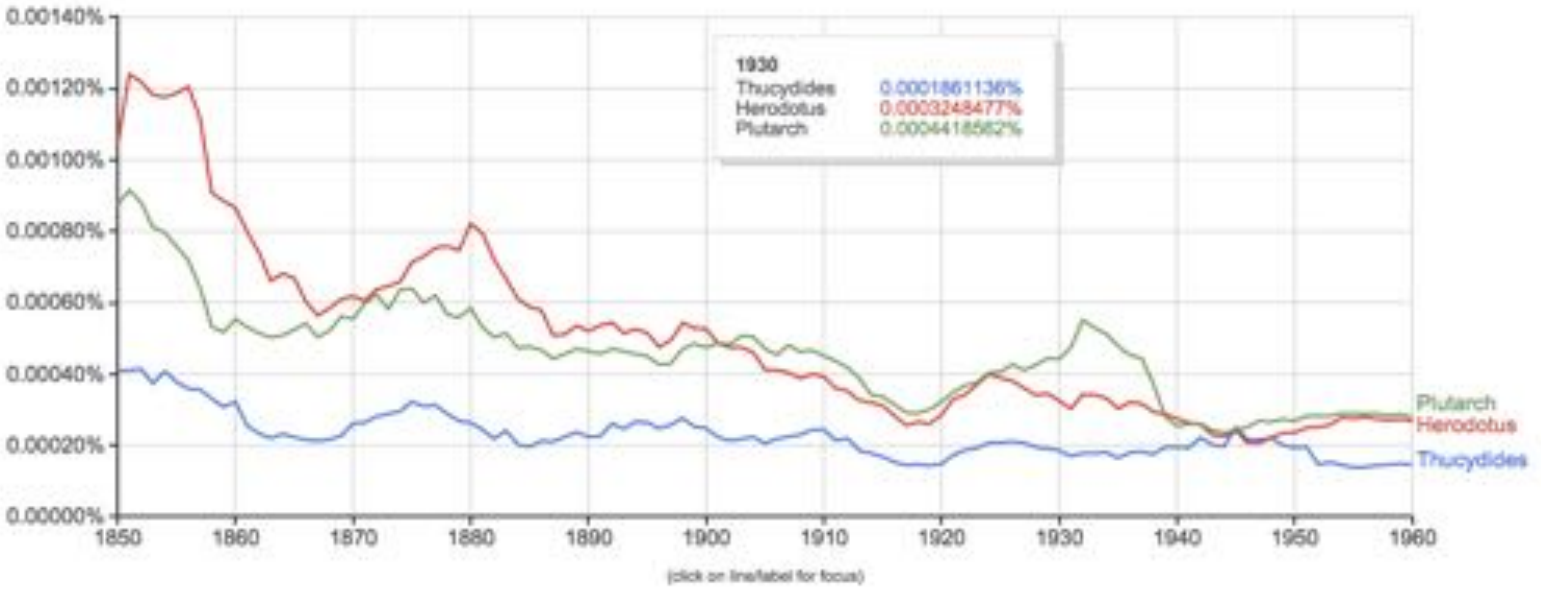
Hellespont Project
`http://gapvis.hellespont.dainst.org/#book/1/read/113/`

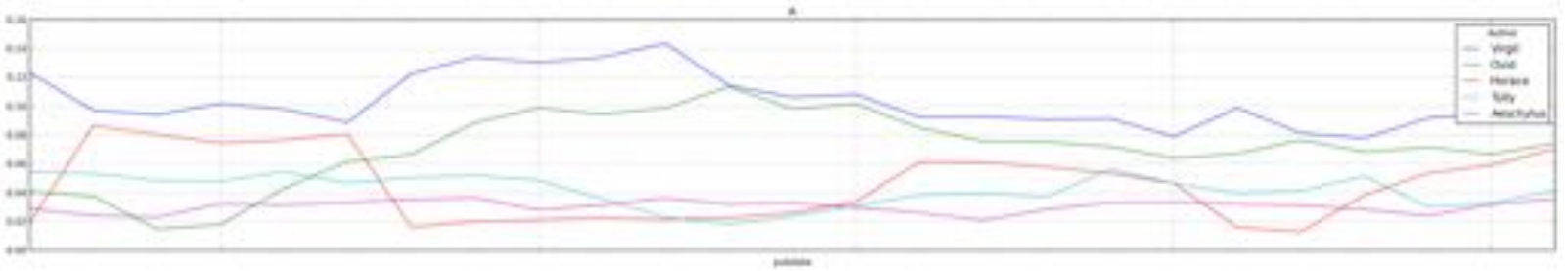
Enhanced Digital Editions (2)

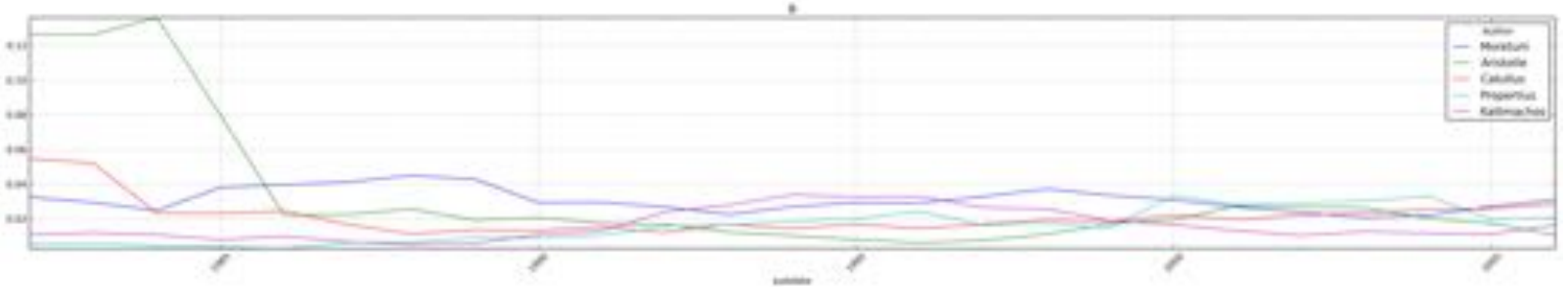


Segetes, `http://segetes.io/aeneid`

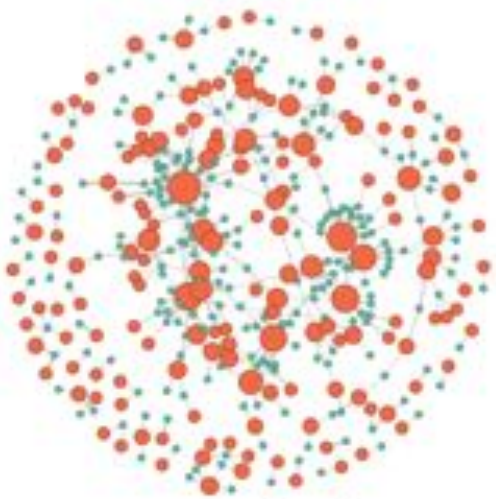
Distant Reading

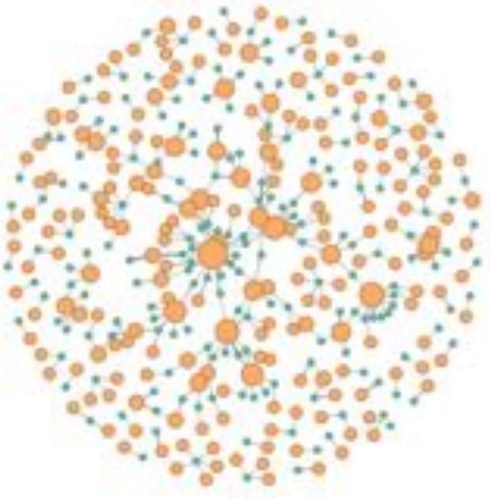






Networks & Information Retrieval





with applications to:

1. search
2. document clustering
3. network analysis and visualisation

Extraction – Approach

Rationale

- beyond string-based search
- references not quotations
- scalable approach:
  - language independent
  - applicable to large amounts of documents
  - easily adaptable to different materials and ways of referencing

Examples:

- In Statius' « Achilleid » (2, 96-102) Achilles describes [...]
- e.g. Vergil, Aen. 12, 101-109 ; Lucan 1, 204-212 ; Statius, Th. 12, 736-740 [...]

Named Entity Recognition

Computer Science > Information Extraction > NER

Question Answering System

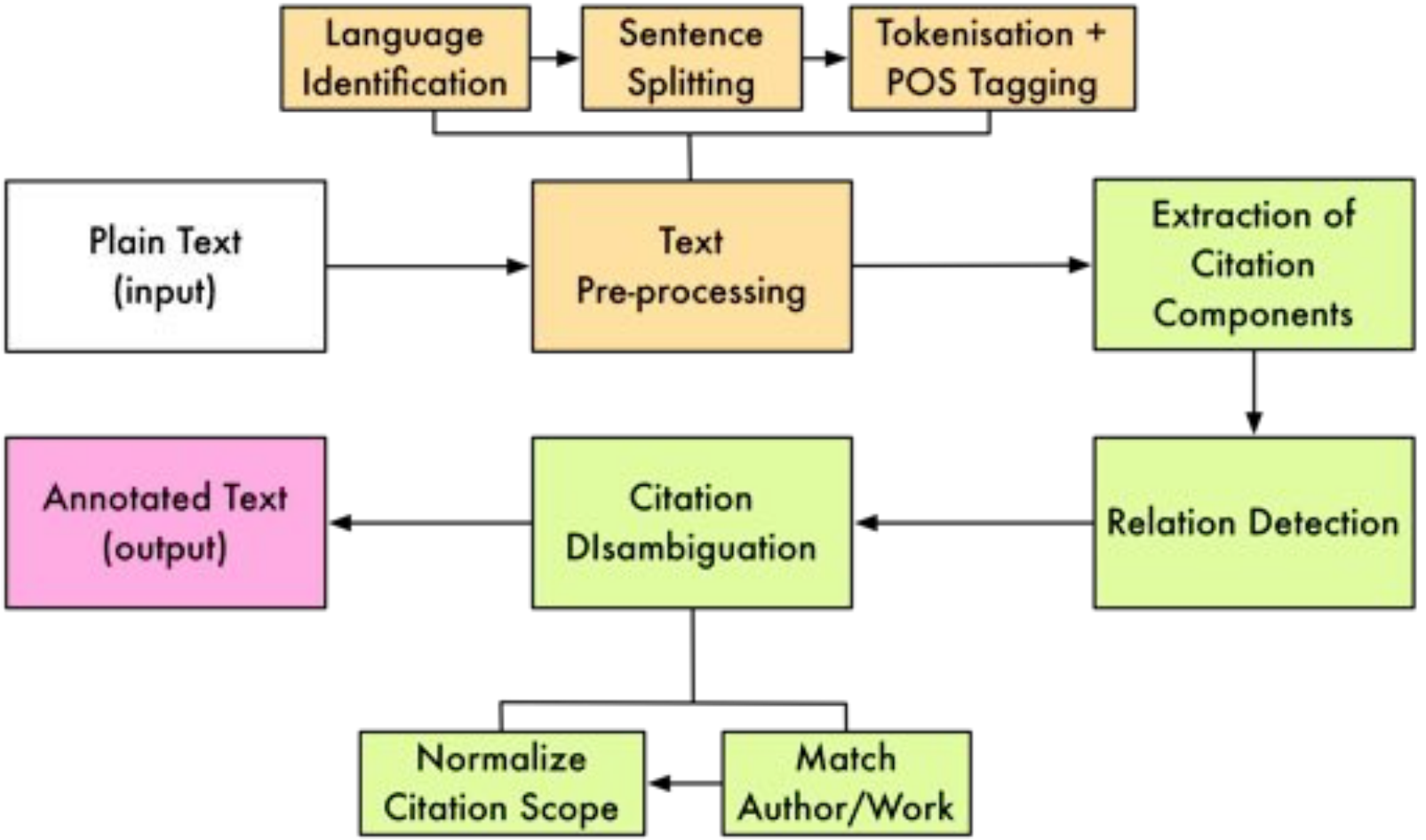
- **Q:** where did Aaron Swartz die?
- **A:** New York

Two days after the prosecution rejected a counter-offer by ***Swartz****,* ***he*** was found dead in *his* ***Brooklyn, New York*** *apartment, where he had hanged himself.*

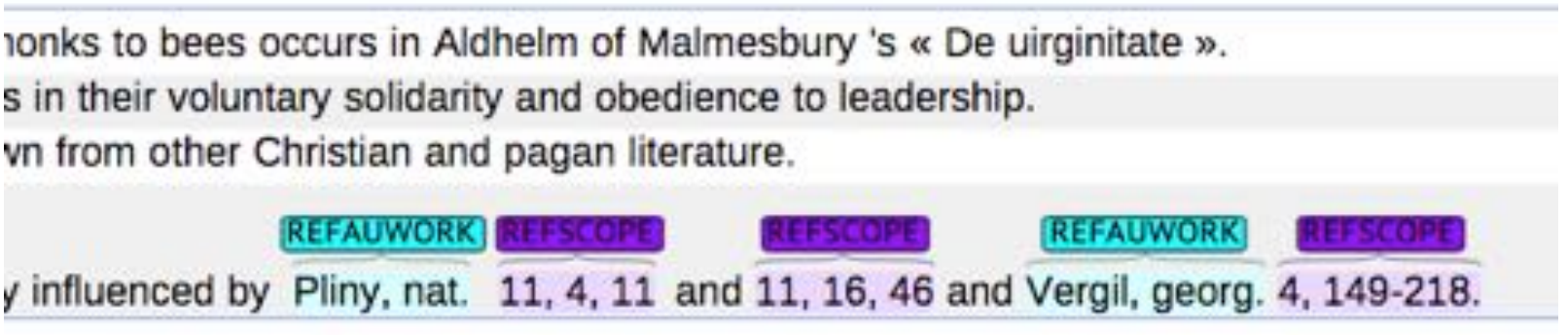
3-step process:

1. Named Entity Recognition and Classification
2. Relation Extraction
3. Named Entity Disambiguation

The Extraction Pipeline



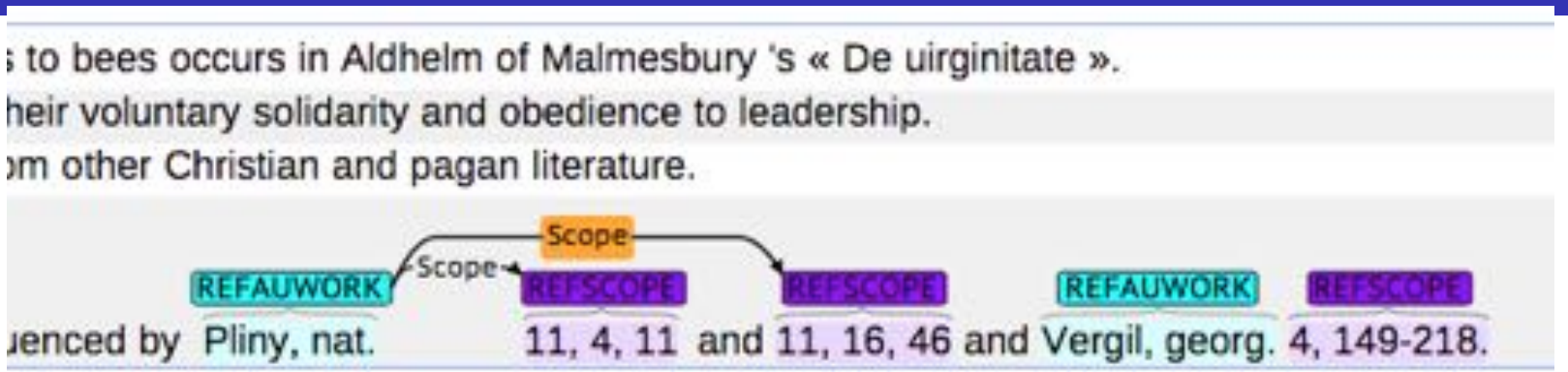
Citation Extraction: Step 1 (NER)



Named Entities (= citation components):

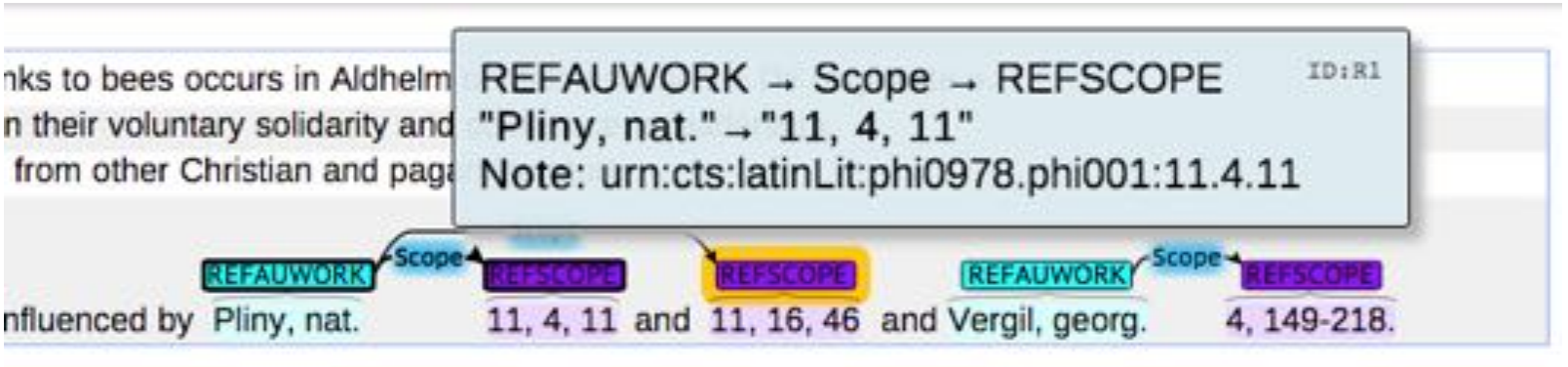
- `AAUTHOR` = ancient author
- `AWORK` = ancient work
- `REFAUWORK` = concise reference to author, work or both ("Pliny, nat.", "Thuc.")
- `REFSCOPE` = indication of the cited passage ("11, 4, 11")

Citation Extraction: Step 2 (Relation Detection)



- reference as *relation* vs. reference as *monolithic entity*
- binary `scope` relation between two entities (arguments)
  - arg1: `aauthor` | `awork` | `refauwork`
  - arg2: `refscope`
- examples:
  - Ammianus (15, 8, 7) {aauthor+refscope}
  - Trabajos 159–173" {awork+refscope}
  - Thuc. 1.89.1 {refauwork+refscope}

Citation Extraction: Step 3 (Disambiguation)



- assign each author/work/canonical reference a unique ID
- IDs are CTS URNs

Extraction – Evaluation

L'Année philologique (APh)

http://www.annee-philologique.com/

APh Example

APh 75-06697 => S. Braund & G. Gilbert. 2004. "An ABC of epic ira: anger, beasts, and cannibalism" *Yale Classical Studies* 32:250-285

> *In Statius ' « Achilleid » (2, 96-102) Achilles describes his diet of wild animals in infancy, which rendered him fearless and may indicate another aspect of his character - a tendency toward aggression and anger.*

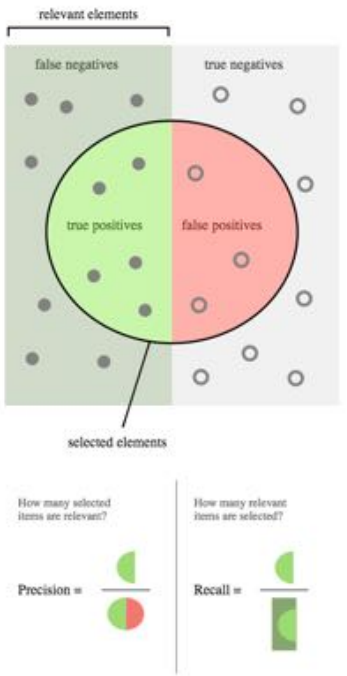
> *The portrayal of angry warriors in Roman epic is effected for the most part not by direct descriptions but indirectly, by similes of wild beasts (e.g. Vergil, Aen. 12, 101-109; Lucan 1, 204-212; Statius, Th. 12, 736-740; Silius 5, 306-315).*

> *These similes may be compared to two passages from Statius (Th. 1, 395-433 and 8, 383-394) that portray the onset of anger in direct narrative. Analysis of these passages demonstrates that the concept of « ira » in epic takes its moral aspect from the context.*

The Data

- **APh**
  - analytical reviews (en, de, fr, es, it)
  - 80 volumes (1924-)
  - autom. processed vol. 75 (2004)
    - 6,694 abstracts (total = 6,946; errors = 252)
    - 350k tokens
    - 3k citations
  - man. corrected ~8 % of vol. 75
    - 366 abstracts
    - 26k tokens
    - 380 citations

Precision, Recall and F1 Score



https://commons.wikimedia.org/wiki/File:Precisionrecall.svg

By Walber (Own work) [CC BY-SA 4.0]

Evaluation Summary

Task	Precision	Recall	F_1 Score
NER	79.24%	69.62%	73.88%
RelEx	93.33%	91.87%	92.60%
NED	61.04%	90.94%	73.05%

- methods:
  - NER: machine learning-based
  - RelEx: rule-based
  - NED: rule-based + knowledge base

NER: Method

1. Linguistic Features (PoS Tag, neighbouring words)
2. Word-level Features:
   - punctuation
     - final_dot, quotation_mark, has_hyphen, bracket
   - case
     - mixed_caps, all_caps, init_caps, all_lower
   - number
     - roman, year, range, mixed_alphanum
   - patterns
     - "Avien." –> "Aaaaa-" (expanded)
     - "Avien." –> "Aa-" (compressed)
3. Semantic Features (matches against dictionaries)

NER: Evaluation

Task: extraction of entities `aauthor`, `awork`, `refauwork`, `refscope`

Algorithm	**Precision**	**Recall**	**F_1 Score**
CRF	**79.24%**	69.62%	**73.88%**
MaxEnt	75.29%	66.75%	70.43%
SVM	74.44%	**70.21%**	71.93%

`Aauthor` : P = 91.15%, R = 39.67%, F_1 = 54.53%

RelEx: Evaluation

- rule-based method

Precision	Recall	F_1 Score
93.33%	91.87%	92.60%

Missed `scope` relations:

- du [REFSCOPE chant 4] de l' [AWORK « Énéide »]
- Le [REFSCOPE livre 13] de la [AWORK « Chronique »]
- les [REFSCOPE v. 9–12] des [AWORK « Acharniens »]

NED: Method

Thuc. I 89, 1s.

1. match reference against knowledge base
   - exact/approximate string matching
     - `edit_distance("Tucidide","Thucydides") = 3`
   - Thuc. → `urn:cts:greekLit:tlg0003.tlg001`
2. normalise the reference scope
   - I 89, 1s. → 1.89.1–1.89.2
   - other possible variants:
     - 1.89.1–2
     - 1, 89, 1–2
3. assign unique ID
   - `urn:cts:greekLit:tlg0003.tlg001:1.89.1-1.89.2`

NED: Evaluation

Matching Type	**Precision**	**Recall**	**F_1 Score**
Exact	58.33%	62.88%	60.52%
Approximate (n=4)	61.04%	90.94%	73.05%
Approximate (n=7)	58.94%	94.76%	72.67%

NED: Error Types

1. abbreviation is highly ambiguous

   *But **Horace** undermines the suggestion that his own poetry will forever represent the Augustan Age. **Carm. 4, 15** in fact [...]*

2. ambiguous author mention

   *Esame dell' esegesi papiracea ad **Aristofane** : permanenza del lavoro degli eruditi alessandrini [...]*

NED: Error Types (contd.)

3 implicit topicalisation

*Dans son **chap. 5** sur le squelette et la respiration, **Lactance** utilise des sources disparates et arrive aux limites de son savoir médical.*

4 ambiguously expressed reference

*Analysis of the pederastic poems in the Theocritean corpus **(12 ; 23 ; 29 ; 30)** reveals that **Theocritus** reflects on mutuality in a relationship [...]*

Future Plans

Large-scale Extraction and Analysis

JSTOR data

- 130k articles, 320m tokens of text
- >200 years of classical scholarship

Analysis

- diachronic citation trends
- longitudinal network analysis

Bottleneck

- processing time (single machine, single process)
- parallelisation of the code, use of EPFL's cluster
  - pre-processing of JSTOR data, 6 months → 6 days

Citation Extraction

Improve Overall Accuracy

- test machine learning methods for RelEx and NED
- get more training data, expand entity set
- expand the knowledge base

Share the Software

- streamline installation
- improve documentation
- offer as web service
- offer as part of a research infrastructure

Thank you for your attention!

matteo.romanello@epfl.ch or
matteo.romanello@dainst.org