

# Mining Citations, Linking Texts

Matteo Romanello (EPFL / DAI) @mr56k

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

Mining  
Citations,  
Linking Texts

Matteo  
Romanello  
(EPFL / DAI)  
@mr56k

Overview

Extraction –  
Approach

Extraction –  
Evaluation

Future Plans

# Overview

## ■ 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

Mining  
Citations,  
Linking Texts

Matteo  
Romanello  
(EPFL / DAI)  
@mr56k

Overview

Extraction –  
Approach

Extraction –  
Evaluation

Future Plans

## 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*

Mining Citations, Linking Texts

Matteo Romanello (EPFL / DAJ) @mr56k

Overview

Extraction – Approach

Extraction – Evaluation

Future Plans

The screenshot displays the 'Dreißiger-Abschnitt 1' section of the Parzival Bibliography. It features a sidebar with navigation options like 'START', 'MENU', and 'HILFE', and a main content area with a list of line numbers (1.1 to 1.14) and their corresponding bibliographic references. A browser window in the background shows the website's URL: wolfram.lexcoll.com/bib/biblio.html#1138.

**Dreißiger-Abschnitt 1**

1.1 Ist zwivel herzen nâchgêr,  
1.2 daz muoz der sîle werden sôr.  
1.3 gemæhet unde gezeret  
1.4 ist, swâ sich parriert  
1.5 unverzaget mannes muot,  
1.6 als ægelstern varwe tuot.  
1.7 der mac dennoch wesen geil:  
1.8 wand in im sint bediu teil,  
1.9 des himels und der helle.  
1.10 der unstatte geselle  
1.11 hât die swarzen varwe gar,  
1.12 und wirt och nâch der vinstar var:  
1.13 sô habet sich an die blanken  
1.14 der mit staten gedanken.

**1, 1** [zu den verpublizierten Ausgaben]

**Abbildungen:** Schöck 1993 197, 199, Palmer, N. 1992 26  
**Aufstellungen:** Hergel 1942 16 [1,1-1,14], 195, 199 [1,1-1,14], 328 Anm. 16 [1,1-1,6], Huby, Michel 1968 40 >  
**Allegorien:** Hains 2000 170-73 [1,1-1,6]  
**Aufführungen:** Guterhoner 1956 36 [1,1-1,14], 38 [1,1-4,8]  
**Ausgabenmethodik:** McCulloh 1983 486, Schweik, G. 1992 97 [1,1-2,14], Schöck 1998b LVIII, Schöck 1999d LVIII  
**Beispielen:** Grimm, N. 1994 565 [1,1-1,14]  
**Bernhard von Clairvaux:** Wilson, H. 1967 9 [1,1-2,22], 13 [1,1-1,14]  
**Bildungsroman:** Gerlach 1926 75 Anm. 2 [1,1-1,14], Gerlach 1968 25 Anm. 2 [1,1-1,14]  
**Charaktere:** Gross, Gertr. 1929 94 [1,1-1,14]  
**Charakterisierung:** Bammes 1966 193 [1,1-1,2], 195 Anm. 1  
**Deutung:** Schröder, Werner 1937 288 [1,1-1,2], 292 [1,1-1,14], Schröder, Werner 1998a 26 [1,1-1,2], 30 [1,1-1,14]  
**Didaktik u. Lehrlinien:** Bosch 1972 31 [1,1-1,2], 174 [1,1-1,14], 176 [1,1-1,2], 177 [1,1-1,5]  
**Edelesteine u. Edelsteinsymbolik:** Ergenz 1978 297 Anm. 10 [1,1-1,14]  
**Entstehungsgeschichte des Parzival:** Hempel, H. 1951/52 162-80 [1,1-4,8], Hatto 1952/53 235 [1,1-1,14], Henrich 1903/06 237-40 [1,1-4,26], Grimm, Walther Ludwig 1897 41 [1,1-4,8], Schneider, A. 1922 122 [1,1-4,8], 124 [1,1-3,24], 150 [1,1-3,24], 151 [1,1-3,24], Karg-Gasterheld 1923 149 [1,1-4,8], Hempel, H. 1956 263-76 [1,1-4,8], Bammes 1970a 289 [1,1-4,26], Bammes 1988 11 [1,1-4,26]  
**Erzähler:** Gutschmann 1971a 61 [1,1-9,27], Förster, U. 1971 133 [1,1-1,2], Hofmann 2000 58 >  
**Erzählweise:** Richey 1923b 86 [1,1-4,30], Levine 1962 187 [1,1-1,14], Stein, A. 1993 170 [1,1-1,2], 176 [1,1-1,14], 177 [1,1-1,14], 235 [1,1-1,14], 240 [1,1-1,14], Gross 1995 2 [1,1-3,27], Bumke 1999a 198 >, 1992 [1,1-1,14], Möhren 2000 158 Anm. 18 [1,1-1,2], Bumke 2001c 1081 >, 1082 [1,1-1,14], Schu 2002 43 [1,1-1,2], 62 [1,1-1,25]

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

# JSTOR Labs: Understanding Shakespeare

Mining Citations, Linking Texts

Matteo Romanello (EPFL / DAI) @mr56k

Overview

Extraction – Approach

Extraction – Evaluation

Future Plans

**BETA**  
*Understanding Shakespeare*

Folger's Digital Text for *Twelfth Night* - Act 1 Scene 1 # of articles

*(with Musicians playing.)*

Text	# of articles
ORSINO	
FTLN 0001 If music be the food of love, play on.	17
FTLN 0002 Give me excess of it, that, surfeiting,	7
FTLN 0003 The appetite may sicken and so die.	8
FTLN 0004 <b>That strain again! It had a dying fall.</b>	12
FTLN 0005 O, it came o'er my ear like the sweet sound	7
FTLN 0006 That breathes upon a bank of violets,	7
FTLN 0007 Stealing and giving odor. Enough; no more.	8
FTLN 0008 'Tis not so sweet now as it was before.	5
FTLN 0009 O spirit of love, how quick and fresh art thou,	8
FTLN 0010 That, notwithstanding thy capacity	6
FTLN 0011 Receivest as the sea, naught enters there,	7
FTLN 0012 Of what validity and pitch soe'er,	8
FTLN 0013 But falls into abatement and low price	8
FTLN 0014 Even in a minute. So full of shapes is fancy	9
FTLN 0015 That it alone is high fantastical.	5
CURIO	
FTLN 0016 Will you go hunt, my lord?	2
FTLN 0017 ORSINO What, Curio?	
FTLN 0018 CURIO The hart.	

**Articles**

- Missing Mourning
- Suzanne Pevul
- Violet's Fall
- Mothers' M
- Garibaldi
- Mario Curi
- Violet's Fall
- Allusion | Cr
- Reversal Puritans
- Paul Yach
- Theater | Vi
- Playwriting |
- Gothic Lit
- Deidre Lyr
- British Iteral
- Lytic poetry

75

for the Topos, and Maria's handwriting for Olivia's report the topos. The twinning in *Twelfth Night* functions as a response to death.<sup>1</sup> A double is most obviously a form of spatial repetition, with one person or image duplicated in another place. However, it can also be chronological repetition: someone from the past is copied into the present, as is the case in the play. This essay will discuss doubling in *Twelfth Night*, its connections to the ambiguously targeted figures of the early modern letter, and the multiple implications of that longing. A response to a specifically post-Reformation hunger, I will argue, the double takes its force from changes in mourning rituals that accompanied the decline of English Catholicism. It serves as a testament to the power of the father-child bond and ultimately as a fantasy of its replacement.

DEATH AND THE DOUBLE

That *Twelfth Night* concerns itself with death is a familiar observation; mortality marks its entrance in the first scene even with the evidently healthy young Orsino, who wishes for music so that "his appetite may sicken and so die" (1.1.12).<sup>2</sup> That strain again," he requests, "it had a dying fall" (1.1.14). With Hamlet, first verbalized around the same time, dying is the alpha and omega of the play. But unlike Hamlet, *Twelfth Night* concludes with the promise of marriage. It also ends not with the more typically comic references to pregnancy but with a song that

<sup>1</sup> For doubling as an actor of death, see Karl Miller, *Double: Studies in Literary History* (New York: Oxford University Press, 1981), 48. For an analysis of the double as death sleeping on a mountain side and another German iteration, see Otto Rank, *The Double: A Psychoanalytic Study*, re Henry Suckler, in *Chicago 1908: University of North Carolina Press, 1913*, 71-76. Sigmund Freud, writing on the topic four years after Rank, claims that the double functions as a denial of death. He notes that "beyond the 'successful' and was the first 'double' of the body" ("The Uncanny" [1919], in *The Standard Edition of the Complete Psychological Works of Sigmund Freud*, vol. 17 [London: Hogarth, 1961], 237).

<sup>2</sup> That Orsino's logic creates fatally severe allusions to death ("Twelfth Night: The Lovers of Antioch," *Studies in English Literature* 11 [1981]: 231). Anne Barton observes that these references increase in frequency as the play progresses. See her introduction to *Twelfth Night*, in *The Complete Shakespeare*, ed. C. Balestrino Evans et al. (Boston: Houghton Mifflin, 1974), 420.

<sup>3</sup> Edward Williams, among others, has commented on the play's focus on deathliness. See his "Machaveli Fall," *Shakespeare Quarterly* 20 (1969): 49. See James Callaghan, *Shakespeare and the Power of Death* (Cambridge: University of Cambridge Press, 1991), for an examination of the ways in which Shakespearean characters seek immortality by joining with large animals; the Amazon I describe in this essay is similar to that discussed by Callaghan, albeit on a more utopian scale.

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

# Research questions

Mining  
Citations,  
Linking Texts

Matteo  
Romanello  
(EPFL / DAI)  
@mr56k

Overview

Extraction –  
Approach

Extraction –  
Evaluation

Future Plans

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

# Research questions

Mining  
Citations,  
Linking Texts

Matteo  
Romanello  
(EPFL / DAI)  
@mr56k

Overview

Extraction –  
Approach

Extraction –  
Evaluation

Future Plans

- 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

Mining  
Citations,  
Linking Texts

Matteo  
Romanello  
(EPFL / DAI)  
@mr56k

Overview

Extraction –  
Approach

Extraction –  
Evaluation

Future Plans

- 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

Mining  
Citations,  
Linking Texts

Matteo  
Romanello  
(EPFL / DAI)  
@mr56k

Overview

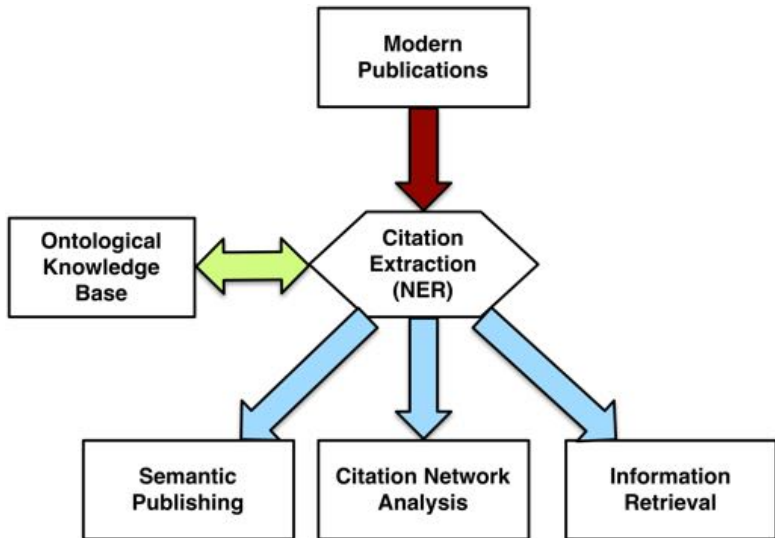
Extraction –  
Approach

Extraction –  
Evaluation

Future Plans

- 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

Mining  
Citations,  
Linking Texts

Matteo  
Romanello  
(EPFL / DAI)  
@mr56k

Overview

Extraction –  
Approach

Extraction –  
Evaluation

Future Plans

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

# Contributions to Research

Mining  
Citations,  
Linking Texts

Matteo  
Romanello  
(EPFL / DAI)  
@mr56k

Overview

Extraction –  
Approach

Extraction –  
Evaluation

Future Plans

- 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

Mining  
Citations,  
Linking Texts

Matteo  
Romanello  
(EPFL / DAI)  
@mr56k

Overview

Extraction –  
Approach

Extraction –  
Evaluation

Future Plans

- 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

Mining  
Citations,  
Linking Texts

Matteo  
Romanello  
(EPFL / DAI)  
@mr56k

Overview

Extraction –  
Approach

Extraction –  
Evaluation

Future Plans

- 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

Mining  
Citations,  
Linking Texts

Matteo  
Romanello  
(EPFL / DAI)  
@mr56k

Overview

Extraction –  
Approach

Extraction –  
Evaluation

Future Plans

- 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

Mining Citations,  
Linking Texts

Matteo  
Romanello  
(EPFL / DAI)  
@mr56k

Overview

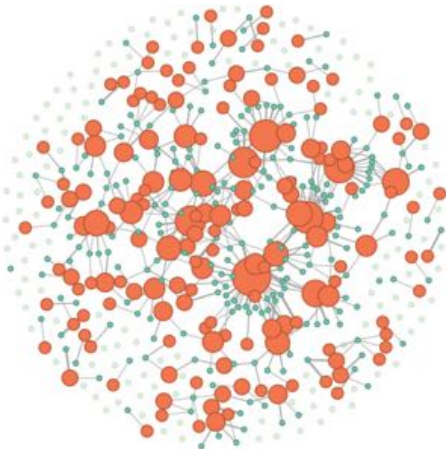
Extraction –  
Approach

Extraction –  
Evaluation

Future Plans

## Index locorum.

Aemilius Asper p. 112 K. Pag.	352	Aristot. Eth. ad Nicom. 1, 3 p. 740
-- p. 120	356	-- Polit. II, c. 11, 1272b 30R. 184
-- p. 127 a. 127b. 128 a.	355	-- III, c. 16, 1294b 30
-- p. 130 b.	354	-- c. 16, 16
Asch. Agon. 561. 1287. 1487.	392	-- c. 17, 1287 b, 37 sqq. B.
-- Sept. 705	42	-- 1268 a, 15 sqq.
Ampoi. c. 2, p. 5, 9 W10813	643	-- 30
-- c. 6, p. 5, 10	700	-- 41
Anacr. fr. 60 Burgk.	700	-- IV, c. 1, 1268b 23
Anth. Pal. V, 106	525	-- c. 8, 1827b 16, c. 8, 1828
-- 170	533	a 25, c. 9, 1828 b, 29, 1829 a
-- VI, 122	534	25
-- 264	160	-- c. 10, 1829 b 33
-- 207	206	-- c. 11, 1830a 24, 41
-- 260	160	-- 1850 b, 2
-- 279	491	-- c. 13, 1831 a 29, 1831 b.
-- 285	536	4
-- VII, 95	534	-- c. 18, 1831 b 1, 1832 b 1
-- 401	527	-- c. 14, 1833 a 2
-- 414	523	-- 1833 a 24
-- 420	528	-- 1833 b 23, c. 15, 1834
-- 460	523	a 18, 1835 b 2, 21
-- 491	161	-- VII, c. 17, p. 1828 a 2 34 sqq.
-- 576	341	-- 1356 b 5, 9
-- 576	491	-- Polit. VIII, c. 1, p. 1337 a
-- 481	491	11, c. 2, 1337 b 5
-- 730	536	-- c. 2, p. 1837 b 11, c. 3,
-- IX, 57	492	1837 b 24, c. 4, 1838 b 26 c.
-- 815	504	b, 1840 a 12
-- 304	161	-- c. 7, 1841 b 19, 20
-- XII, 130	150	Artes. XIII, p. 560 A
-- XIII, 22	492	Cass. B. G. III, 28
-- Append. Platon. 160, 6 296	341	Cic. pro Caes. 5, 11
Aristoph. Pan. 620	560	-- 12, 29, 29, 47
-- Plat. 117 v. Testes.		-- de Divin. I, c. 11, 12, 13



# From Footnotes/References to Links

Mining  
Citations,  
Linking Texts

Matteo  
Romanello  
(EPFL / DAI)  
@mr56k

Overview

Extraction –  
Approach

Extraction –  
Evaluation

Future Plans

- *footnotes*: only backwards
- *references*: only outwards
- *links*:
  - future-proof footnotes
  - reverse references
- Pelagios:
  - focus on place names/geographical data
  - common unique identifiers to express 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)

Mining Citations,  
Linking Texts

Matteo Romanello  
(EPFL / DAI)  
@mr56k

Overview

Extraction – Approach

Extraction – Evaluation

Future Plans

A machine-readable syntax for canonical references

`urn:cts:greekLit:tlg0012.tlg001.perseus-grc1:1.1-1.10`

<u>urn</u>	<u>:cts</u>	<u>:greekLit</u>	<u>:tlg0012</u>	<u>.tlg001</u>	<u>.perseus-grc1</u>	<u>:1.1-1.10</u>
prefix		namespace	textgroup	work	exemplar	passage
			work			

- Homer

- `urn:cts:greekLit:tlg0012`

- Homer's *Iliad*

- `urn:cts:greekLit:tlg0012.tlg001`

- Hom., // 1.1-10

- `urn:cts:greekLit:tlg0012.tlg001:1.1-1.10`

# Enhanced Digital Editions (1)

Mining Citations,  
Linking Texts

Matteo Romanello  
(EPFL / DAI)  
@mr56k

Overview

Extraction –  
Approach

Extraction –  
Evaluation

Future Plans

The screenshot displays the GapVis for Hellenist web application. At the top, there is a navigation bar with the project name, a search bar, and links for 'About Hellenist Project', 'GapVis', 'Linked Data Interface', and 'Downloads and Links'. Below this, the main content area is titled 'The Pentecontaetia (Thuc. 1.89 - 1.118)' and includes a sub-header 'Thucydides: The Peloponnesian War'. A link to the 'Current passage on Perseus' is provided. The main text area shows 'Chapter 113' with three sections of Greek text. The text is annotated with colored boxes and links for entities like 'Athens', 'Boeotians', 'Corinthians', 'Thebes', and 'Athenians'. To the right, a 'Secondary Literature' section lists a citation by Martin Ostwald and provides links to the original text on Perseus and a JSTOR article.

Hellespont Project

<http://gapvis.hellespont.dainst.org/#book/1/read/113/>

# Enhanced Digital Editions (2)

Mining  
Citations,  
Linking Texts

Matteo  
Romanello  
(EPFL / DAI)  
@mr56k

Overview

Extraction –  
Approach

Extraction –  
Evaluation

Future Plans

≡ segetes

LOGIN



## Aeneid

I

Arma virumque cano, Troiae qui primus ab oris

2 Italiam, fato profugus, Laviniaque venit



Scansion

Definitions

Commentary

**Criticism**

Related Passages

Entries

Media



### Criticism

**RICHARD H. LANSING.**

'VERGIL'S HOMAGE TO HOMER IN "AENEID" 1.1-7:  
Vergilius (1959-), Vol. 54, (2008), pp. 3-8

**Harry L. Levy.**

'Teaching Latin and Greek: New Approaches.'  
The Classical Journal, Vol. 57, No. 5 (Feb., 1962), pp. 202-230

**Elise Bartosik-Vélez.**

'TRANSLATIO IMPERII: VIRGIL AND PETER MARTYR'S COLUMBUS.'

**Leo M. Kaiser.**

'An Aspect of Hugh Henry Brackenridge's Classicism.'

*Aeneid*

I

1

TRANSLATIONS

Segetes, <http://segetes.io/aeneid>

# Distant Reading

Mining Citations,  
Linking Texts

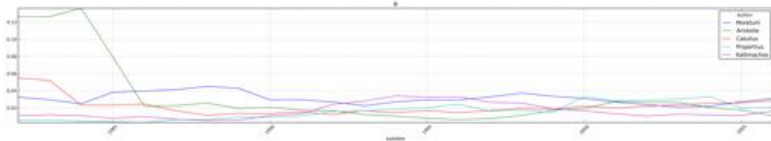
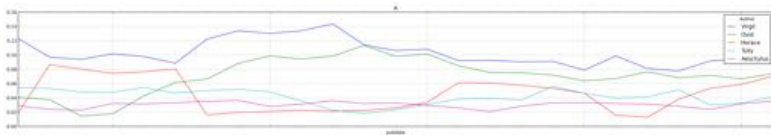
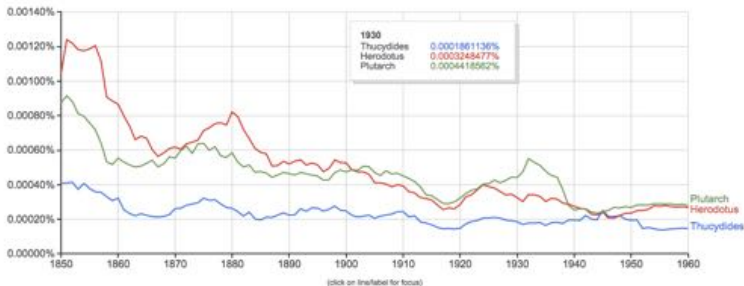
Matteo  
Romanello  
(EPFL / DAI)  
@mr56k

Overview

Extraction –  
Approach

Extraction –  
Evaluation

Future Plans



# Networks & Information Retrieval

Mining Citations,  
Linking Texts

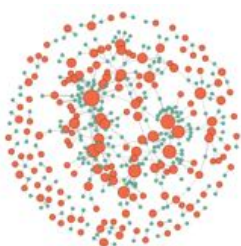
Matteo  
Romanello  
(EPFL / DAI)  
@mr56k

Overview

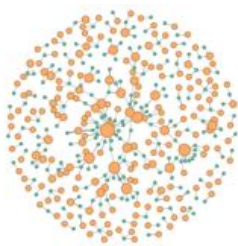
Extraction –  
Approach

Extraction –  
Evaluation

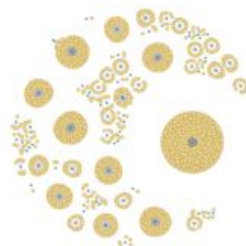
Future Plans



macro-level



meso-level



micro-level

with applications to:

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

Mining  
Citations,  
Linking Texts

Matteo  
Romanello  
(EPFL / DAI)  
@mr56k

Overview

Extraction –  
Approach

Extraction –  
Evaluation

Future Plans

# Extraction – Approach



# Rationale

Mining  
Citations,  
Linking Texts

Matteo  
Romanello  
(EPFL / DAI)  
@mr56k

Overview

Extraction –  
Approach

Extraction –  
Evaluation

Future Plans

- 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

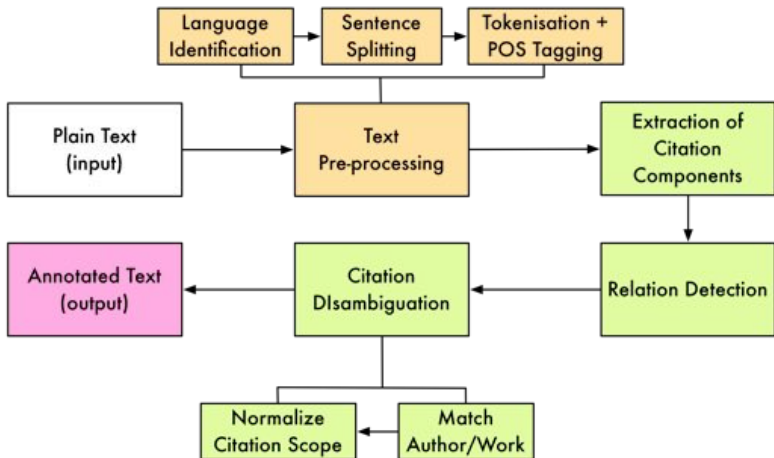
- **Q:** where did Aaron Swartz die?
- **A:** New York

*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



Mining Citations,  
Linking Texts

Matteo  
Romanello  
(EPFL / DAI)  
@mr56k

Overview

Extraction –  
Approach

Extraction –  
Evaluation

Future Plans

# Citation Extraction: Step 1 (NER)

Mining  
Citations,  
Linking Texts

Matteo  
Romanello  
(EPFL / DAI)  
@mr56k

Overview

Extraction –  
Approach

Extraction –  
Evaluation

Future Plans

honks to bees occurs in Aldhelm of Malmesbury 's « De uirginitate ».  
s in their voluntary solidarity and obedience to leadership.  
vn from other Christian and pagan literature.

y influenced by **REFAUWORK** Pliny, nat. **REFSCOPE** 11, 4, 11 and **REFSCOPE** 11, 16, 46 and **REFAUWORK** Vergil, georg. **REFSCOPE** 4, 149-218.

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)

is to bees occurs in Aldhelm of Malmesbury's « De uirginitate ».  
their voluntary solidarity and obedience to leadership.  
on other Christian and pagan literature.

enced by Pliny, nat. 11, 4, 11 and 11, 16, 46 and Vergil, georg. 4, 149-218.

- 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)

Mining Citations,  
Linking Texts

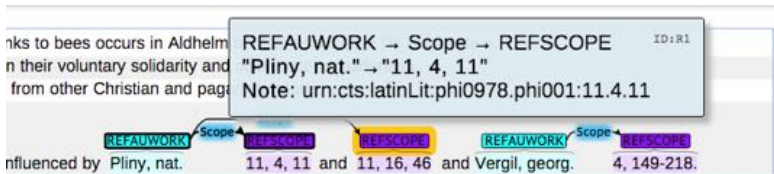
Matteo  
Romanello  
(EPFL / DAI)  
@mr56k

Overview

Extraction –  
Approach

Extraction –  
Evaluation

Future Plans



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

Mining  
Citations,  
Linking Texts

Matteo  
Romanello  
(EPFL / DAI)  
@mr56k

Overview

Extraction –  
Approach

Extraction –  
Evaluation

Future Plans

# Extraction – Evaluation

# L'Année philologique (APh)

Mining  
Citations,  
Linking Texts

Matteo  
Romanello  
(EPFL / DAI)  
@mr56k

Overview

Extraction –  
Approach

Extraction –  
Evaluation

Future Plans



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



# APh Example

Mining  
Citations,  
Linking Texts

Matteo  
Romanello  
(EPFL / DAI)  
@mr56k

Overview

Extraction –  
Approach

Extraction –  
Evaluation

Future Plans

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.*

## ■ 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

Mining Citations,  
Linking Texts

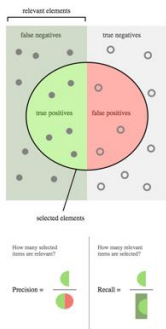
Matteo Romanello  
(EPFL / DAI)  
@mr56k

Overview

Extraction –  
Approach

Extraction –  
Evaluation

Future Plans



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

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

# Evaluation Summary

Mining  
Citations,  
Linking Texts

Matteo  
Romanello  
(EPFL / DAI)  
@mr56k

Overview

Extraction –  
Approach

Extraction –  
Evaluation

Future Plans

Task	Precision	Recall	F <sub>1</sub> 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

## 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

Mining  
Citations,  
Linking Texts

Matteo  
Romanello  
(EPFL / DAI)  
@mr56k

Overview

Extraction –  
Approach

Extraction –  
Evaluation

Future Plans

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

Algorithm	Precision	Recall	F <sub>1</sub> Score
CRF	<b>79.24%</b>	69.62%	<b>73.88%</b>
MaxEnt	75.29%	66.75%	70.43%
SVM	74.44%	<b>70.21%</b>	71.93%

Aauthor : P = 91.15%, R = 39.67%, F<sub>1</sub> = 54.53%

# RelEx: Evaluation

Mining  
Citations,  
Linking Texts

Matteo  
Romanello  
(EPFL / DAI)  
@mr56k

Overview

Extraction –  
Approach

Extraction –  
Evaluation

Future Plans

- rule-based method

Precision	Recall	F <sub>1</sub> 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 » ]

## 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

Mining  
Citations,  
Linking Texts

Matteo  
Romanello  
(EPFL / DAI)  
@mr56k

Overview

Extraction –  
Approach

Extraction –  
Evaluation

Future Plans

Matching Type	Precision	Recall	F <sub>1</sub> 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

Mining  
Citations,  
Linking Texts

Matteo  
Romanello  
(EPFL / DAI)  
@mr56k

Overview

Extraction –  
Approach

Extraction –  
Evaluation

Future Plans

- 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.)

Mining  
Citations,  
Linking Texts

Matteo  
Romanello  
(EPFL / DAI)  
@mr56k

Overview

Extraction –  
Approach

Extraction –  
Evaluation

Future Plans

## 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 [...]*

Mining  
Citations,  
Linking Texts

Matteo  
Romanello  
(EPFL / DAI)  
@mr56k

Overview

Extraction –  
Approach

Extraction –  
Evaluation

Future Plans

# Future Plans

# Large-scale Extraction and Analysis

Mining  
Citations,  
Linking Texts

Matteo  
Romanello  
(EPFL / DAI)  
@mr56k

Overview

Extraction –  
Approach

Extraction –  
Evaluation

Future Plans

## 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

Mining  
Citations,  
Linking Texts

Matteo  
Romanello  
(EPFL / DAI)  
@mr56k

Overview

Extraction –  
Approach

Extraction –  
Evaluation

Future Plans

## Improve Overall Accuracy

- test machine learning methods for ReEx 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](mailto:matteo.romanello@epfl.ch) or  
[matteo.romanello@dainst.org](mailto:matteo.romanello@dainst.org)