Citation Mining of Humanities Journals

A comparison of the projects *Cited Loci* and *Linked Books*

Giovanni Colavizza & Matteo Romanello

TOC

1. Citation indexes
2. Citation mining
3. The humanities
4. Cited Loci and Linked Books
5. The citation mining framework
6. ...

Citation Indexes

Google Scholar

Articles (include patents) Case law

Stand on the shoulders of giants

Citation Mining

Dopo la ritirata di Pipino l'imperatore di Costantinopoli inviò ad Aquisgrana, presso Carlo Magno, il legato Arsafio, il quale nell'812 stipulò un trattato col sovrano franco. In esso fu stabilito il riconoscimento a Carlo della dignità imperiale (ma non di imperatore dei "romani") da parte del *basileus*, in cam-

3. Giovanni Diacono, *Istoria Veneticorum*, a cura di L.A. Berto in *Cronache*, a cura di L.A. Berto e G. Fedalto, in *Corpus Scriptorum Ecclesiae Aquileiensis*, XII/2, Roma 2003, pp. 72-74..

4. Ivi, p. 74. G. Ostrogorsky, *Storia dell'Impero bizantino*, Torino 1968, p. 177; A. Carile, "*La presenza bizantina dell'alto Adriatico tra il VII e il IX secolo*" in AAA., XXVII (1985), pp. 107-129. Negli *Annales*. Eginardo attribuisce, invece, la vittoria ai franchi.

Grado originata dall'invasione longobarda (568), perché l'interrogativo posto da Massenzio riguardava direttamente la predicazione dell'evangelista Marco nella città di Aquileia[5]. Infatti, la risposta del rappresentante del patriarca di Grado Venerio fu dimessa e rassegnata: "*Negare non possumus et omnes sciunt Aquileiam civitatem primam et metropolim a beato Marco*

5. Sulla questione marciana cf. G. Fedalto, *Aquileia. Una chiesa due patriarcati*, Roma 1999, pp. 44-54; ID., *San Marco tra Babilonia, Roma e Aquileia, nuove ipotesi e ricerche*, in *San Marco: aspetti storici e agiografici*, Atti a cura di A. Niero, Venezia 1996, pp. 35-50. G. Biasutti, *La tradizione marciana aquileiese*, Udine 1950; G.C. Menis, *La XII lettera attribuita a sant'Ambrogio e la questione marciana aquileiese*, in *Rivista di storia della Chiesa in Italia*, XVIII (1964), pp. 243-253.

The Humanities

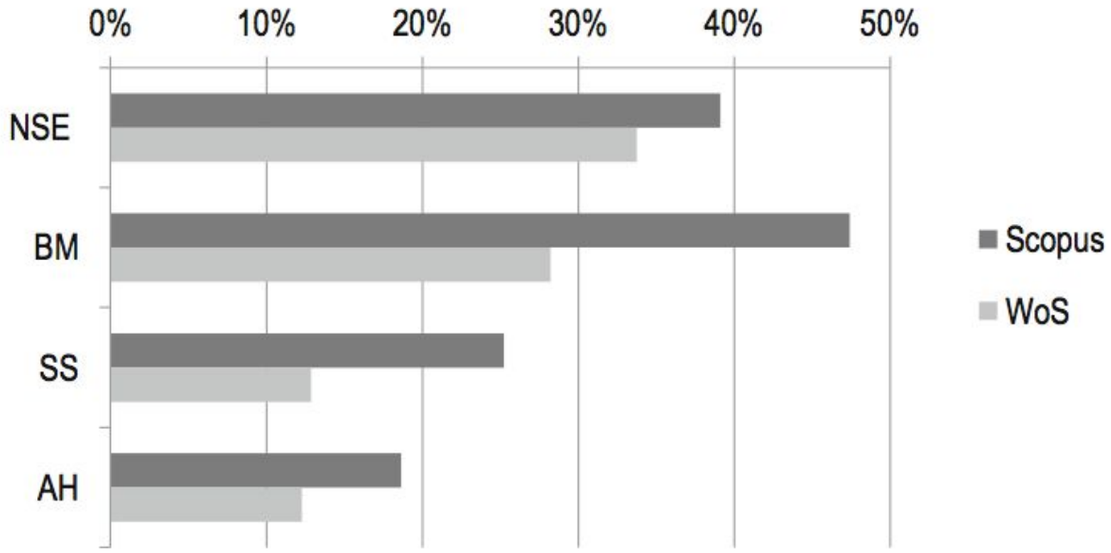
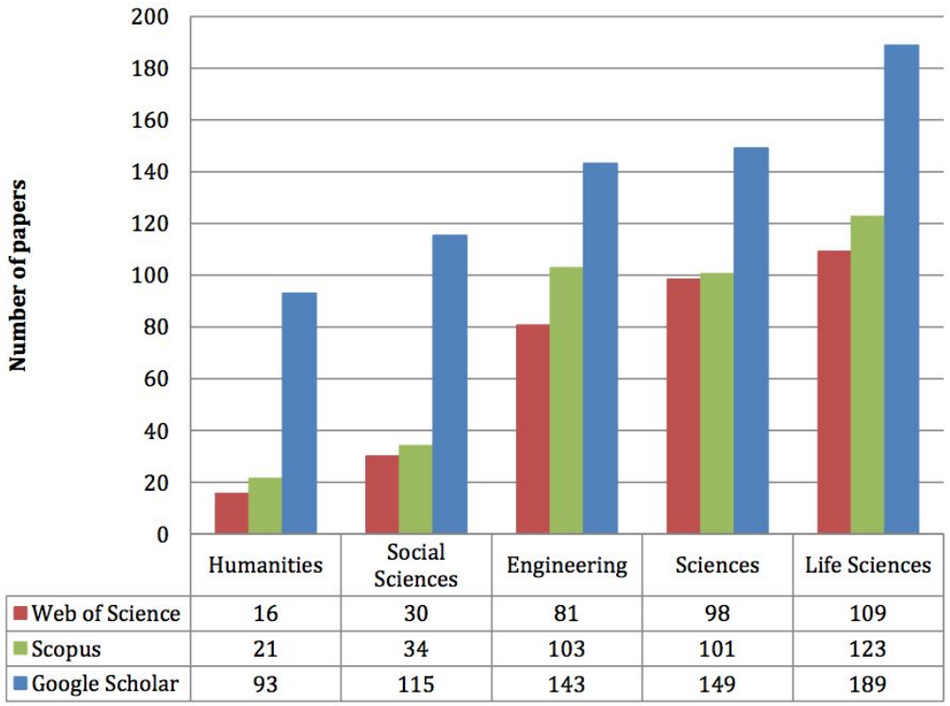


Fig. 1 Proportion of Ulrich academic journals indexed in Web of Science and Scopus



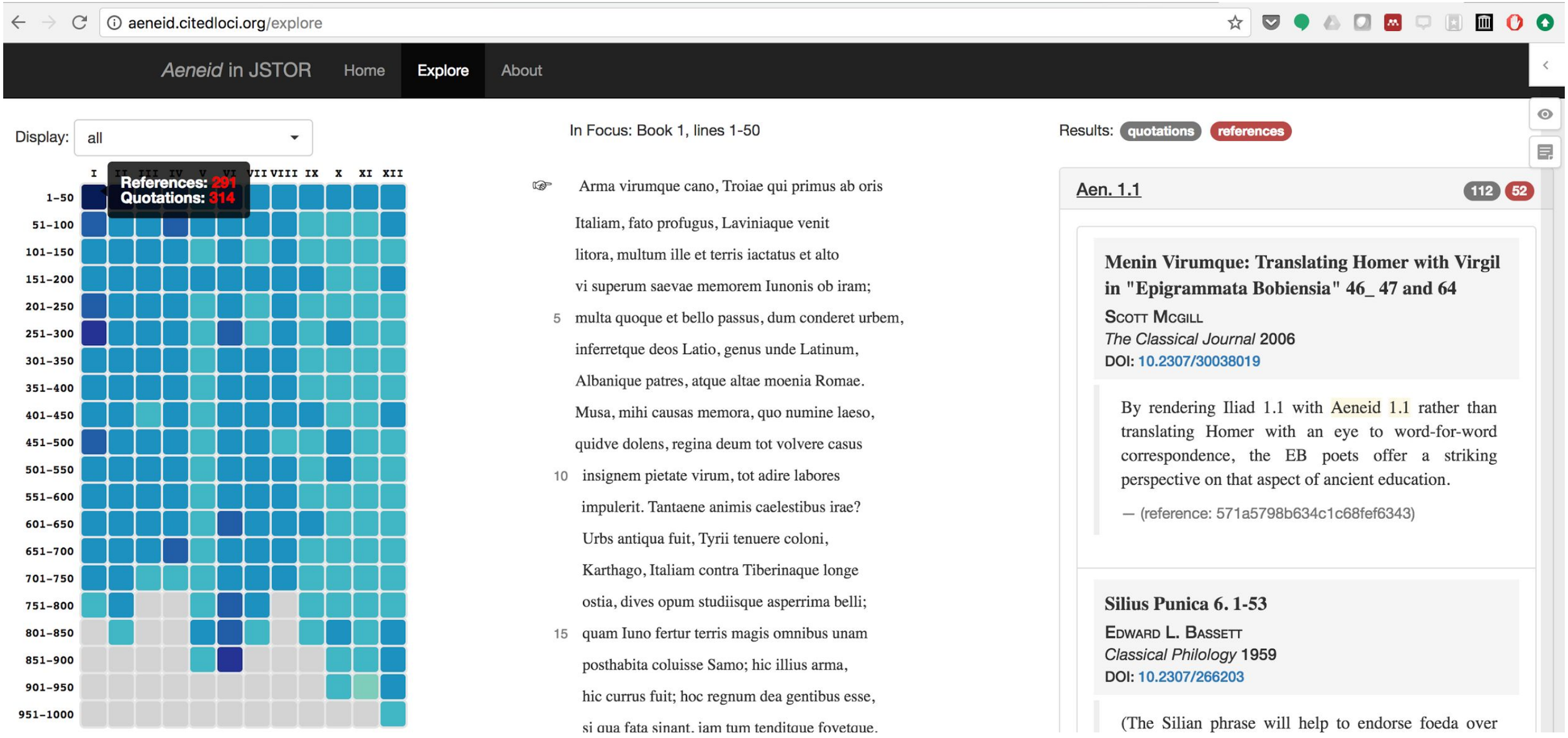
	Humanities	Social Sciences	Engineering	Sciences	Life Sciences
Web of Science	16	30	81	98	109
Scopus	21	34	103	101	123
Google Scholar	93	115	143	149	189

Fig. 3 Average number of papers per academic across five disciplines and three databases, July 2015

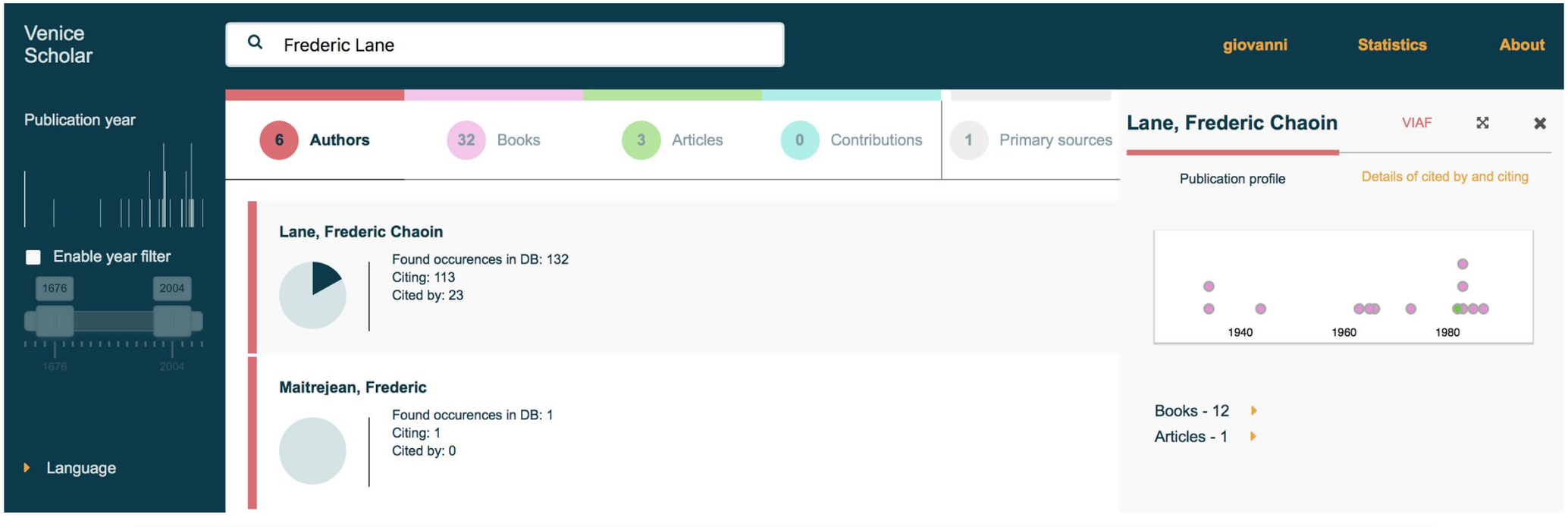
Harzing, Anne-Wil, and Satu Alakangas. 2016. “Google Scholar, Scopus and the Web of Science: A Longitudinal and Cross-Disciplinary Comparison.” *Scientometrics* 106 (2): 787–804. doi:10.1007/s11192-015-1798-9.

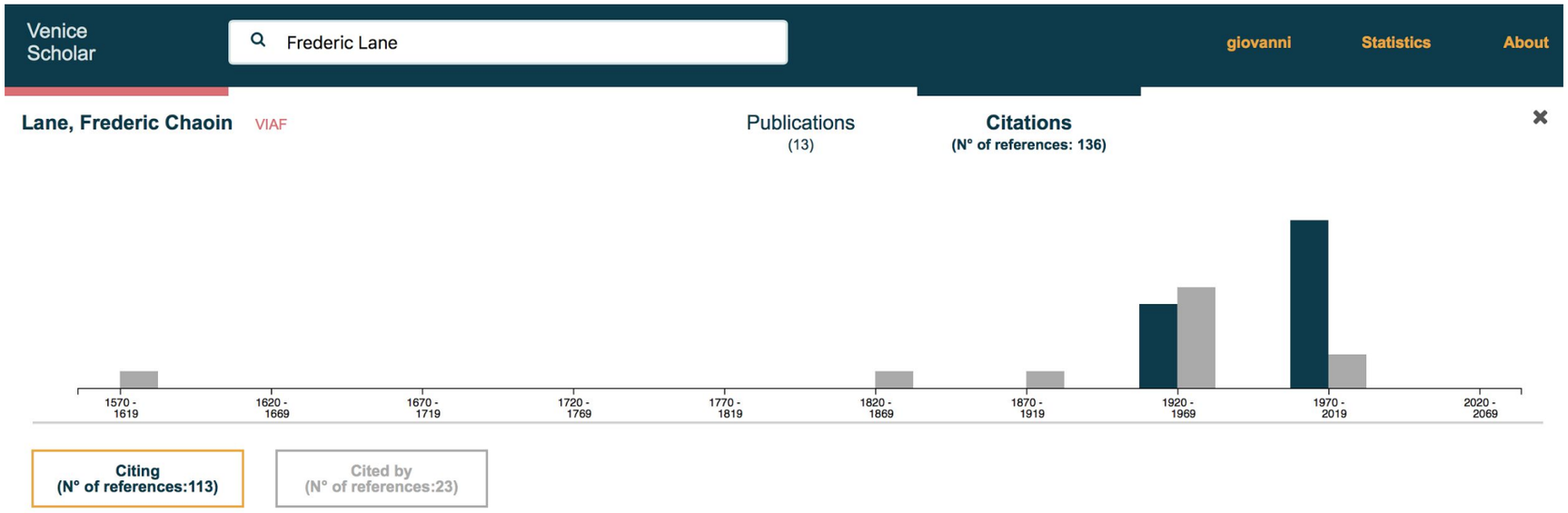
Mongeon, Philippe, and Adèle Paul-Hus. 2016. “The Journal Coverage of Web of Science and Scopus: A Comparative Analysis.” *Scientometrics* 106 (1): 213–28. doi:10.1007/s11192-015-1765-5.

Cited Loci


aeneid.citedloci.org/explore
Aeneid in JSTOR
Home
Explore
About
Display: all
I II III IV V VI VII VIII IX X XI XII
References: 291
Quotations: 314
1–50
51–100
101–150
151–200
201–250
251–300
301–350
351–400
401–450
451–500
501–550
551–600
601–650
651–700
701–750
751–800
801–850
851–900
901–950
951–1000
In Focus: Book 1, lines 1-50
Arma virumque cano, Troiae qui primus ab oris
Italiam, fato profugus, Laviniaque venit
litora, multum ille et terris iactatus et alto
vi superum saevae memorem Iunonis ob iram;
multa quoque et bello passus, dum conderet urbem,
inferretque deos Latio, genus unde Latinum,
Albanique patres, atque altae moenia Romae.
Musa, mihi causas memora, quo numine laeso,
quidve dolens, regina deum tot volvere casus
insignem pietate virum, tot adire labores
impulerit. Tantaene animis caelestibus irae?
Urbs antiqua fuit, Tyrii tenuere coloni,
Karthago, Italiam contra Tiberinaque longe
ostia, dives opum studiisque asperrima belli;
quam Iuno fertur terris magis omnibus unam
posthabita coluisse Samo; hic illius arma,
hic currus fuit; hoc regnum dea gentibus esse,
si qua fata sinant, iam tum tenditque fovetque.
Results: quotations references
Aen. 1.1
112
52
Menin Virumque: Translating Homer with Virgil in "Epigrammata Bobiensia" 46_ 47 and 64
Scott McGill
The Classical Journal 2006
DOI: 10.2307/30038019
By rendering Iliad 1.1 with Aeneid 1.1 rather than translating Homer with an eye to word-for-word correspondence, the EB poets offer a striking perspective on that aspect of ancient education.
— (reference: 571a5798b634c1c68fef6343)
Silius Punica 6. 1-53
Edward L. Bassett
Classical Philology 1959
DOI: 10.2307/266203
(The Silian phrase will help to endorse foeda over

Linked Books

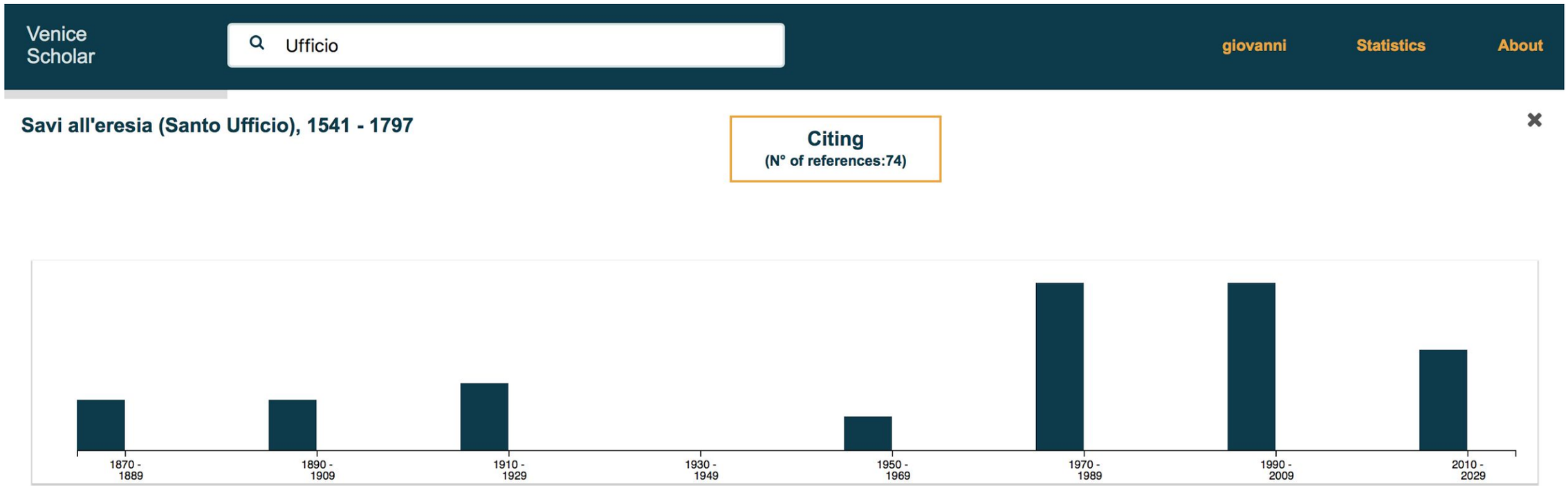




Books - 33

Title		Links	Author	Publisher	Year
Dalla galea alla nave di linea : le trasformazioni della marina veneziana (1572-1699)	Show reference (1)	VScholar | Catalog	Candiani, Guido	Città del Silenzio,	2012
Venice's Mediterranean colonies : architecture and urbanism	Show references (2)	VScholar | Catalog	Geōrgopoúlou, María 1961-	Cambridge university press,	2001
Glassmaking in Renaissance Venice : the fragile craft	Show reference (1)	VScholar | Catalog	McCray, Patrick	Ashgate,	1999
Glassmaking in Renaissance Venice : the fragile craft	Show reference (1)	VScholar | Catalog	McCray, Patrick	Ashgate,	1999
Money and banking in medieval and Renaissance Venice	Show reference (1)	VScholar | Catalog	Mueller, Reinhold Christopher	Johns Hopkins University pres...	1997

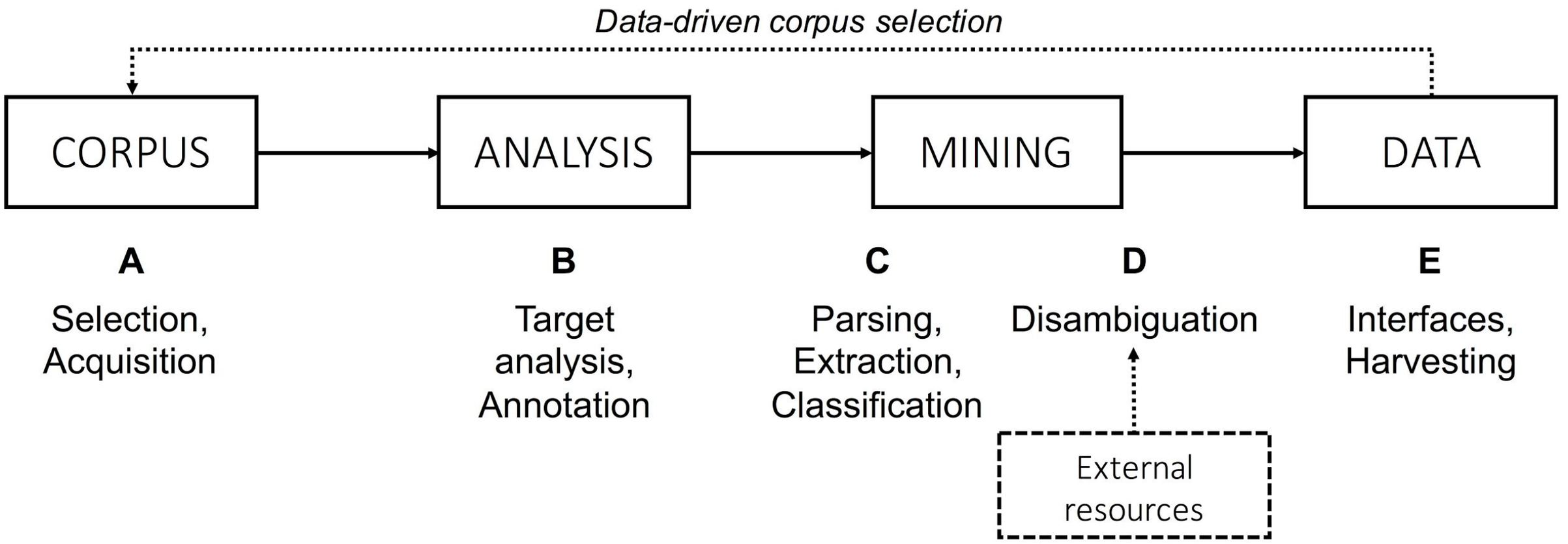
Linked Books



Articles - 66

Title		Links	Journal	Issue n°	Year
Una guaritrice processata dal Tribunale del Sant'Uffizio di Venezia nel Seicento per eresia ed esorc…	Show references (3)	VScholar | Catalog	Studi veneziani		2013
Donne imprenditrici nella Terraferma Veneta della prima età moderna (secoli XV-XVI)	Show reference (1)	VScholar | Catalog	Archivio veneto		2012
Il terzo visitatore nella biblioteca di Marin Sanudo il Giovane e nelle sue camere	Show reference (1)	VScholar | Catalog	Studi veneziani		2011

The Citation Mining Framework



A. Corpus Selection & Acquisition

	Linked Books	**Cited Loci**
Corpus' boundaries	Open corpus	JSTOR = corpus
Publication types	Monographs, journal articles	Journal articles
Corpus size	~2k monographs ~5.5k articles	~130k articles
Metadata source	Opac SBN, later enriched (e.g. with article-level metadata)	JSTOR API
Data	Digital images (.tiff, .jpg) + OCR	Plain text
Legal framework	Holding libraries keep copyright on materials	License agreement with JSTOR, DfR scheme

B. Target Analysis and Annotation

	Linked Books	**Cited Loci**
Types of cited sources	Primary, secondary (books, articles)	Primary (canonical texts)
Ad-hoc annotation campaign	Yes, 40k annotations	No
Reference style classification	Yes	No

- Archive of Venice, *Procuratori di San Marco*, *de citra*, *commissarie*, b. 1, c. 7
- ASVe, PSM, *de citra, commissarie*, b. 1, c. 7.
- Ivi, c. 8.

- canonical texts: e.g. Virgil, *Aen.* 12, 101-109, Hom. *Il*. 7.180
- fragmentary texts: e.g. FGrHist 688 F 13; Alc. fr. 34,1 Voigt
- inscriptions (e.g. CIL 3, 6174; AE 1991, 1405)
- papyri (e.g. PCair. inv. 10750)
- manuscripts (e.g. Vendôme, Bibl. mun. 31)
- museum objects (e.g. coins, vases, etc.)

C. Parsing, Extraction, Classification (LB)

1

(5) A.S.V., Provveditori sopra monasteri, b. 280; Riformatori dello Studio di Padova, f. 272.

2

- (5) ← out of reference
- Provveditori sopra monasteri ← record group
- b. 280; ← box
- Riformatori dello Studio di Padova ← record group
- f. 272 ← sheet

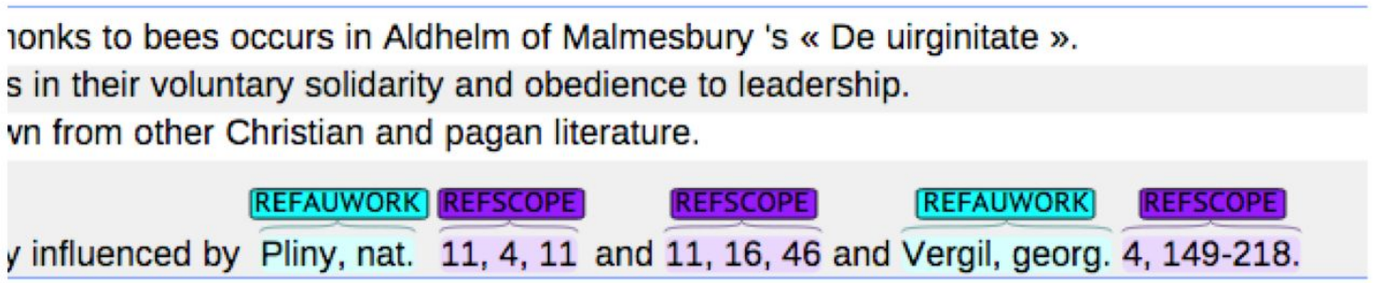
3

- (5) ← out of reference
- A.S.V., Provveditori sopra monasteri, b. 280; ← primary source
- Riformatori dello Studio di Padova, f. 272. ← primary source

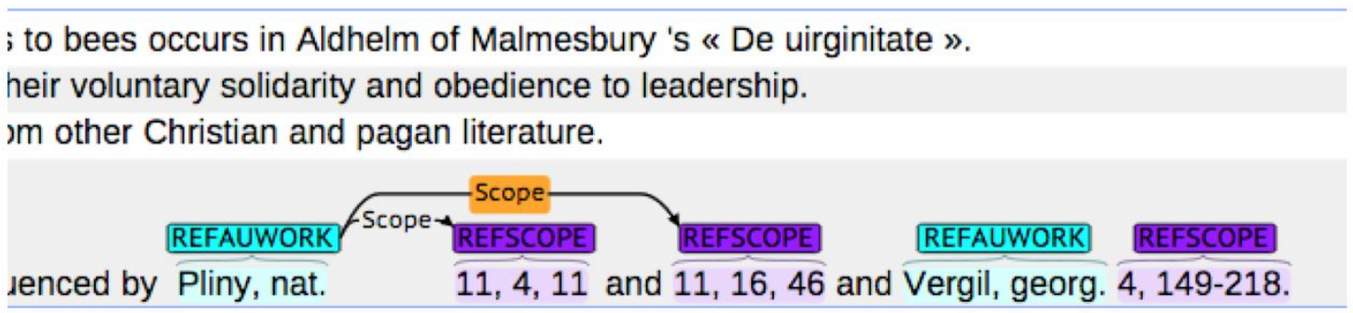
1 → **specific category tagger** → 2 → **generic category tagger** → 3

C. Parsing, Extraction, Classification (CL)

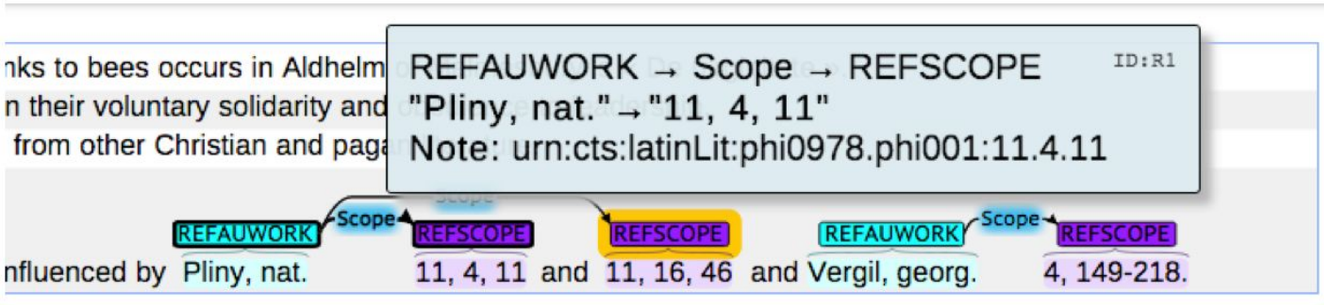
1



2



3



0 → **extraction of cit. component** → 1 → **relation extraction** → 2 → **disambiguation** → 3

D. Disambiguation of References (LB)

Author mentions:

1. Lookup on internal DB
2. Search in VIAF (API)
3. Otherwise, add new record to DB

Primary references:

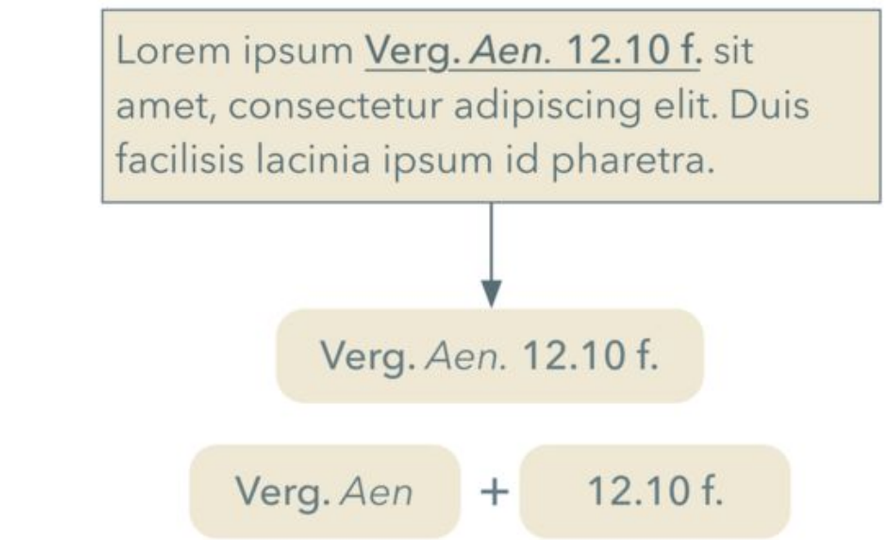
1. Lookup on internal DB
2. SiASVe lookup
   - ML-based classifier
   - If low confidence, roll back to Rule Based classifier

Secondary references:

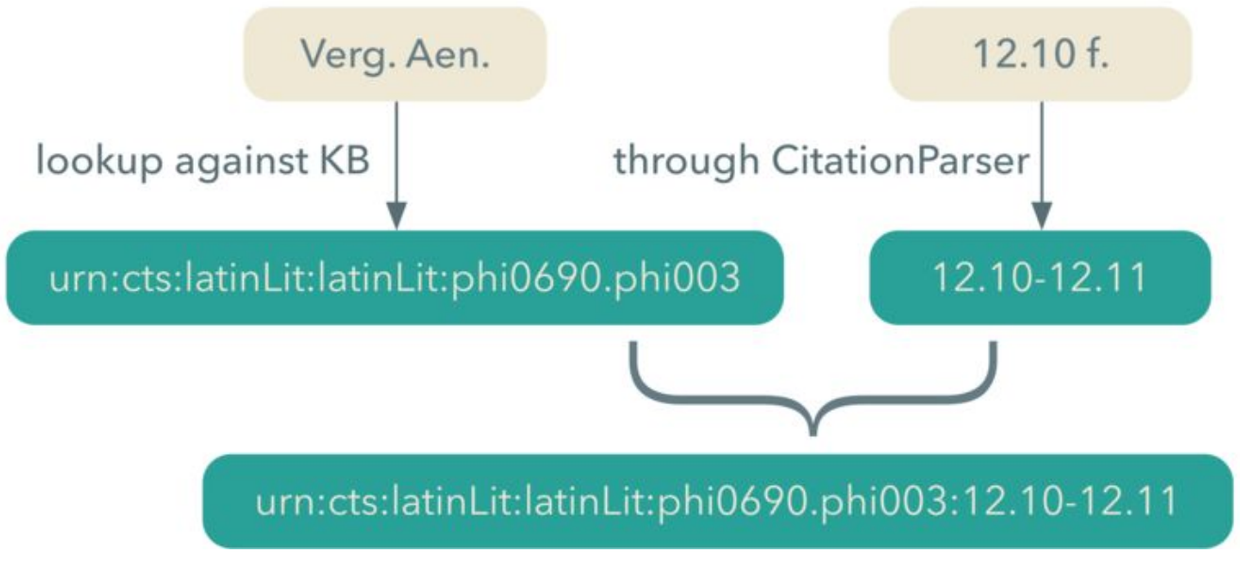
1. Lookup on internal DB
2. Lookup on OPAC ISBN
3. Otherwise, add new record to DB

D. Disambiguation of References (CL)

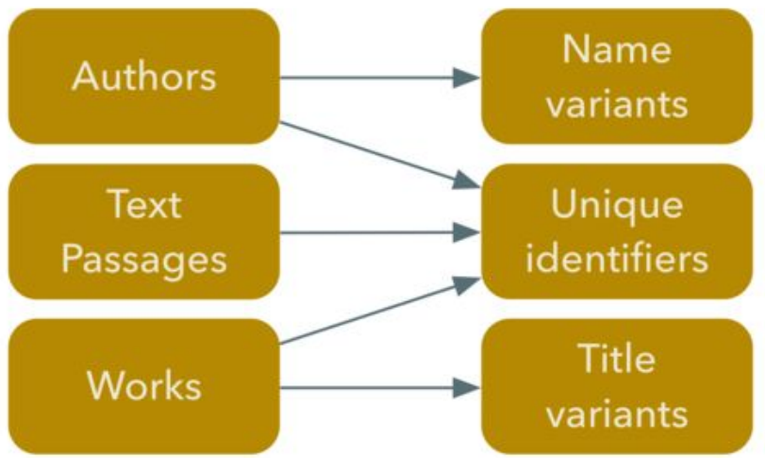
1 Citation Extractor



2 Citation Matcher

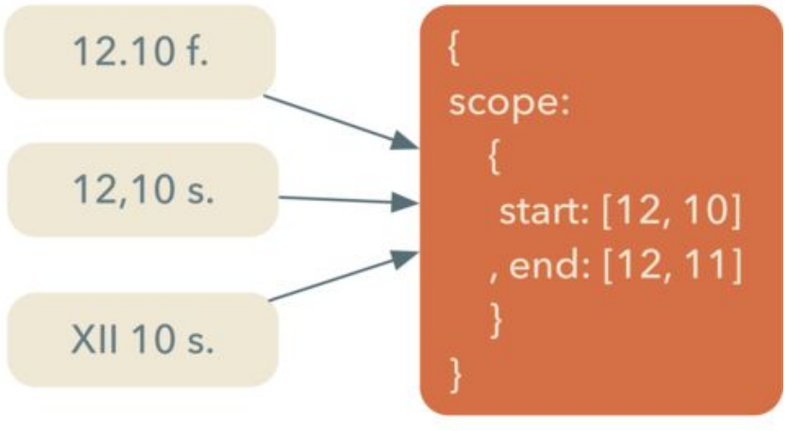


3 Knowledge Base



(1,500 authors; 5,500 works)

4 Citation Parser

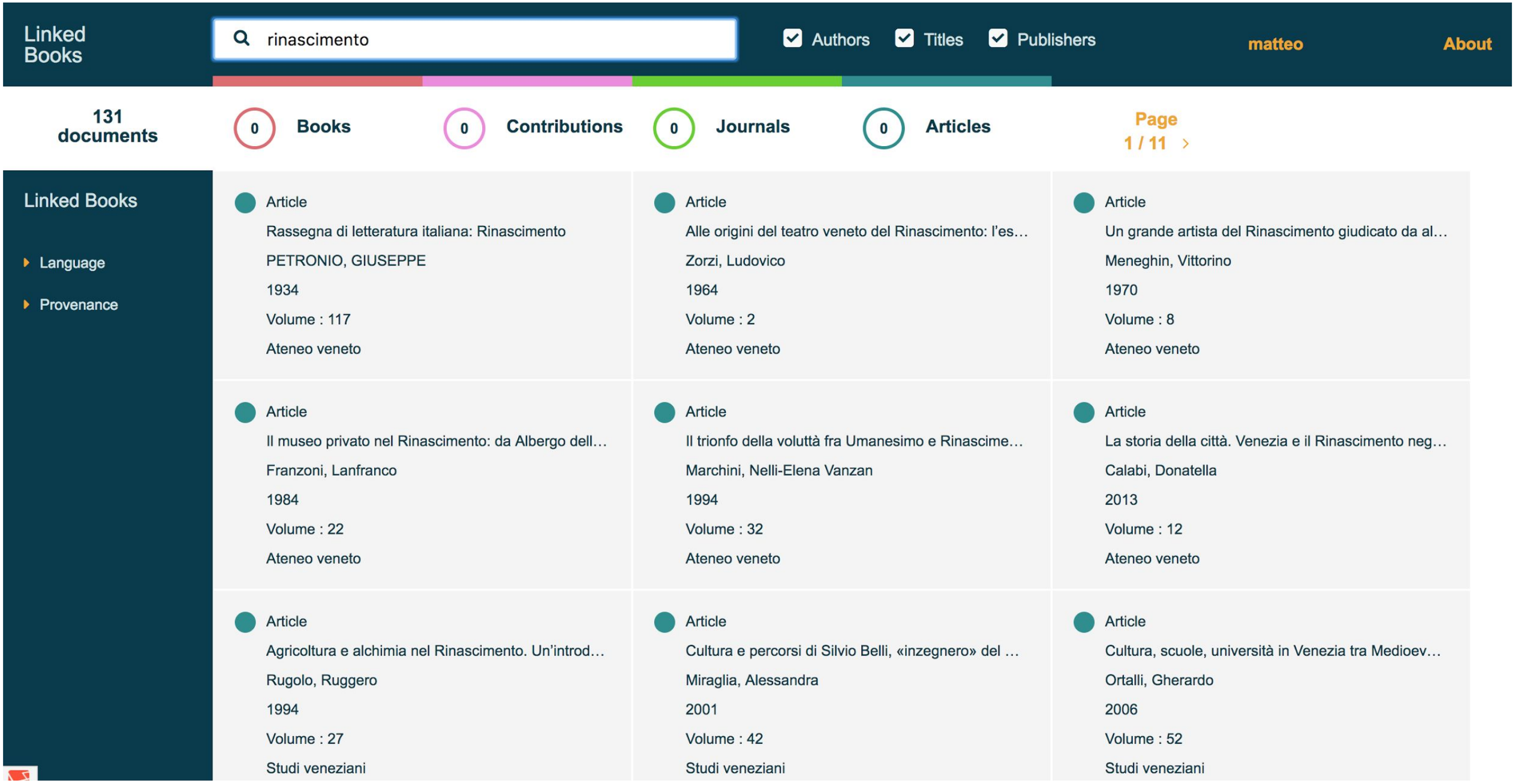


(Context-Free Grammar)

E. Publication & Project Outputs

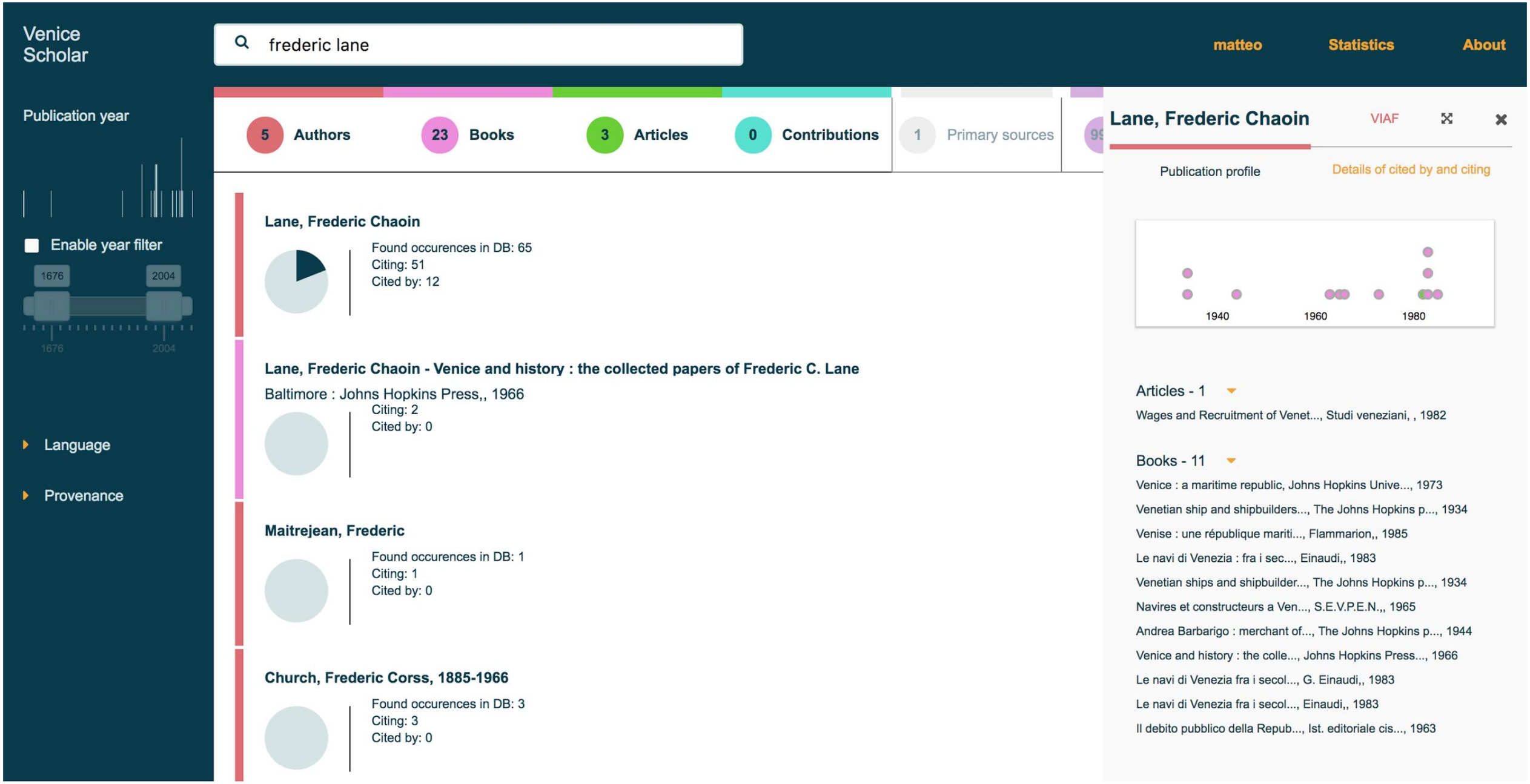
	Linked Books	**Cited Loci**
Code	Yes (partly)	Yes
Annotation Data	Yes	No
Citation dataset	Yes (in progress)	Yes (in progress)
API	Yes	Yes (in progress)

E. Publication & Project Outputs (LB)



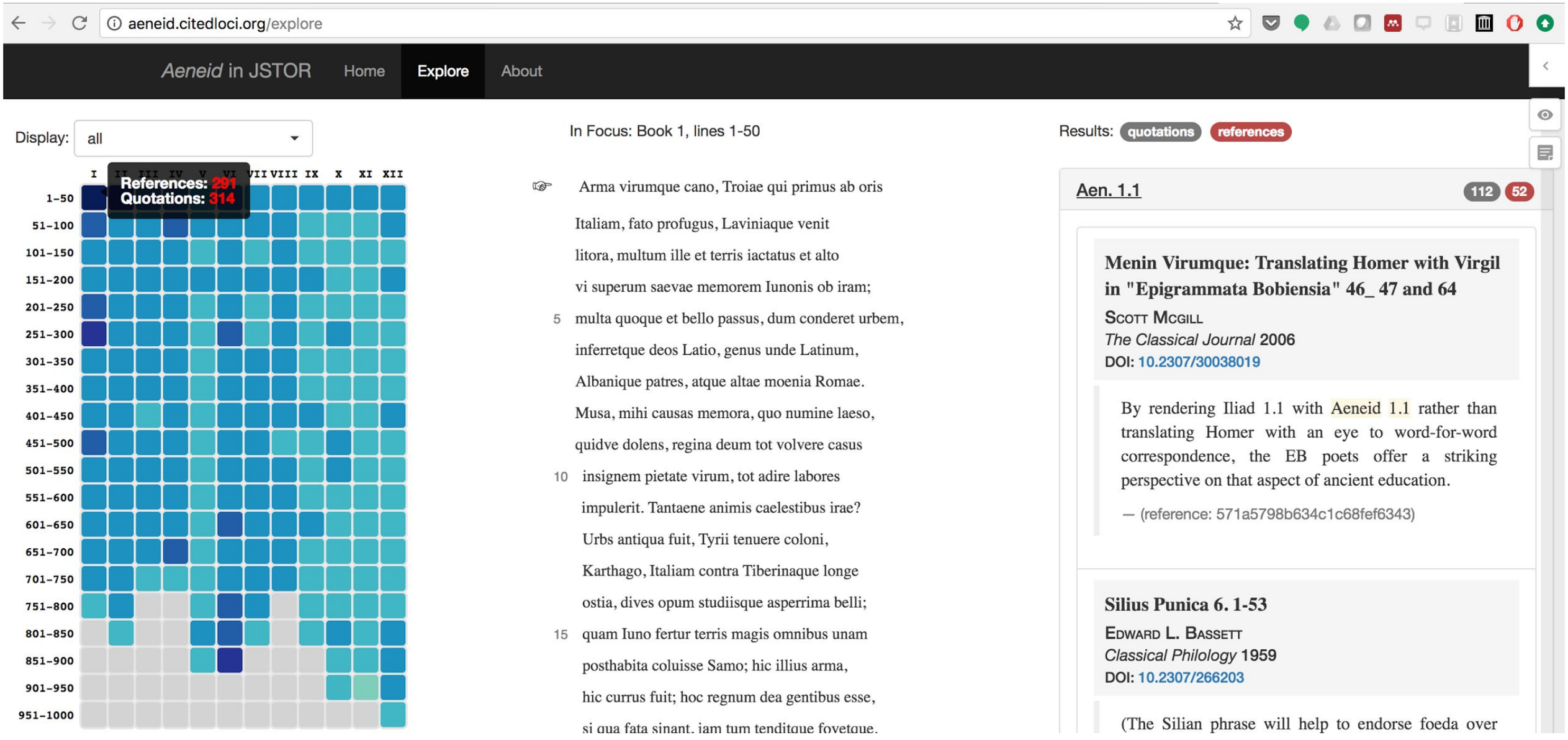
Linked Books Catalog

E. Publication & Project Outputs (LB)



Venice Scholar

E. Publication & Project Outputs (CL)



http://aeneid.citedloci.org/

Open Challenges

1. Issues intrinsic to the Humanities:
   - Fragmentation and variety of literature
   - Lack of incentives
   - Lack of central actors either private or public

2. Infrastructural issues:
   - Availability of digital corpora
   - Accessibility of catalog metadata
   - Little adoption of standards to expose metadata uniformly
   - Limited availability of APIs for harvesting (or data dumps)
   - Legal frameworks for the use and publication of data
     - issues related to copyright on digitised materials
     - No copyright exception for text mining on contemporary scholarly publications

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Thank you!

https://www.researchgate.net/project/Linked-Books
https://www.researchgate.net/project/Cited-Loci

giovanni.colavizza
matteo.romanello
{@epfl.ch}

Giovanni Colavizza & Matteo Romanello