What's in a corpus? Utilizing metadata in Latin and Greek text collections

Neven Jovanović

University of Zagreb

neven.jovanovic@ffzg.hr

Greek and Latin text collections

Greek and Latin

Perseus (internet, free access)

Greek

TLG (Thesaurus linguae Graecae; CD + internet); **PHI** (Greek inscriptions, documentary papyri; CD + internet, commercial)

Latin

Bibliotheca Teubneriana Latina (CD, commercial); *Library of Latin Texts* (CLCLT5; CD, commercial); *PHI Latin library* (CD + internet, commercial); *IntraText Digital Library* (internet, free access); *The Latin Library* (internet, free access); *Itinera electronica* (internet, free access); *Thesaurus Linguae Latinae* (a dictionary; CD, commercial)

What Greek and Latin text collections are not

A corpus is a collection of pieces of language text in electronic form, selected according to external criteria to represent, as far as possible, a language or language variety as a source of data for linguistic research.

(Sinclair 2005)

Maximize number of users
Maximize number of uses

… a library?

But libraries have *catalogues*.
Catalogues *enhance* libraries.

Users of Greek and Latin text collections

Learners

Researchers

A learner's experience

A researcher's experience

A proposal

Design a collection of texts in such a way to:

a) help learners orientate, and learn what is inside

b) help researchers ask complex questions

Questions expected

— In which metre are those poems?

— How do I search just the poems in hendecasyllables?

— Which texts in the collection are letters?

— How do I search just the letters in the collection?

— Which texts in the collection were produced in first century b. C?

— How do I search just the texts produced in first century b. C?

Problems expected

— There are too many texts!

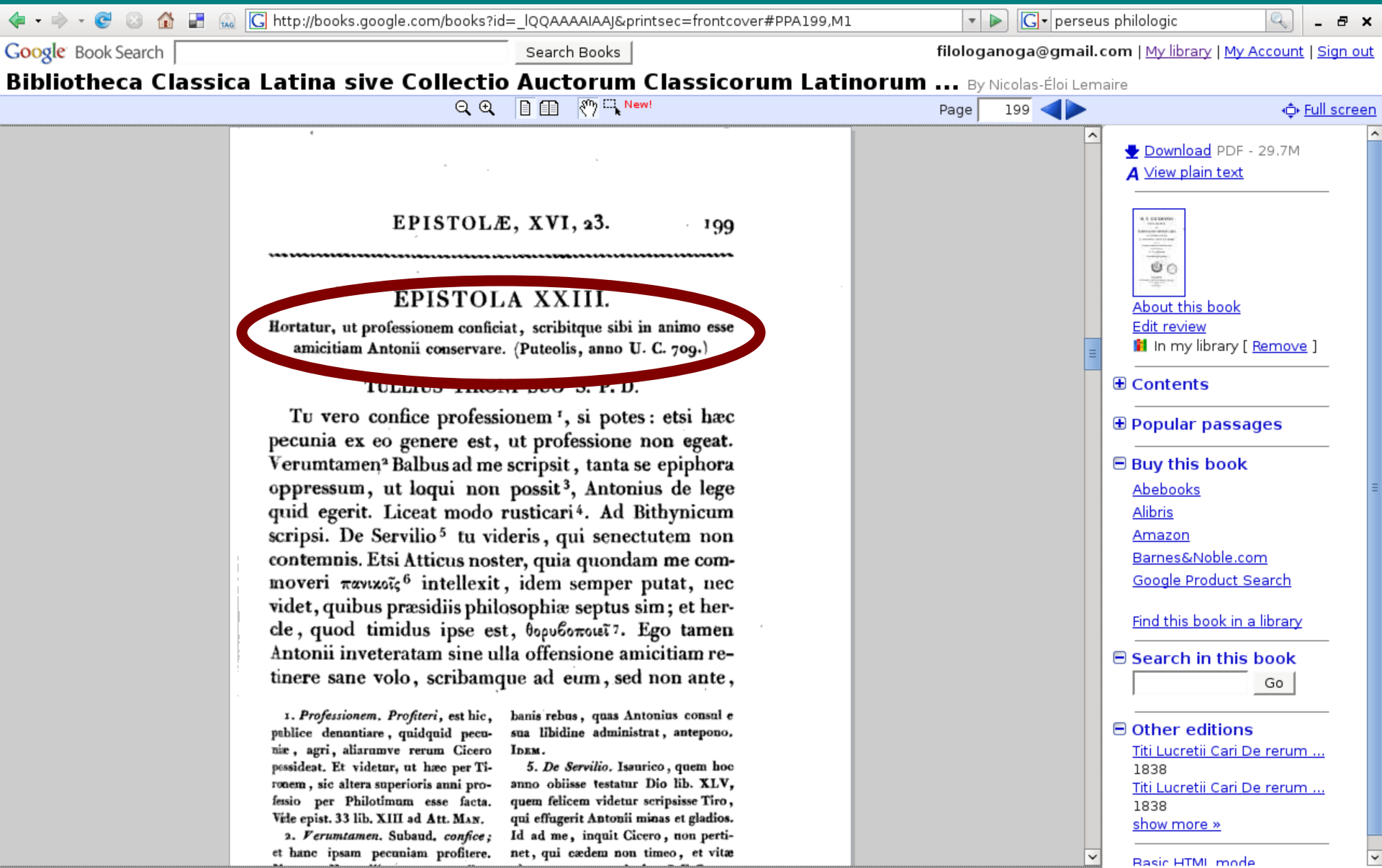
— How do we find metadata?

— How do we actually *do* it?

— Where do we find an army of coders?

What is already around?

(Old) scholarship as source of metadata



EPISTOLA XXIII.

Hortatur, ut professionem conficiat, scribitque sibi in animo esse amicitiam Antonii conservare. (Puteolis, anno U. C. 709.)

TULLIUS TIRONI SUO S. P. D.

Tu vero confice professionem [1], si potes: etsi hæc pecunia ex eo genere est, ut professione non egeat. Verumtamen [2] Balbus ad me scripsit, tanta se epiphora oppressum, ut loqui non possit [3], Antonius de lege quid egerit. Liceat modo rusticari [4]. Ad Bithynicum scripsi. De Servilio [5] tu videris, qui senectutem non contemnis. Etsi Atticus noster, quia quondam me commoveri πανικοῖς [6] intellexit, idem semper putat, nec videt, quibus præsidiis philosophiæ septus sim; et hercle, quod timidus ipse est, θορυβοποιεῖ [7]. Ego tamen Antonii inveteratam sine ulla offensione amicitiam retinere sane volo, scribamque ad eum, sed non ante,

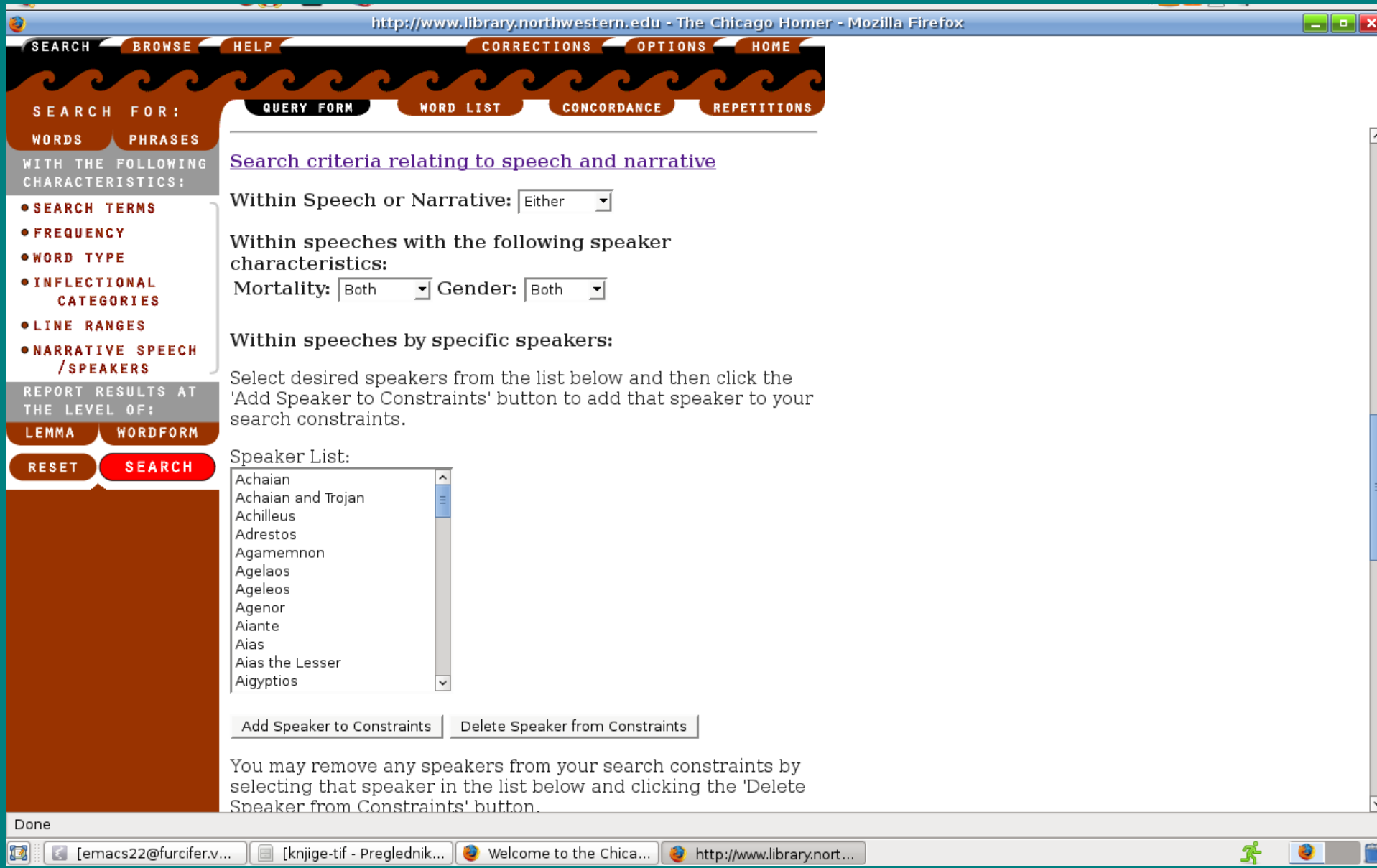
1. *Professionem. Profiteri*, est hic, publice denuntiare, quidquid pecuniæ, agri, aliarumve rerum Cicero possideat. Et videtur, ut hæc per Tironem, sic altera superioris anni professio per Philotimum esse facta. Vide epist. 33 lib. XIII ad Att. Man.

2. *Verumtamen.* Subaud. *confice*; et hanc ipsam pecuniam profitere.

banis rebus, quas Antonius consul e sua libidine administrat, antepono. Idem.

5. *De Servilio.* Isaurico, quem hoc anno obiisse testatur Dio lib. XLV, quem felicem videtur scripsisse Tiro, qui effugerit Antonii minas et gladios. Id ad me, inquit Cicero, non pertinet, qui cædem non timeo, et vitæ

Chicago Homer



Search criteria relating to speech and narrative

Within Speech or Narrative: Either

Within speeches with the following speaker characteristics:

Mortality: Both **Gender:** Both

Within speeches by specific speakers:

Select desired speakers from the list below and then click the 'Add Speaker to Constraints' button to add that speaker to your search constraints.

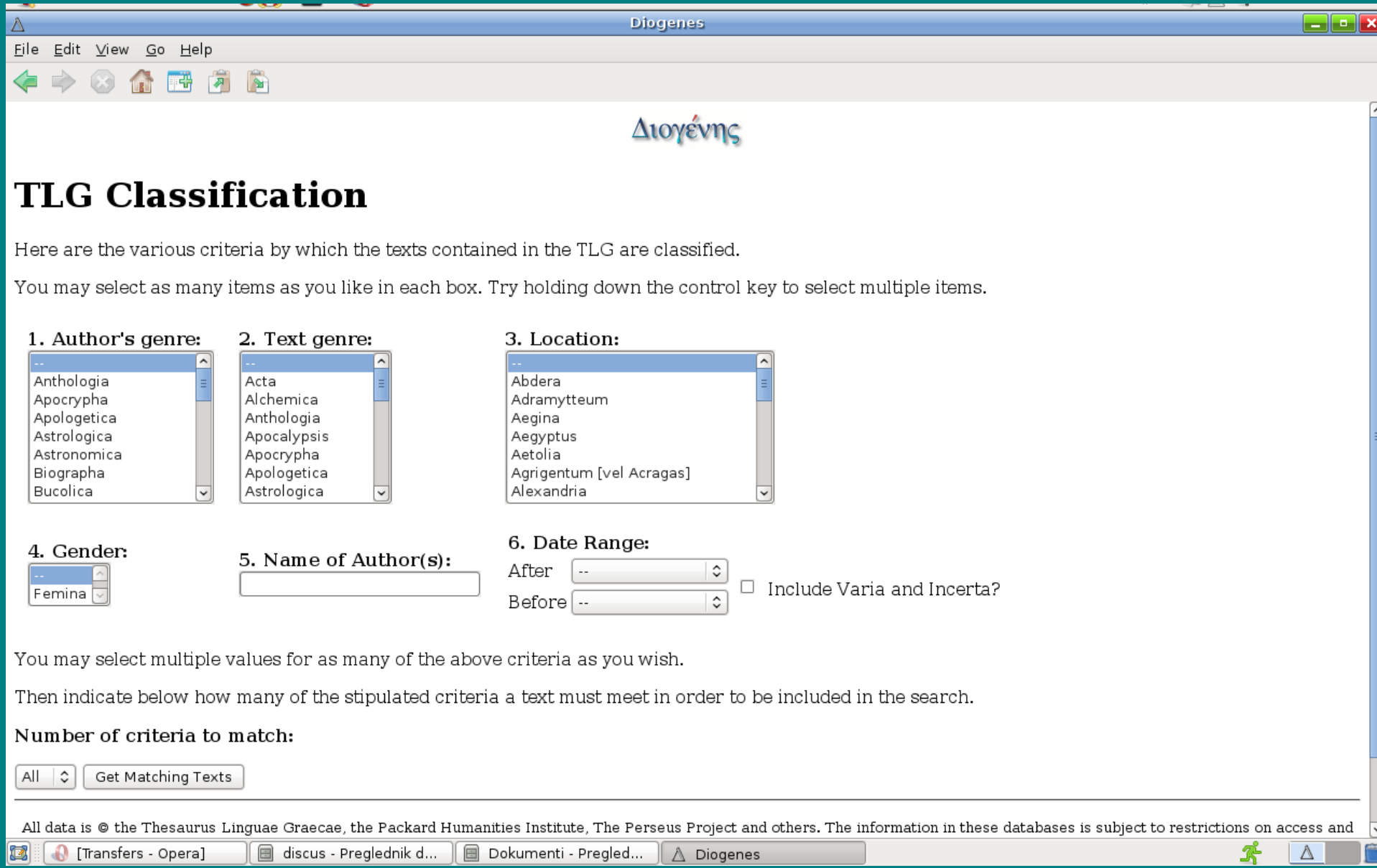
Speaker List:

- Achaian
- Achaian and Trojan
- Achilleus
- Adrestos
- Agamemnon
- Agelaos
- Ageleos
- Agenor
- Aiante
- Aias
- Aias the Lesser
- Aigyptios

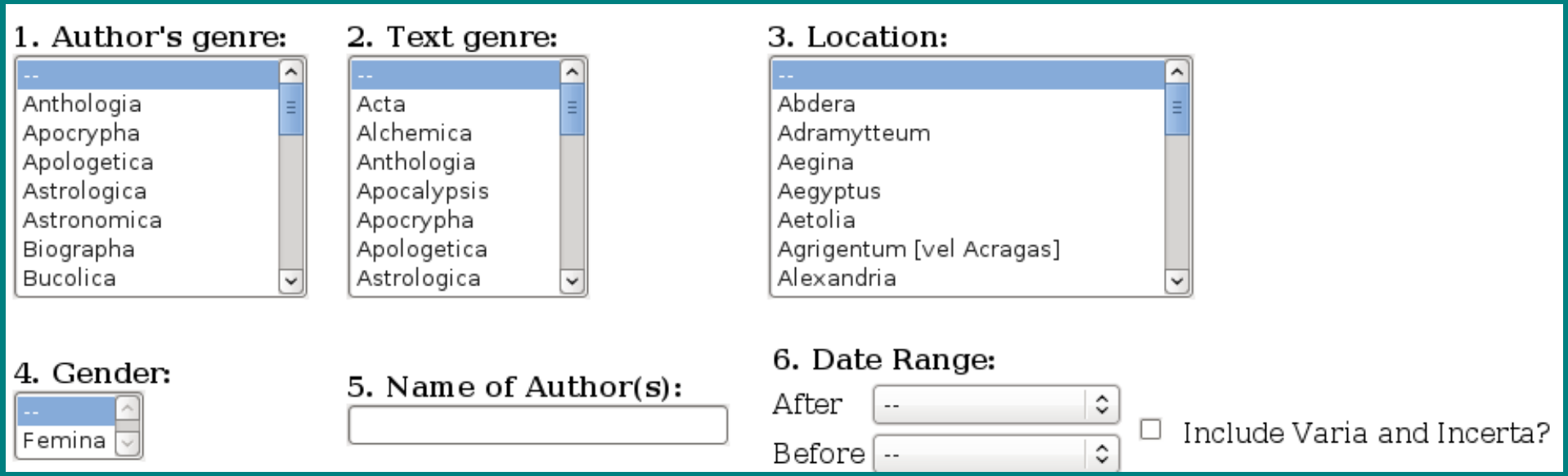
Add Speaker to Constraints | Delete Speaker from Constraints

You may remove any speakers from your search constraints by selecting that speaker in the list below and clicking the 'Delete Speaker from Constraints' button.

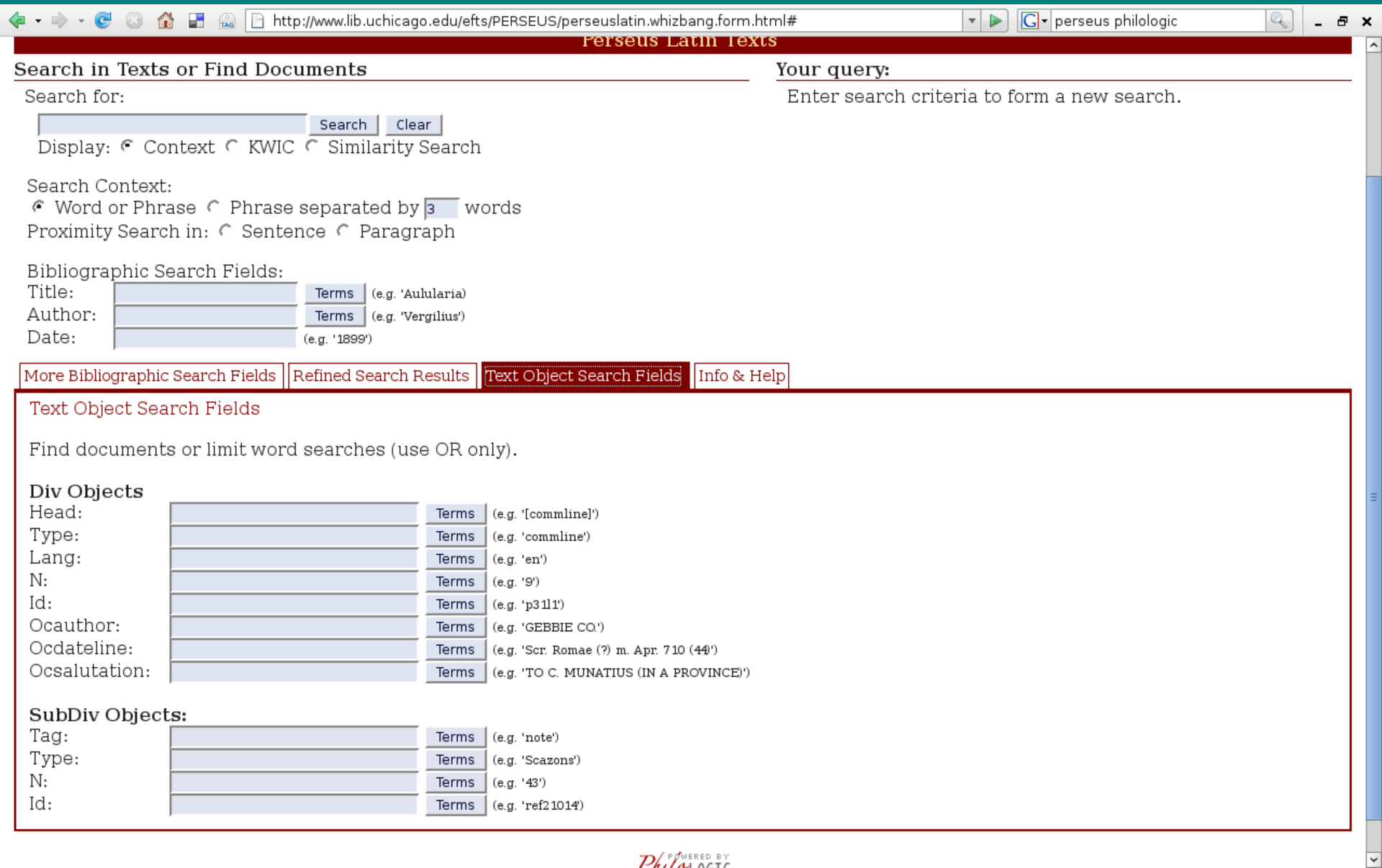
TLG / PHI with Diogenes



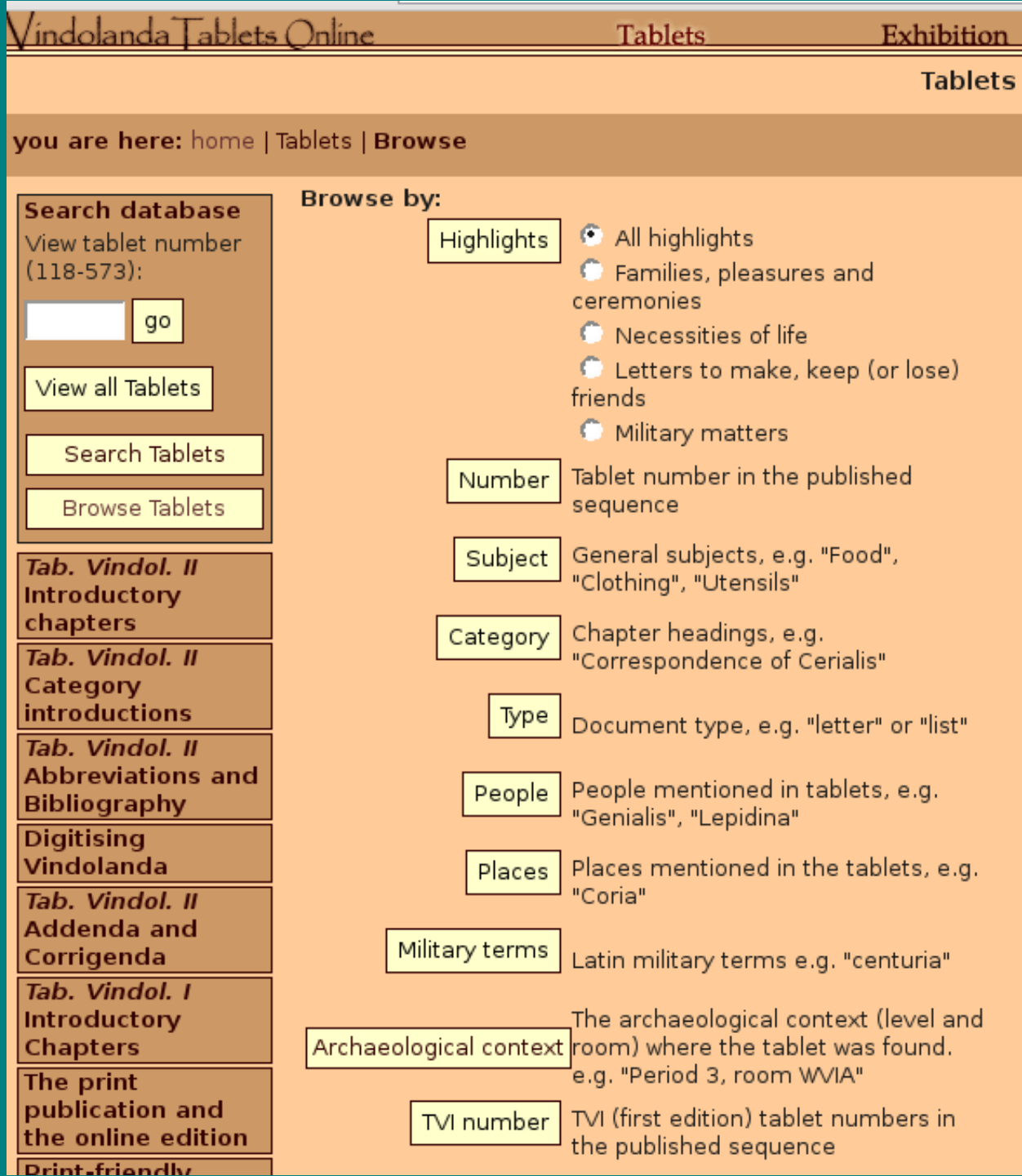
TLG / PHI with Diogenes



Perseus under PhiloLogic



Vindolanda tablets online



Croatiae auctores Latini (CAuLa)

- ca. 300.000 words pilot
- short texts, long texts, poetry, prose, literature, functional texts (e. g. notarial documents)
- until now: uncentralised, undigitised, sometimes unindexed, not easily (world-wide) accessible or searchable, not always reliably edited…

Croatiae auctores Latini (CAuLa)

Search and browse by:

- Auctores (A-Z)
- Tempora (e. g. 1400-1950)
- Loca (e. g. Dubrovnik, Split, Trogir)

Croatiae auctores Latini (CAuLa)

Search and browse by:

- Genera
  - Poesis
  - Prosa

Croatiae auctores Latini (CAuLa)

- Genera
  - Poesis
    - epica
    - elegiaca
    - epigrammata
    - eclogae
    - saturae

Croatiae auctores Latini (CAuLa)

- Themata
  - funeraria
  - amicitia
  - amores
  - antiturcica
  - …

Croatiae auctores Latini (CAuLa)

- Damjan Beneša (Dubrovnik, around 1500),
- *De morte Christi* (10 books, 8300+ verses)
  - Liber I
    - Opening scene *(vv. 1-30).* Before Easter: everywhere sorrow. The poet thinks about faraway places, about Christ's passion and death. Jerusalem: Christ is being taken to Pilates' palace. The poet sees a vision of Christ hanging on the cross, his Mother grieving
    - Invocation *(vv. 31-43):* one who sings about Christ will earn a place in heaven; why did the Virgin bear a son, etc.

CAuLa: sample queries

What did people write about when they wrote in Latin in Split between 1500 and 1600?

How did poetry about friendship look like in Dubrovnik between 1500 and 1600?

In what types of texts is word *arma* used? Are there types of texts that do not use this word?

...

What do ~~we~~ I need?

— *Caveat*: a theoretically simple task may get quite untractable in real life (standards? searches? references? openness? computer science? etc.)

— If possible, use tools that already exist (learn about them)

— If possible, connect with projects that already exist (idem)

— Attract users, who will also help keep the project alive (corrections? reviews? research? teaching?)

— Hear what others think!