

✦ Finding our LODestone: Evaluating Linked Open Data for Qualitative Research



Kate Topham, Michigan State University ✦
ALA Linked Data Interest Group March 2023

Schoenberg Database of Manuscripts

The Database:

- Observations of premodern manuscripts
- 14,000 catalogs, inventories, and crowdsourced observations
- SPARQL Endpoint

The Goals:

- Investigate transfer of Classical Latin premodern manuscripts
- Visualize networks of provenance
- Upload authority data to Wikidata
- Clean and improve Schoenberg authority data



<https://sdbm.library.upenn.edu/>

✦ The Data

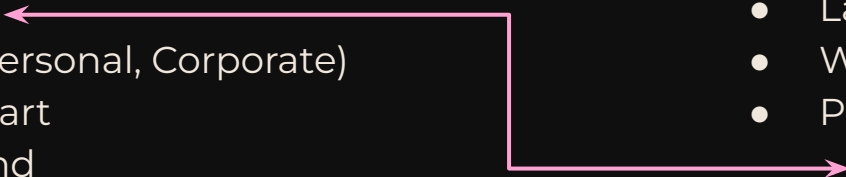


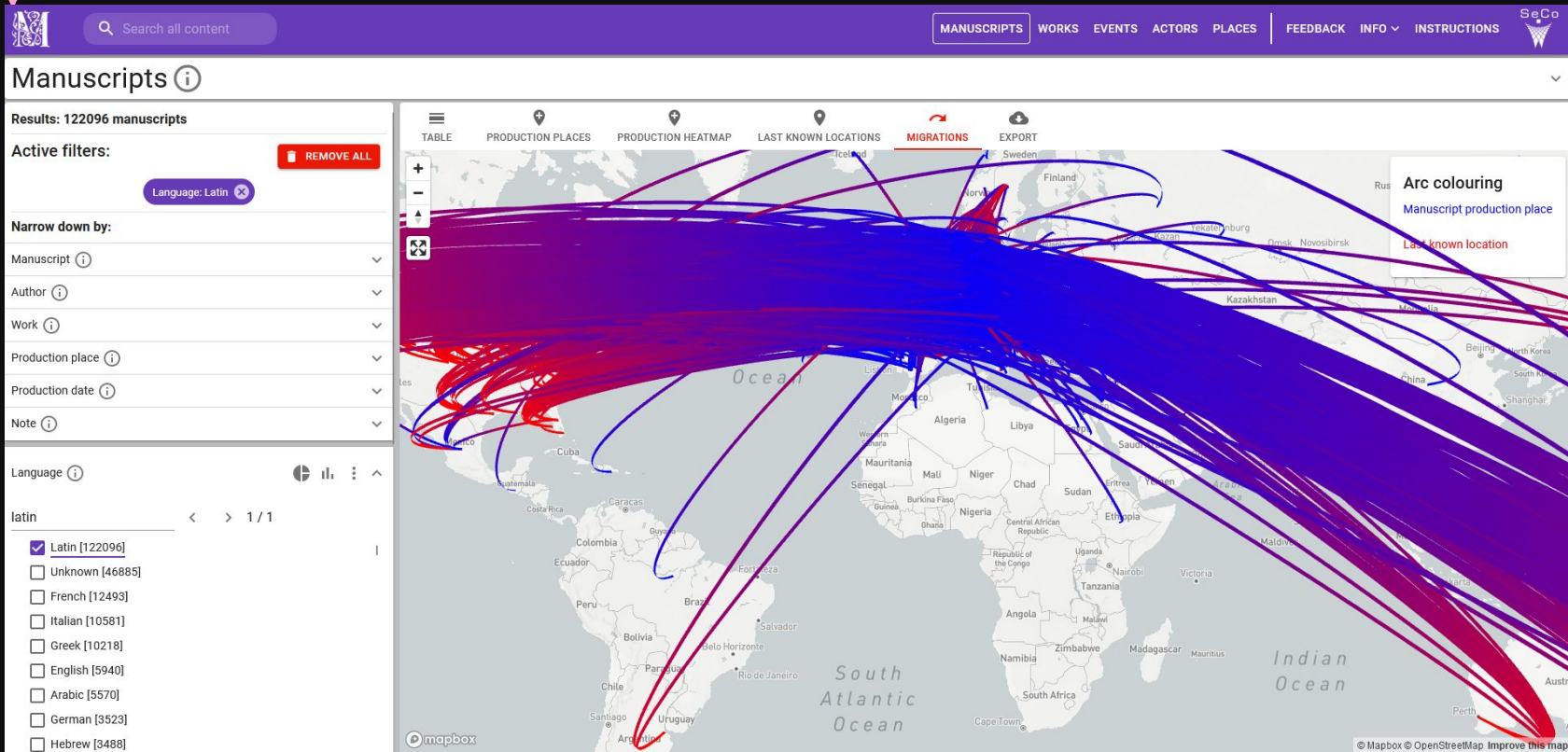
Agents

- Name
- Type (Personal, Corporate)
- Date Start
- Date End
- Places/Nationalities (dates)
- VIAF

Manuscripts

- Language
- Work(s) & Authors
- Provenance
 - Buyer & Seller
 - Date
 - Place





Mapping Manuscript Migrations Project

SPARQL Queries

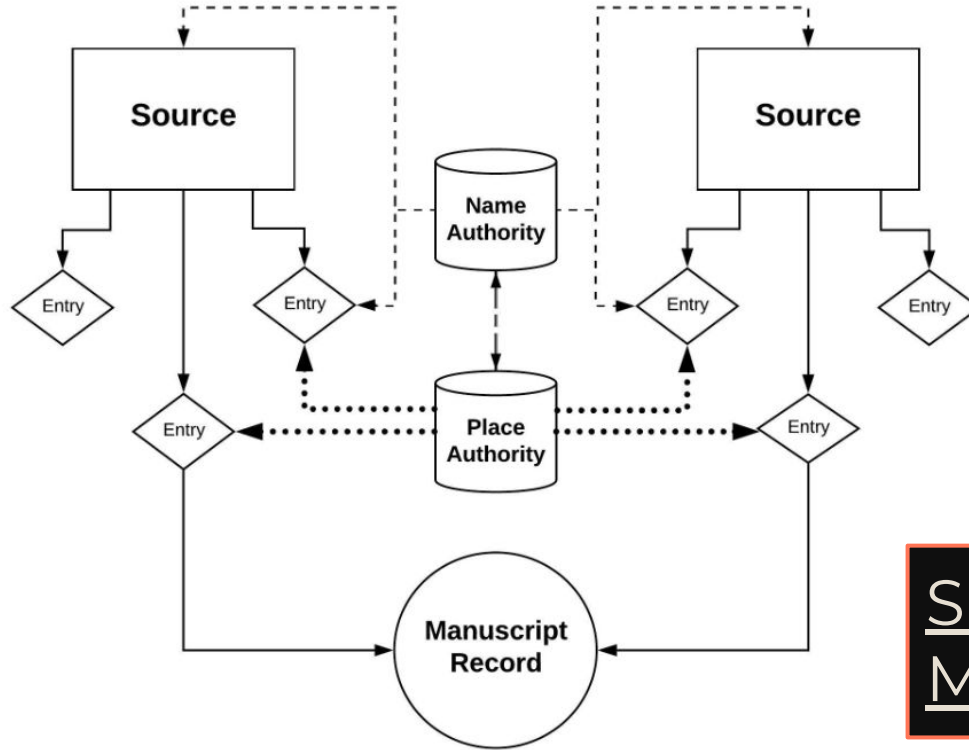
- Agent name & id
- Agent dates
- Agent place (& dates)
- Author name & id
- Manuscript id

Filtered by manuscripts in Latin, with
authors born between 190 BCE - 180 CE
and died between 75 BCE - 300 CE

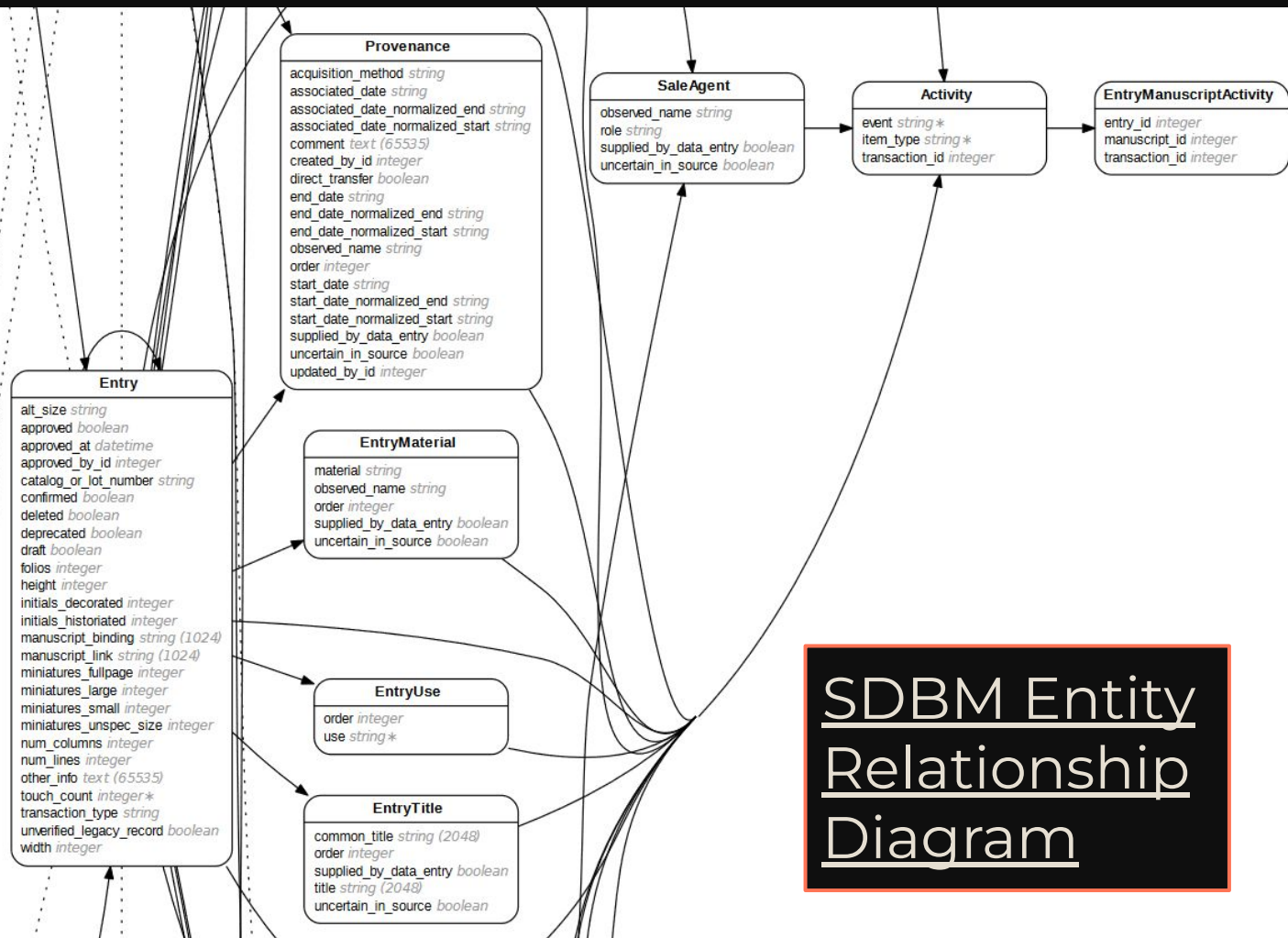
prefix xsd: <<http://www.w3.org/2001/XMLSchema#>>
prefix sdbm: <<https://sdbm.library.upenn.edu/>>

```
SELECT ?authorid ?authorname ?author_start ?author_end  
WHERE {  
    BIND (<https://sdbm.library.upenn.edu/language>  
        ?entry_lang_id sdbm:entry_languages_language_id  
        record.get entry ids for latin Triple is: entry_lang_id entry_lang_id  
        ?entry_lang_id sdbm:entry_languages_entry_id ?entry_lang_id  
        ?entry_author_id sdbm:entry_authors_author_id  
        ?entry_author_id sdbm:entry_authors_entry_id ?entry_authors_entry_id  
        ?authorid sdbm:names_name ?authorname .  
    OPTIONAL {  
        ?authorid sdbm:names_startdate ?author_start  
        ?authorid sdbm:names_enddate ?author_end .  
        ?author_nameplace_id sdbm:name_places_nameplace_id  
        ?author_nameplace_id sdbm:name_places_place_id  
        ?author_placeid sdbm:places_name ?author_placeid .  
    }  
  
    FILTER(xsd:integer(substr(concat(replace(?author_start,' ','0'),  
02000000)) .  
    FILTER(xsd:integer(substr(concat(replace(?author_end,' ','0'),  
    FILTER(?author_start!='0')  
}
```

SDBM Data Model



SDBM Data
Model



SDBM Entity Relationship Diagram

Wikidata

- OpenRefine!
- Cleaning
- Disambiguating
- VIAF IDs
 - Misattribution
 - Outdated ids
 - Duplicates

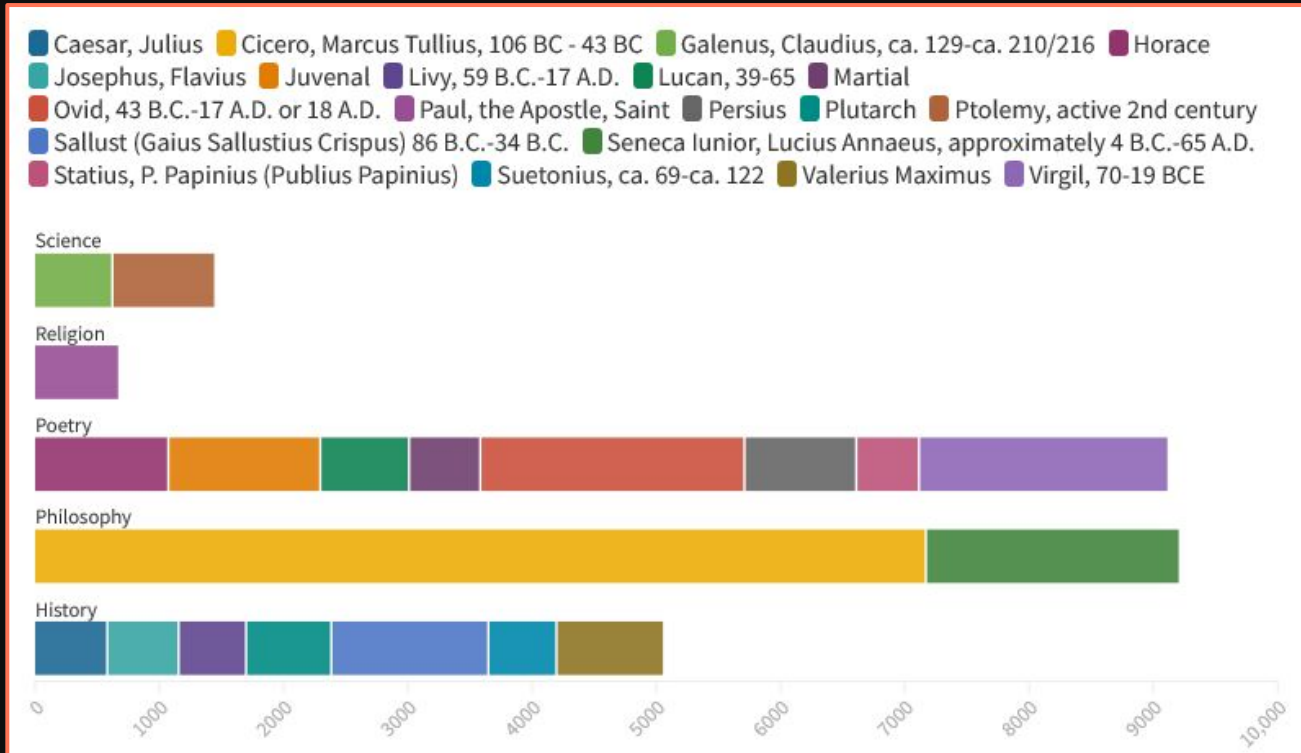




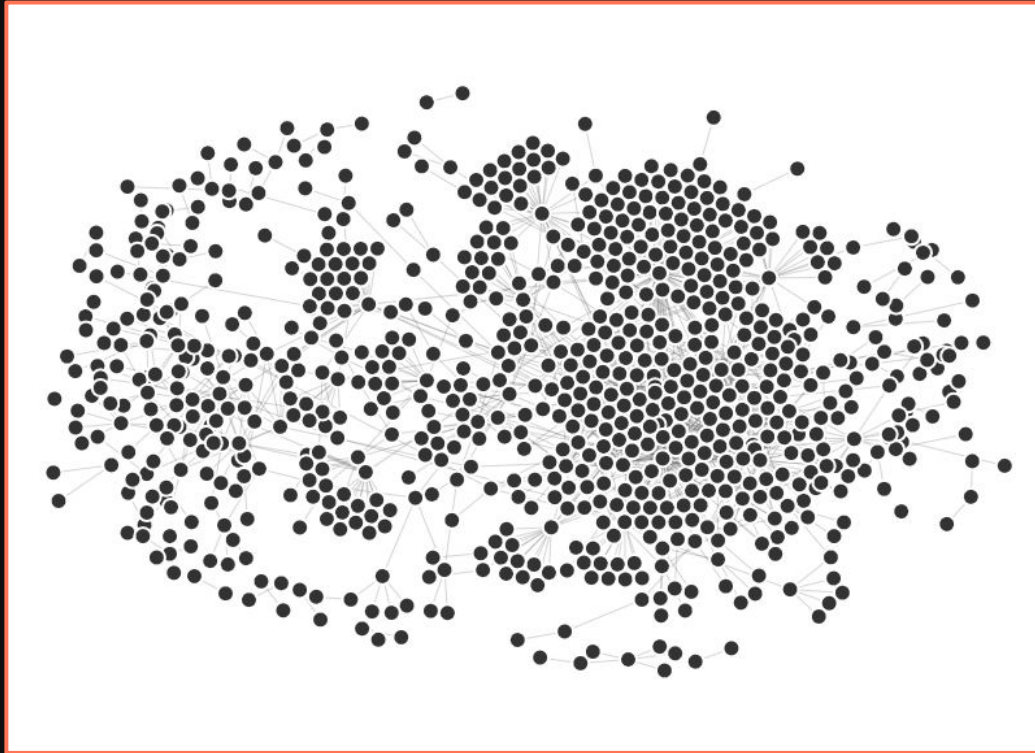
Visualizations
& ~~their~~ my
discontents



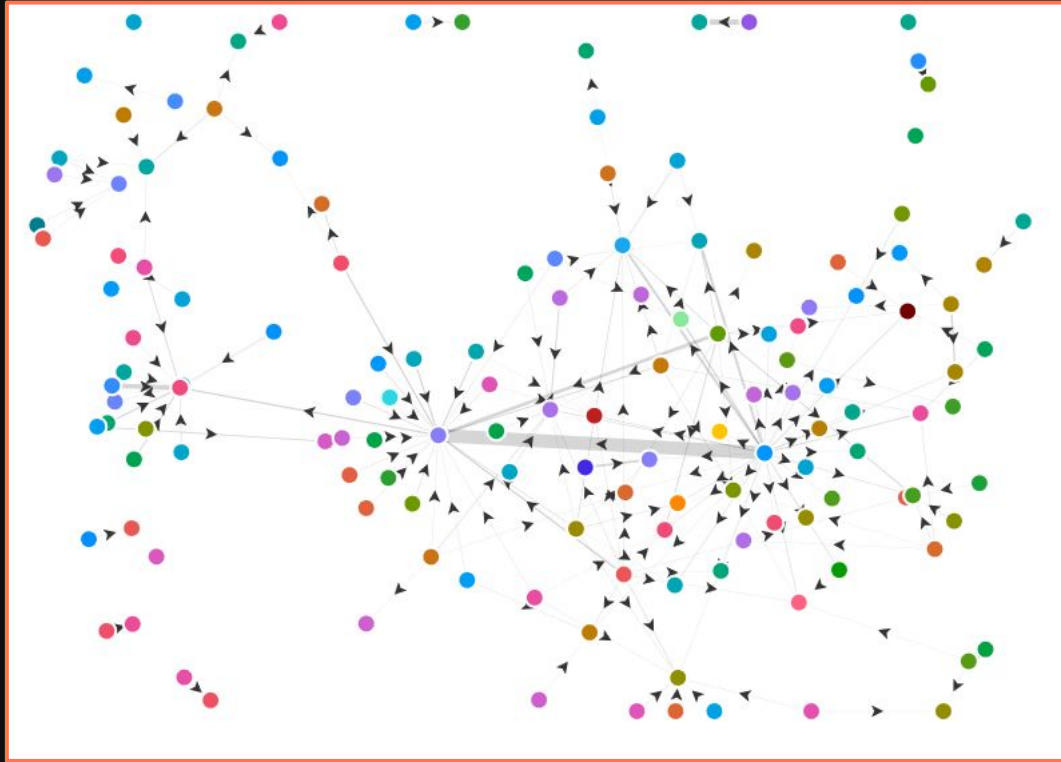
Circulation of Authors by Genre



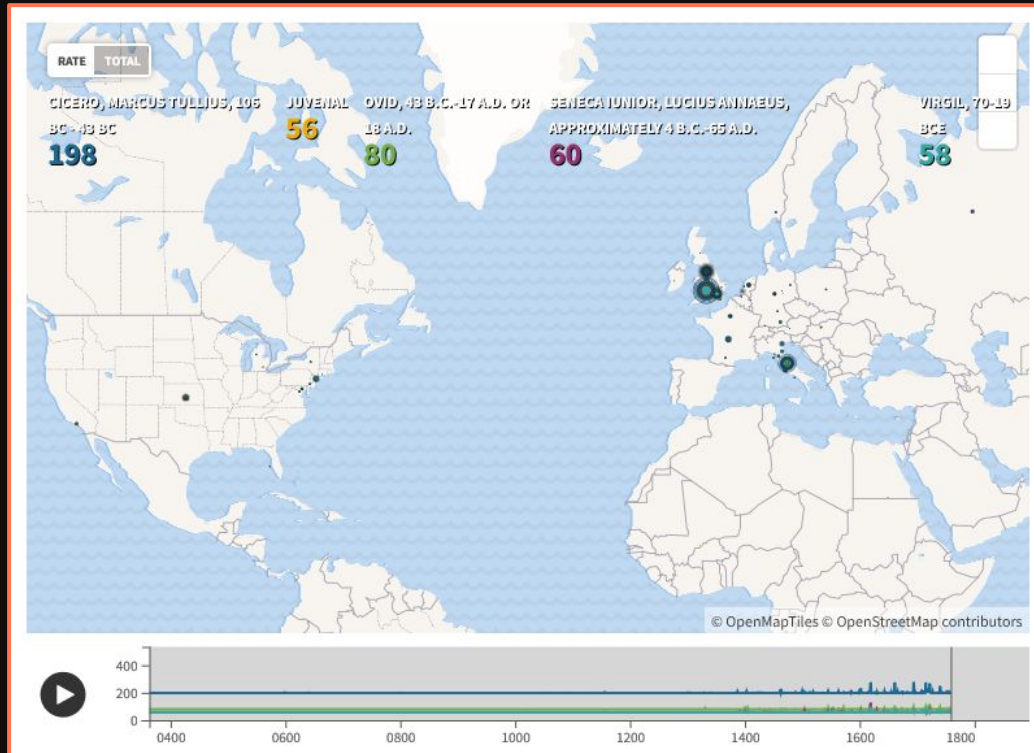
Buyer-Seller Network



Buyer-Seller Network



Agents and authors over time



Challenges & Drawbacks

- ◆ So much missing data!
- ◆ Complexity
- ◆ Unclear references
- ◆ Ambiguities
- ◆ Did I mention missing data?





What does this
mean for LOD
research?



((Is this anything??))



Dataset Quality



Accuracy & Consistency

Is the data **reliable**?

Completeness & Relevance

Is it the **right** data for this question?



✦ Dataset Quality



Accuracy & Consistency

Is the data **reliable**?

Completeness & Relevance

Is it the **right** data for this question?

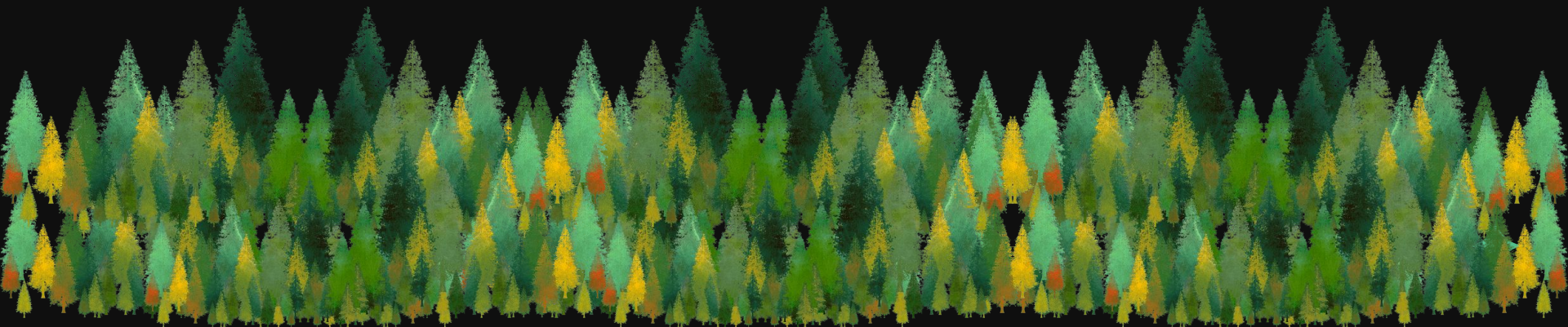


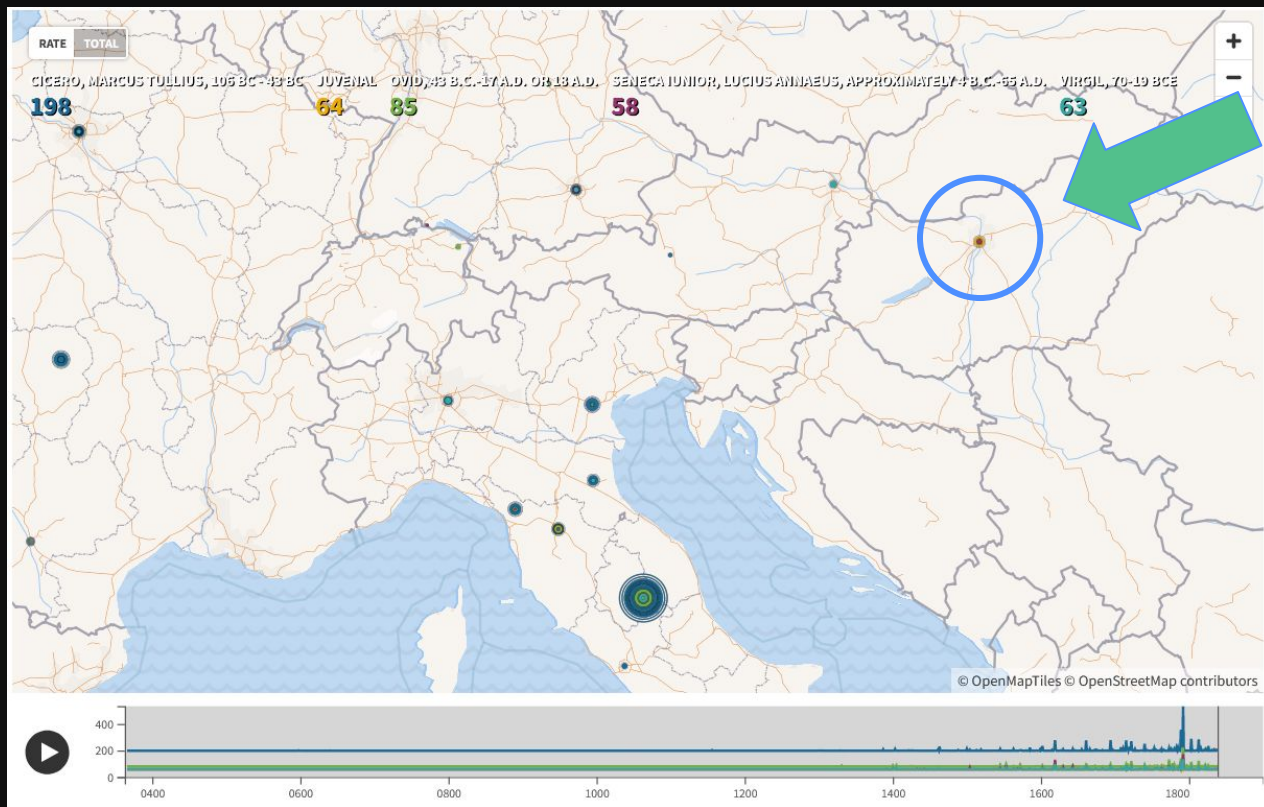


The SDBM as a Dataset

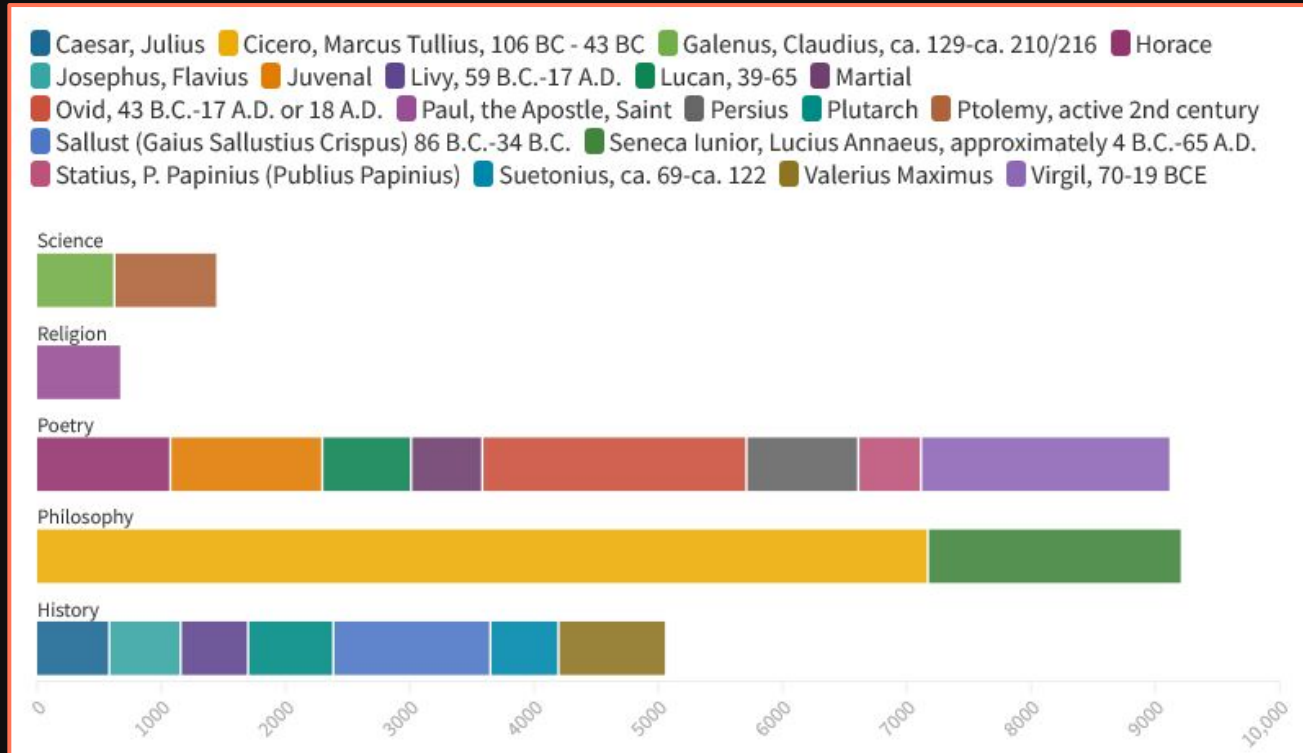


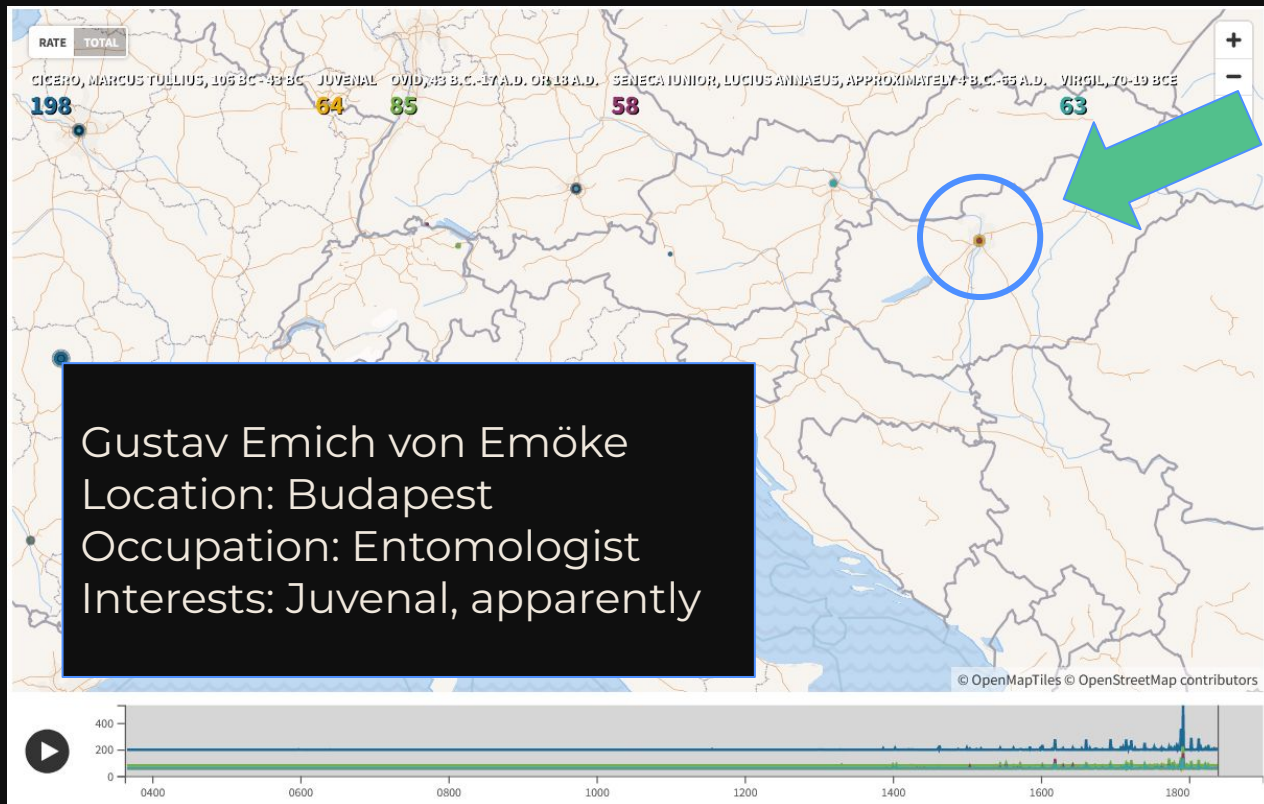
- ◆ So much missing data!
- ◆ Complexity
- ◆ Unclear references
- ◆ Ambiguities
- ◆ Did I mention missing data?








Circulation of Authors by Genre







Gustav Emich von Emöke
Location: Budapest
Occupation: Entomologist
Interests: Juvenal, apparently

Image c/o Entomologists of the World



The visualization is not the
answer, it's the **question**



✧ Data in the Humanities ✧



“Data”

That which is *given*

“objective”



“Capta”

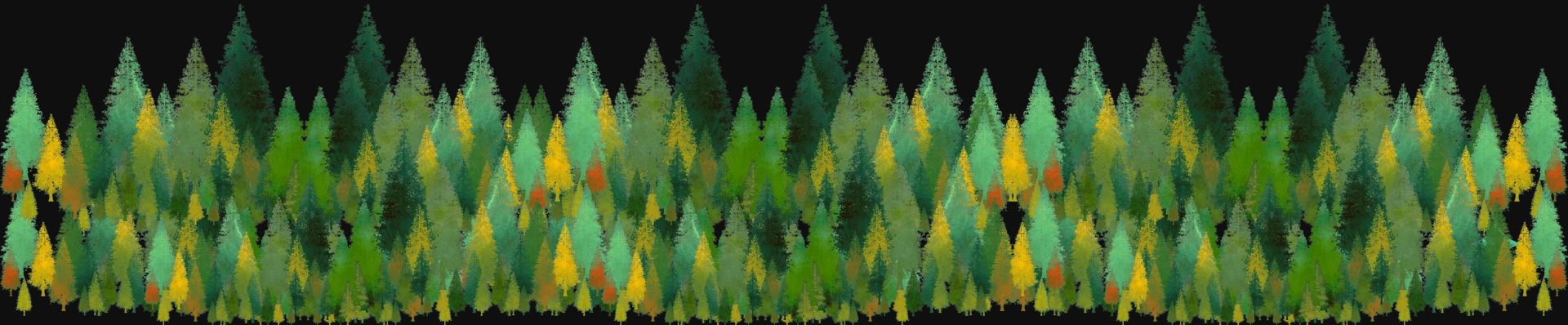
That which is *taken*

constructed

What does this mean??

LOD is trying to structure a complex, nuanced, messy universe

Humanists thrive on complexity, nuance, and ambiguity



Approaching LOD by Embracing Freedom



Integrations

Linked **OPEN** Data

Borrow data (& structures!)
from other sources

Verification

More data

Cleaner data

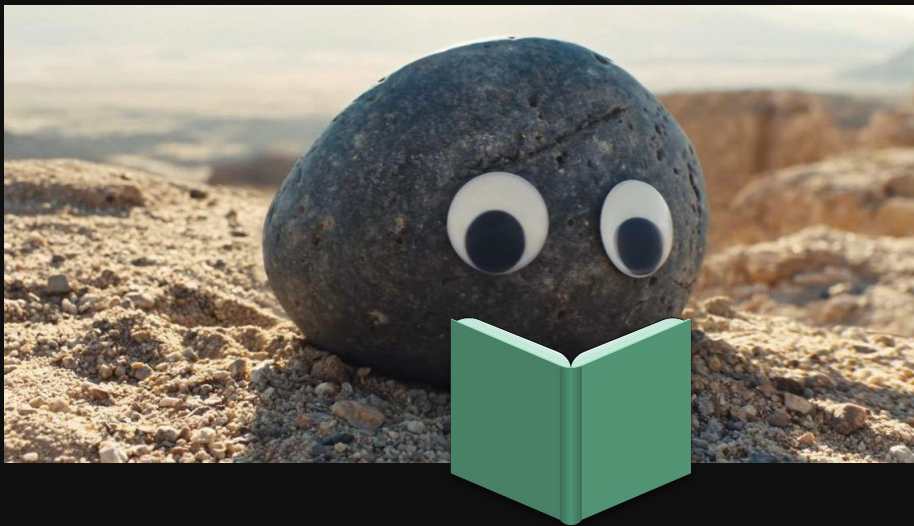
Better data



Close reading

Looking for **small** patterns
with **big** implications

Finding **stories** in the noise

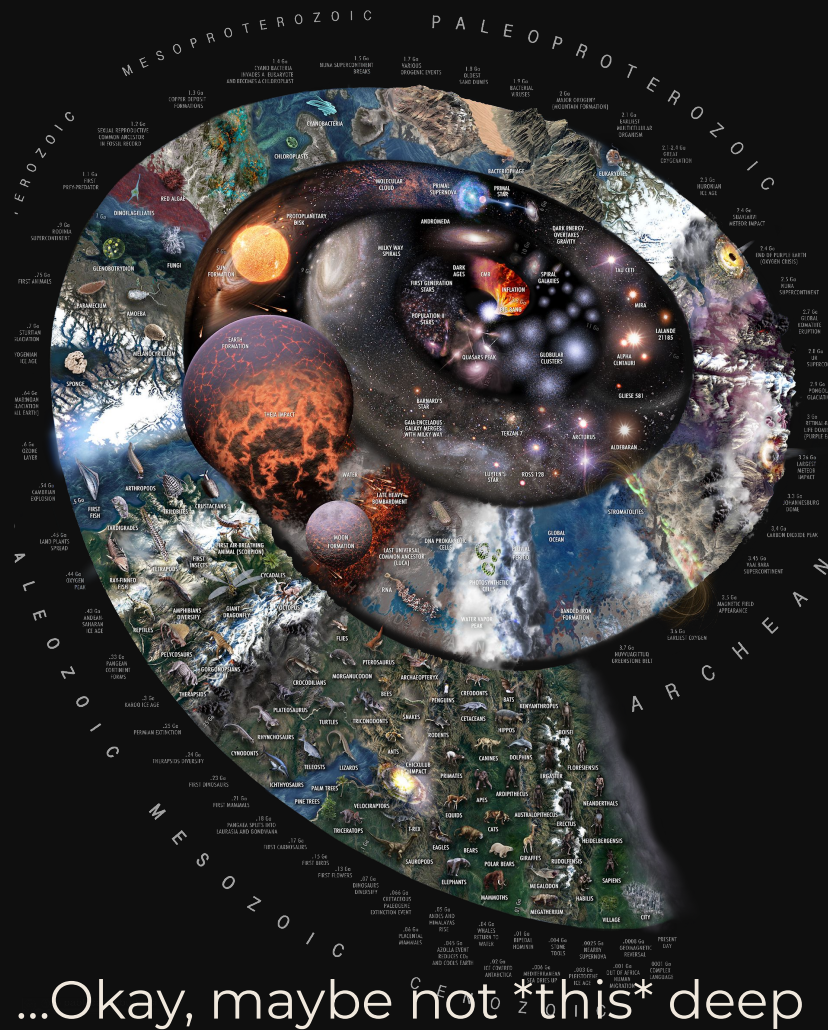


Deep Time

Long-term crowdsourcing and curation

Integration with other LOD systems (when relevant)

Slowly, the dataset improves



I have a
particular
set of
skills...



The visualization is not the
answer, it's the **question**

Thank you!

I have a
particular
set of
skills...



Kate Topham

Digital Humanities Archivist
Michigan State University
2022 LEADING Fellow
tophamka@msu.edu
[@tophkat](https://twitter.com/tophkat)

Data Visualizations

bit.ly/LeadingViz



LEADING & Schoenberg

LEADING: LIS Education
And Data Science
Integrated Network
Group

Schoenberg Database of
Manuscripts