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.
SDBM Entity Relationship Diagram
- OpenRefine!
- Cleaning
- Disambiguating
- VIAF IDs
  - Misattribution
  - Outdated ids
  - Duplicates
Visualizations & their my discontents
Circulation of Authors by Genre

- Caesar, Julius
- Cicero, Marcus Tullius, 106 BC - 43 BC
- Galenus, Claudius, ca. 129-ca. 210/216
- Horace
- Josephus, Flavius
- Juvenal
- Livy, 59 B.C.-17 A.D.
- Lucan, 39-65
- Martial
- Ovid, 43 B.C.-17 A.D. or 18 A.D.
- Paul, the Apostle, Saint
- Persius
- Plutarch
- Ptolemy, active 2nd century
- Sallust [Gaius Sallustius Crispus] 86 B.C.-34 B.C.
- Seneca Junior, Lucius Annaeus, approximately 4 B.C.-65 A.D.
- Statius, P. Papinius (Publius Papinius)
- Suetonius, ca. 69-ca. 122
- Valerius Maximus
- Virgil, 70-19 BCE

- Science
- Religion
- Poetry
- Philosophy
- History
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

- Caesar, Julius
- Cicero, Marcus Tullius, 106 BC - 43 BC
- Galenus, Claudius, ca. 129-ca. 210/216
- Horace
- Josephus, Flavius
- Juvenal
- Livy, 59 B.C.-17 A.D.
- Lucan, 39-65
- Martial
- Ovid, 43 B.C.-17 A.D. or 18 A.D.
- Paul, the Apostle, Saint
- Persius
- Plutarch
- Ptolemy, active 2nd century
- Sallust [Gaius Sallustius Crispus] 86 B.C.-34 B.C.
- Seneca Junior, Lucius Annaeus, approximately 4 B.C.-65 A.D.
- Statius, P. Papinius [Publius Papinius]
- Suetonius, ca. 69-ca. 122
- Valerius Maximus
- Virgil, 70-19 BCE
Gustav Emich von Emöke
Location: Budapest
Occupation: Entomologist
Interests: Juvenal, apparently
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

...Okay, maybe not *this* deep
The visualization is not the answer, it’s the question. I have a particular set of skills...
Thank you!

I have a particular set of skills...

Kate Topham
Digital Humanities Archivist
Michigan State University
2022 LEADING Fellow
tophamka@msu.edu
@tophkat

Data Visualizations
bit.ly/LeadingViz

LEADING & Schoenberg
LEADING: LIS Education And Data Science Integrated Network Group
Schoenber Database of Manuscripts