BUILDING A SCHOLARLY KNOWLEDGE BASE WITH WIKIDATA: A CASE STUDY FROM VANDERBILT UNIVERSITY
MOTIVATIONS

Why are we doing this?
DATA INTEGRATION
DATA FREEDOM
CROWDSOURCING
LIBRARIANSHIP
NUTS & BOLTS
Using OpenRefine

to add bibliographic data to Wikidata
### OpenRefine for Single Vanderbilt Divinity School Author

<table>
<thead>
<tr>
<th>Item Type</th>
<th>Publication Year</th>
<th>Author</th>
<th>Authors 2</th>
<th>Authors 3</th>
<th>Authors 4</th>
<th>Authors 5</th>
<th>Authors 6</th>
<th>Authors 7</th>
<th>Authors 8</th>
<th>Authors 9</th>
<th>DOI</th>
<th>10.1002/9780470391564</th>
<th>10.1002/9780470391571</th>
<th>10.1002/9780470391588</th>
<th>10.1002/9780470391595</th>
<th>10.1002/9780470391601</th>
<th>10.1002/9780470391618</th>
<th>10.1002/9780470391625</th>
<th>10.1002/9780470391632</th>
<th>10.1002/9780470391649</th>
</tr>
</thead>
</table>

The table above shows various records from the Vanderbilt Divinity School. Each row represents a different academic record, including the author(s), publication year, and DOI. The table is used to manage and refine the data in a structured format.
BENEFITS OF USING OPENREFINE

• Built for large-scale cleaning of bibliographic data
• Reconciliation with Wikidata
  • Helpful for Identifying Duplicates
• Schemas built in OR are exportable/importable
AND THE DRAWBACKS OF OPENREFINE

- Rigidity of Schemas
  - Documents or Authors
  - Difficulty with rich datasets
- Scalability
- Future use for small edits?
- Incorporating automation?
FUTURE DIRECTIONS

• Train Liaisons
  • Bibliographic citation collection
  • Input citations into CSV
  • Clean data in OpenRefine
  • Import pre-built schemas
  • Upload to Wikidata
• Investigate Methods for Automation
• Provide Training to Departments outside Library
INTERACTIONS MODEL

- Human
  - Manually imports data
  - Imports data from spreadsheet
- Spreadsheet
  - Manually imports data
  - Imports data from Quickstatements and OpenRefine
- Wikidata
  - Interacts with Python (API)
  - Sends SPARQL requests
- Wikibase
  - Interacts with Python (API)
  - Sends SPARQL requests
WIKIDATA ALIGNMENT ISSUES

• Researcher ≠ faculty (postdoc, staff)

• Some important properties missing
  • college/school
  • department

• Congruence? fieldOfWork, interestedIn
DISAMBIGUATION

• Matching researchers with IDs
• Matching authors with researchers

• Fuzzy string matching (Python fuzzywuzzy library)

• Eventual predominance of ORCID?
"Generating RDF from Tabular Data on the Web"

JSON-LD
CSV table schema

Python script:
• Local Wikibase mirror of data in Wikidata?
• How to map Wikibase properties to "real" Wikidata properties?
UPDATING

- Versioning?
- Sources? Duplication.
- Harvest/push frequency
- Acceptance of Wikidata data
- Vetting by researchers?
WikiData as a unifying resource for data on collectors of botanical and zoological specimens

Notes on STSM-Mobilise project.

Objectives

The main objectives of this STSM are:

- Find the total number of biological specimen collectors in Wikidata. Ideally, this would be done through SPARQL queries which can be replicated afterwards as Wikidata is in constant flux.
- Assess the comprehensiveness of data on these collectors, using both shape expressions and more in depth analysis in R for specific properties of interest.
- Describe regional and temporal patterns for these collectors, as well as gender balance and the extent of external links.
- Data validation by matching through external sources such as Bloodhound.
PHILOSOPHICAL QUESTIONS

• Does this project belong in Wikidata?
• Is Wikidata the place to build or just "identifier central"?
• Is it the community or the platform?
• How much do we care about the motivators vs. existing options?
NEXT STEPS

Where do we go from here?
COMMUNICATING
NETWORKING
QUESTIONS?

• Clifford Anderson
  Associate University Librarian
  @andersoncliffb

• Steve Baskauf
  Data Science and Data Curation Specialist
  @baskaufs

• Elisabeth Shook
  Librarian for Copyright and Scholarly Communications
  @Elisabrarian