NAMED ENTITY DISAMBIGUATION FOR ARCHIVAL COLLECTIONS: Metadata, Wikidata, & Linked Data

Presentation by
Katherine Polley

Paper also by
Vivian Tompkins, Brendan Honick, Dr. Jian Qin

Syracuse University School of Information Studies
Linked Archives Project

Modelling item-level metadata for three special collections as linked data
The Belfer Cylinders Collection: music and spoken word recordings dating from 1890 to 1929.


PERSON RECORDS

Person

Related to

Subject

Related to

Item

First name
Middle name
Last name
Role(s)
Metadata Enrichment

Using Wikidata to get additional information/properties about people related to the collections
**Wikidata**: free and open knowledge base for storing the structured/linked data related to other Wikimedia projects like Wikipedia

**Wikibase**: the database software behind Wikidata, available to create your own instances

**Why metadata enrichment**: initial metadata is limited, so adding additional properties would help researchers/users of the collections
**Named Entity Disambiguation:** also called entity linking, involves matching named entities in text to unique identities in a knowledge base.

**OpenTapioca:** recognizes named entities in free text and matches them to Wikidata entities, with scores for how likely a match is correct.

**Why this linker:** easy to use, accessible to librarians and archivists without technical knowledge.
Experiment & Results

Testing out OpenTapioca entity linking for person names in archival collections
### CHOOSING DOMAIN

<table>
<thead>
<tr>
<th>Collection</th>
<th>Type of People</th>
<th>Matches</th>
<th>In Wikidata</th>
</tr>
</thead>
<tbody>
<tr>
<td>The Ronald G. Becker Collection of Charles Eisenmann Photographs</td>
<td>Specific type of people</td>
<td>✔️</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Less than 200 names</td>
<td>❌</td>
<td></td>
</tr>
<tr>
<td></td>
<td>People not well known</td>
<td>❌</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Easy to confirm matches</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Faster but less significant</td>
<td>❌</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Less likely to be in Wikidata</td>
<td>❌</td>
<td></td>
</tr>
<tr>
<td>The Ted Koppel Collection</td>
<td>Wide variety of people</td>
<td>❌</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Over 32,000 names</td>
<td>❌</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Recognizable people</td>
<td>✔️</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Difficult to confirm matches</td>
<td>❌</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Too big for manual process</td>
<td>❌</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Likely to be in Wikidata</td>
<td></td>
<td></td>
</tr>
<tr>
<td>The Belfer Cylinders Collection</td>
<td>Specific type of people</td>
<td>✔️</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Around 700 names</td>
<td>✔️</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Many recognizable people</td>
<td>✔️</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Easy to confirm matches</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Reasonable scope for paper</td>
<td>✔️</td>
<td></td>
</tr>
<tr>
<td></td>
<td>More likely to be in Wikidata</td>
<td>✔️</td>
<td></td>
</tr>
</tbody>
</table>
BELFER RESULTS

Manual matching
- Match Confirmed: 79%
- No Match: 13%
- Unknown: 8%

OpenTapioca matching
- Correct: 85%
- Incorrect: 14%
- Not Recognized: 1%

Syracuse University 2021
Confidence Score Distribution – Correct Matches

Confidence Score Distribution – Incorrect Matches
Conclusions & Insight

What can we learn from this experimentation?
Choosing a knowledge base
- Understand your collection and the scope of the KB

Finding useful metadata elements for entity linking
- What information makes a person recognizable? Time period? Occupation? Description of related items?

Limits of existing algorithms
- Entity linking typically done with free text not metadata, more geared towards popular or contemporary figures

Need for future work/research
- Bridge gap between libraries/archives and computer scientists
THANK YOU!

Any Questions?