



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

COLLECTIONS

The Belfer Cylinders Collection:
music and spoken word recordings
dating from 1890 to 1929.



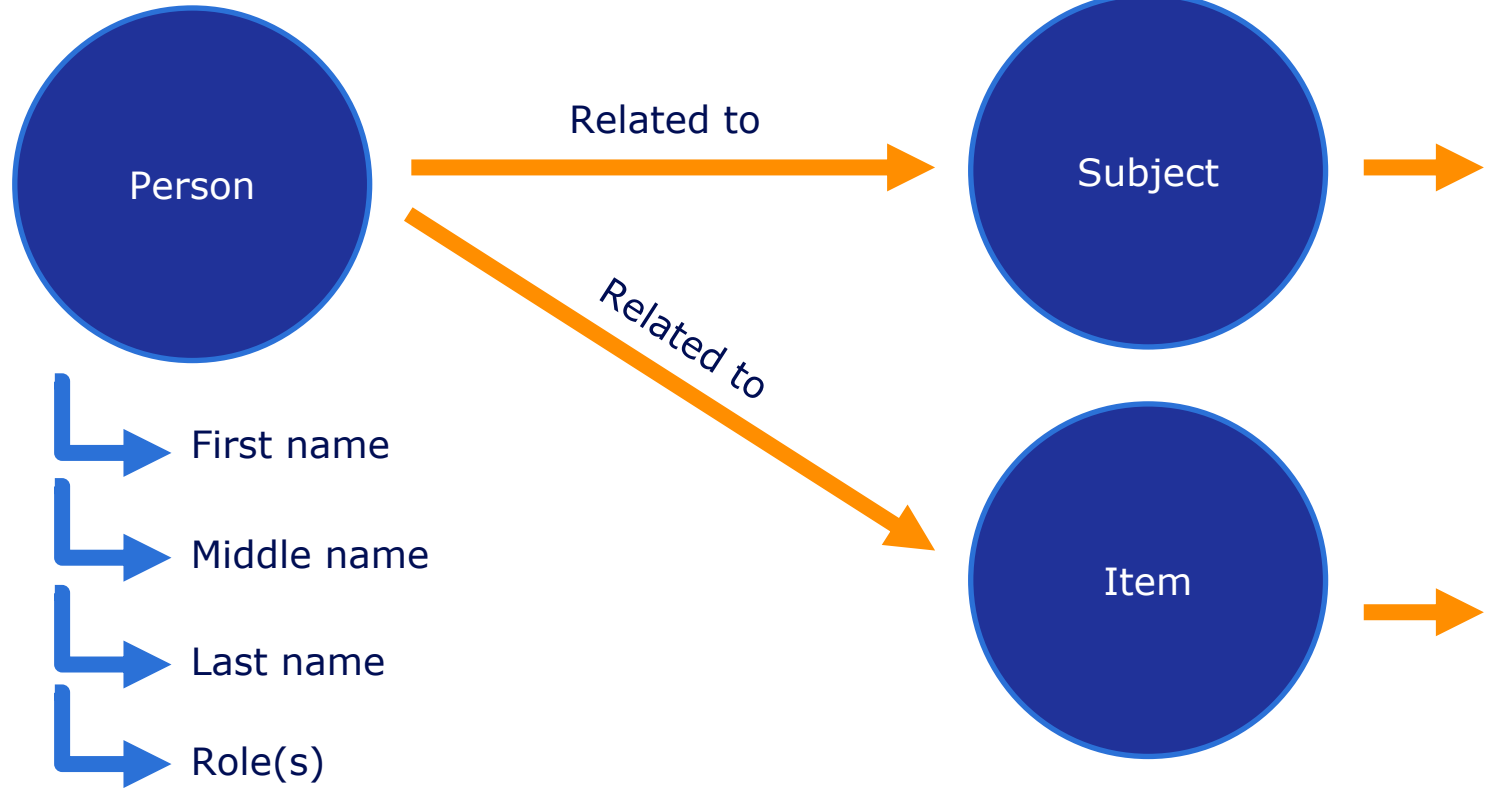
**The Ronald G. Becker
Collection of Charles
Eisenmann Photographs:**
photographs of 19th century sideshows,
circuses, and performers.



The Ted Koppel Collection:
videos of ABC News television
programming with Ted Koppel.



PERSON RECORDS



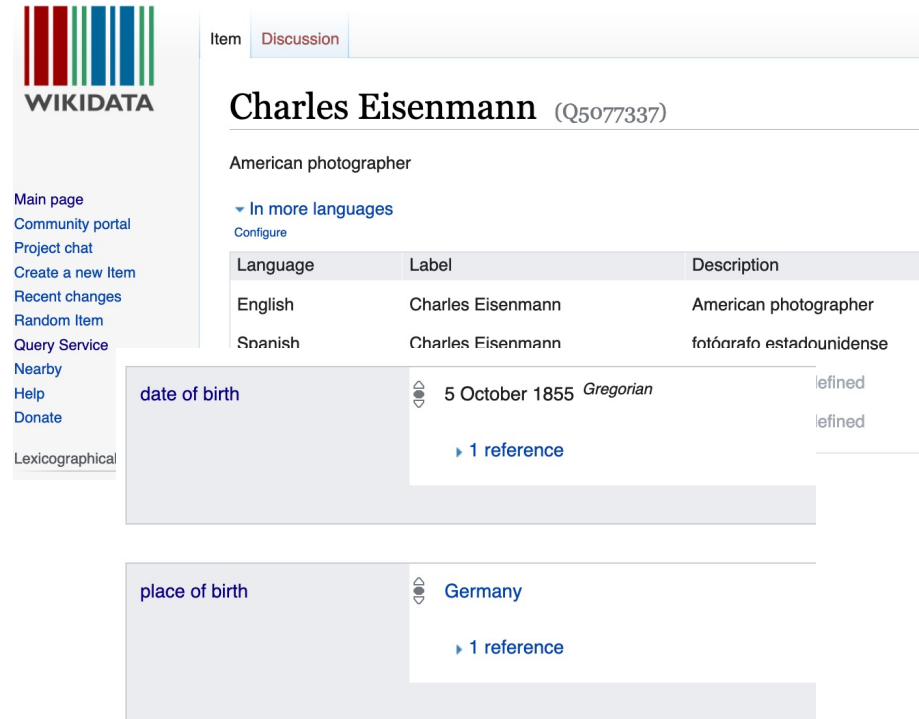
Metadata Enrichment

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



Item [Discussion](#)

Charles Eisenmann (Q5077337)

American photographer

[In more languages](#)
[Configure](#)

Language	Label	Description
English	Charles Eisenmann	American photographer
Spanish	Charles Eisenmann	fotógrafo estadounidense

date of birth 🗳️ 5 October 1855 *Gregorian* 🗳️ [1 reference](#)

place of birth 🗳️ Germany 🗳️ [1 reference](#)

ENTITY LINKING

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

[[OpenTapioca^β]]

OpenTapioca annotates text with locations, organizations and people from Wikidata.

It is synchronous with Wikidata.

This is a prototype. See the [GitHub project](#) for more info.

Associated Press writer Julie Pace contributed from Washington.

Annotate


[[Associated Press]] writer [[Julie Pace]] contributed from [[Washington]].

Experiment & Results

Testing out OpenTapioca entity linking for person names
in archival collections

CHOOSING DOMAIN




The Ronald G. Becker Collection of Charles Eisenmann Photographs

- Specific type of people  Easy to confirm matches
- Less than 200 names  Faster but less significant
- People not well known  Less likely to be in Wikidata

The Ted Koppel Collection

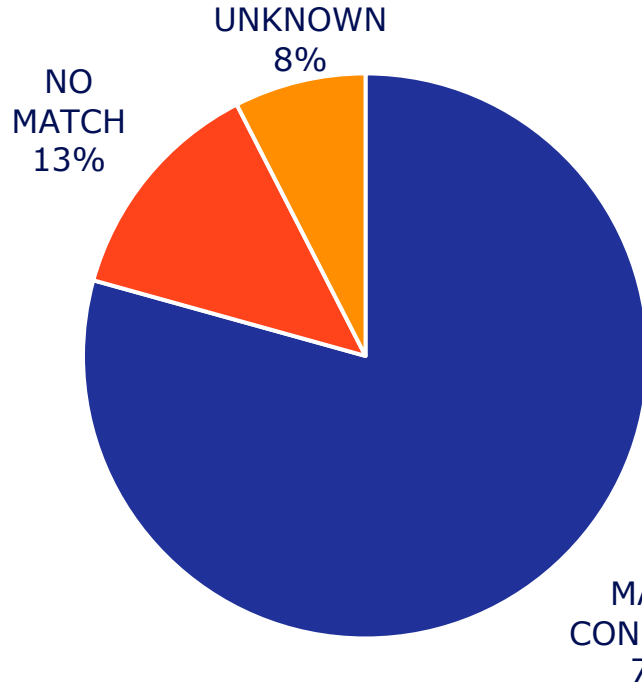
- Wide variety of people  Difficult to confirm matches
- Over 32,000 names  Too big for manual process
- Recognizable people  Likely to be in Wikidata

The Belfer Cylinders Collection

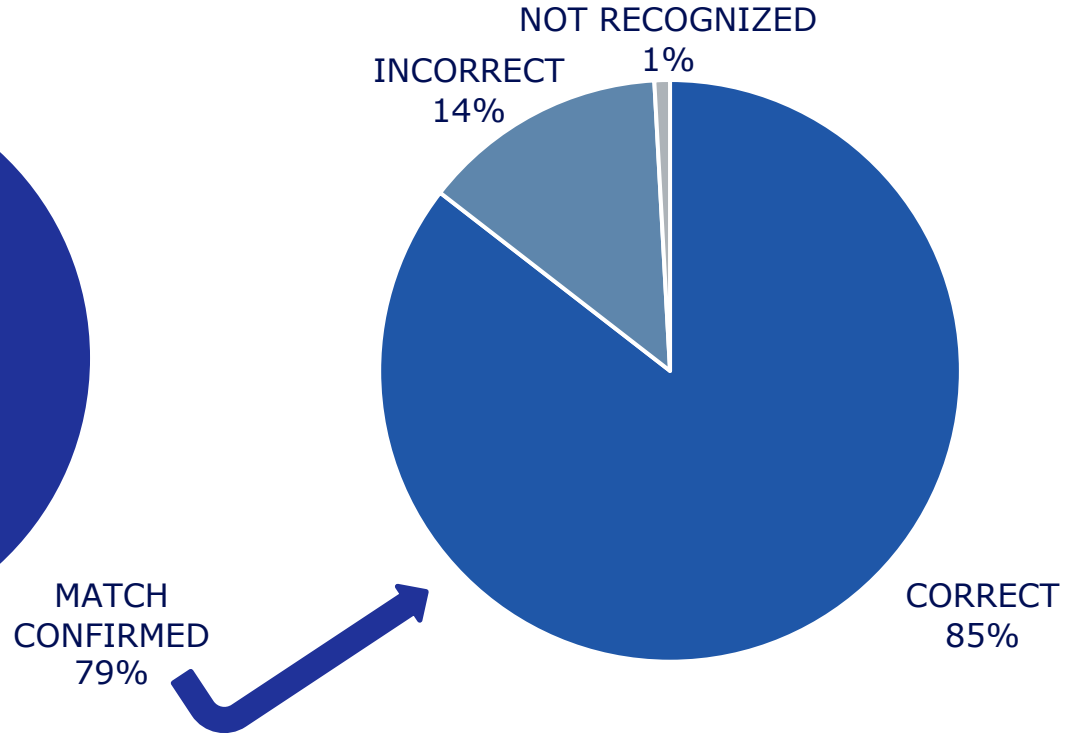
- Specific type of people  Easy to confirm matches
- Around 700 names  Reasonable scope for paper
- Many recognizable people  More likely to be in Wikidata

BELFER RESULTS

Manual matching

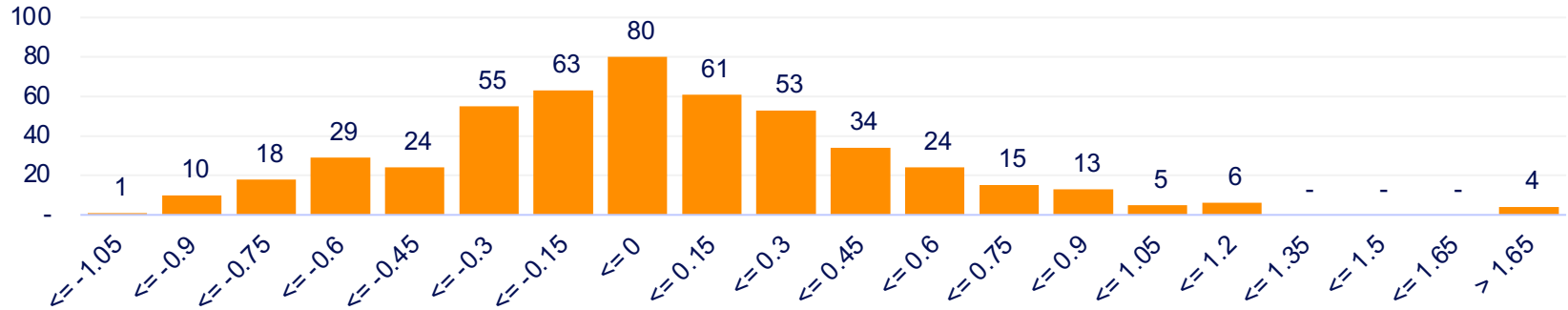


OpenTapioca matching

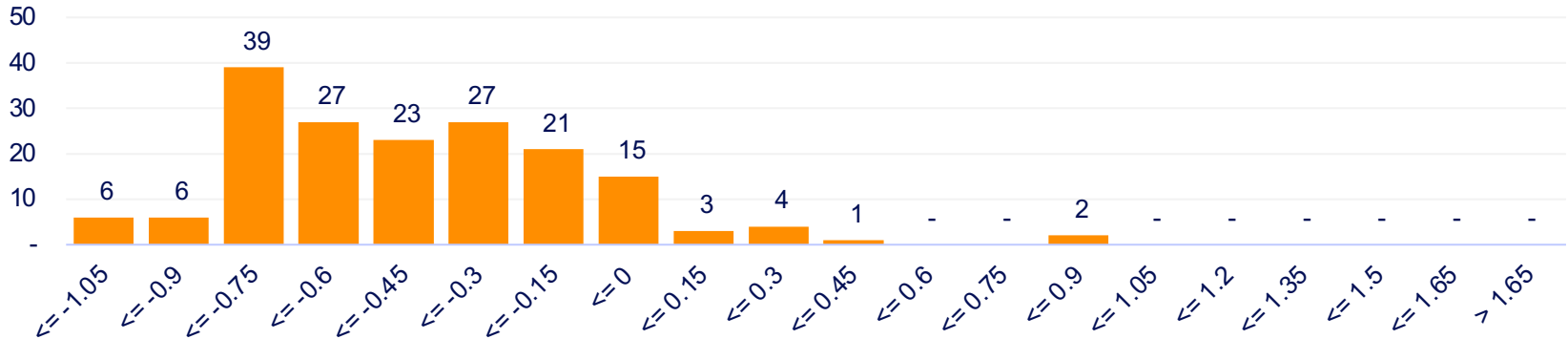


BELFER RESULTS

Confidence Score Distribution – Correct Matches



Confidence Score Distribution – Incorrect Matches

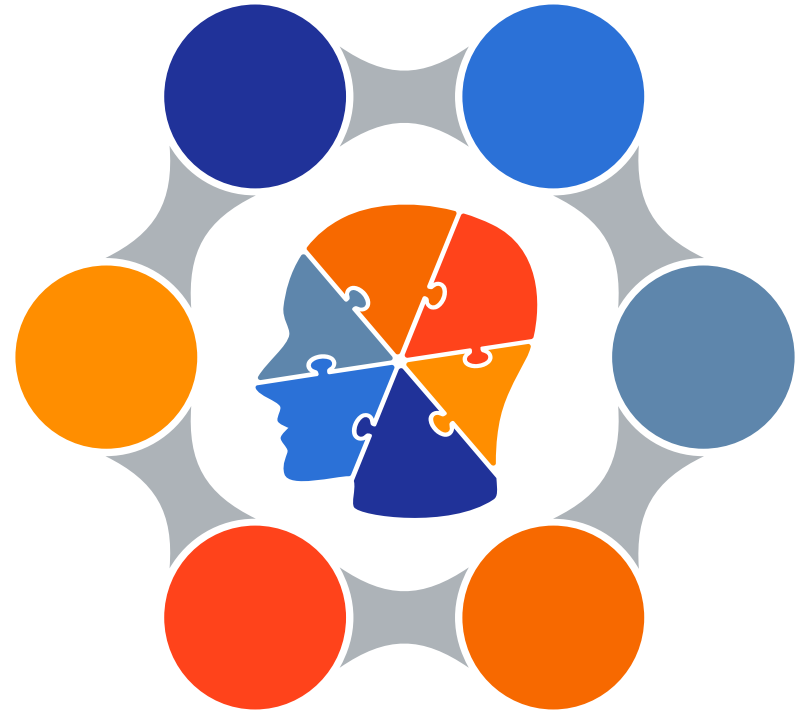


Conclusions & Insight

What can we learn from this experimentation?

CONSIDERATIONS

- **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?