WIKIDATA MISMATCH CREATION

Ethan Dawes, Henry Lee, Maggie Gao, Nanda Binod, Meredith Steever

WIKIMEDIA DEUTSCHLAND · WIKIDATA · WIKIPEDIA The Free Encyclopedia · PURDUE UNIVERSITY

OVERVIEW

Wikidata is a public domain database of people, places, and facts that anyone can contribute to like Wikipedia.

Discrepancies between Wikidata and trusted external sources have been identified that could propagate to downstream projects, like Google knowledge boxes and Apple's Siri.

To tackle this, the Mismatch Finder was established as a site where downstream projects and the broader Wikimedia community can report disparities between Wikidata and other trusted data sources. Wikidata contributors then manually reconcile the differences.

The project's goal is to deliver new useful mismatches for the Mismatch Finder.
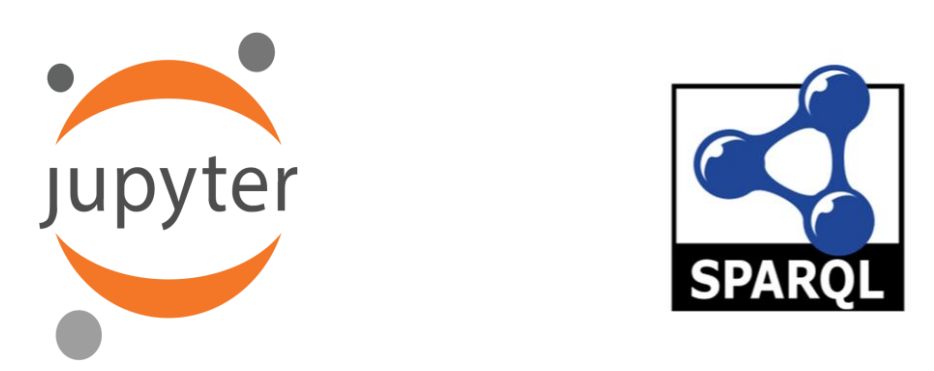
CENTRAL QUESTIONS

1. What strategies can be employed to reconcile discrepancies between Wikidata and external trusted sources and how can these strategies mitigate potential errors propagated to downstream projects?

2. How can the process of identifying mismatches be improved and made easier?

OUTCOMES

- Better understand Wikidata mismatch generation process from the user perspective to improve the platform

- Document this process to be shared with the Wikidata community as examples

- Mismatches added to the Mismatch Store, thus improving the quality of Wikidata's data for all downstream users
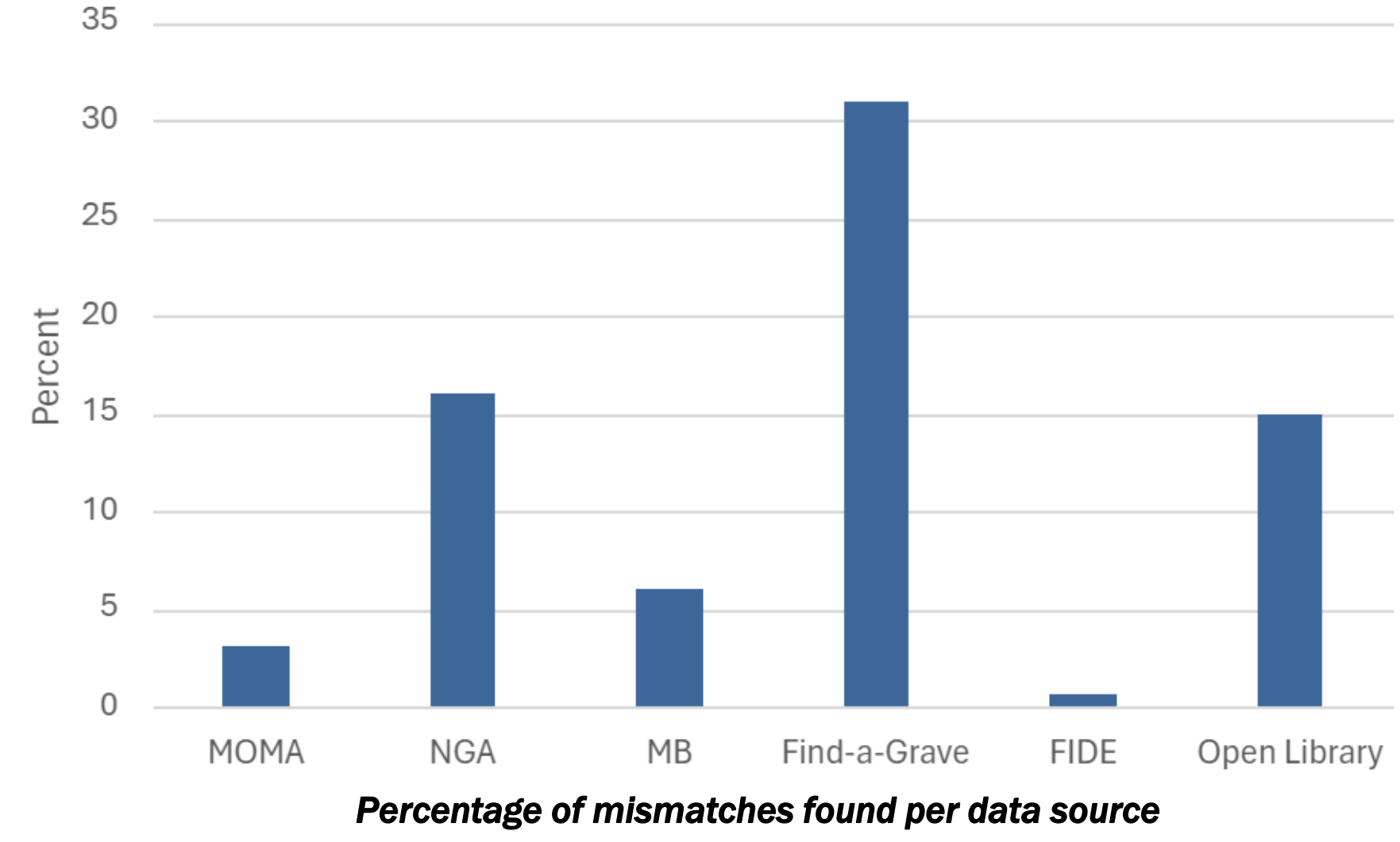
TOOLS USED

Tools used:

- Jupyter Notebook

- SPARQL: Language for querying values with relations
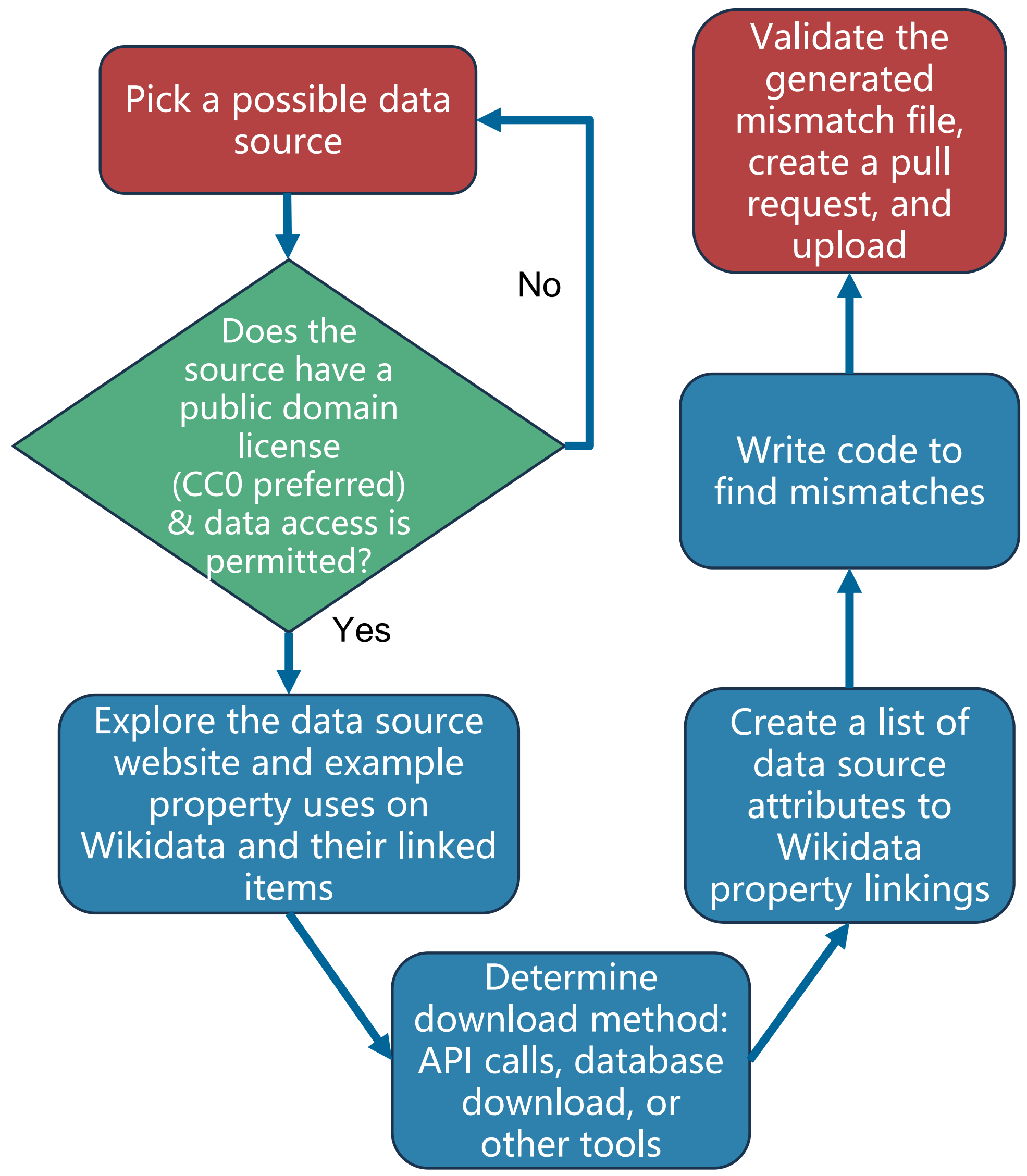
- QLever: Fast SPARQL query service

DATA EXPLORATION

Sources Chosen:

- FIDE: Chess Player Database

- MOMA: Museum Art Database

- National Gallery of Art: Museum Art Database

- Open Library: Book Database

- Mix'N'Match: Community Website and Database for Finding Mismatches

- MusicBrainz: Open Music Database

- Find A Grave: Database of Death/Birth Date/Place

MOMA MUSEUM OF MODERN ART · Find a Grave · OPEN LIBRARY · FIDE · NATIONAL GALLERY OF ART U.S.A.

Percentage of mismatches found per data source

METHODOLOGY

Pick a possible data source → Does the source have a public domain license (CC0 preferred) & data access is permitted? → No (back to Pick a possible data source) / Yes → Explore the data source website and example property uses on Wikidata and their linked items → Determine download method: API calls, database download, or other tools → Create a list of data source attributes to Wikidata property linkings → Write code to find mismatches → Validate the generated mismatch file, create a pull request, and upload

FUTURE PLANS

- Bots: We did not utilize bots. Bots can help generate mismatches and even increase the number of mismatches that are being found.

- Project Expansion: Alternative ways could be used to find mismatches between Wikidata and other external data sources.

REFERENCES

Wikidata Introduction
Wikidata Data Access Guide
Wikidata Query Service
Database Download Docs
Mismatch Finder Docs

ACKNOWLEDGEMENTS