Workshop: Reconciling to Wikidata with OpenRefine
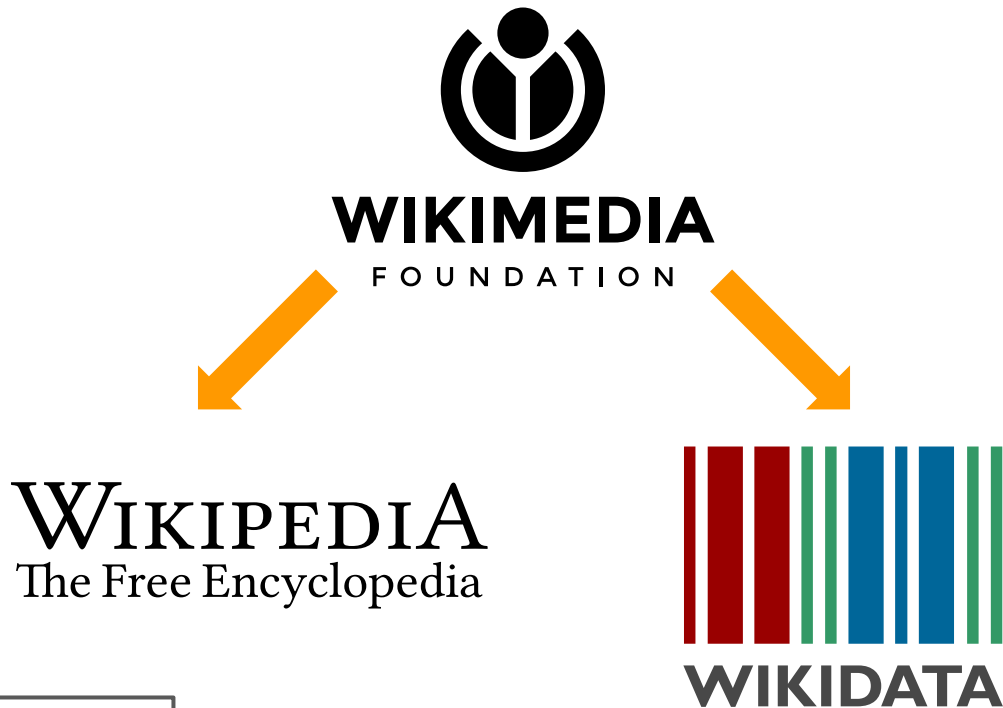
Wikidata

- Free
- Collaborative
- Structured data
- Items can describe people, places, organizations, things, publications, ideas, and more

Barack Obama (Q76)

president of the United States from 2009 to 2017

free will (Q9476)

ability of agents to make choices free from certain kinds of external constraints

Wikidata interfaces

Manual editing interface:
human-readable

OpenRefine interface:
machine-readable/transformable

Wikidata structure

item	property	value
ID begins with 'Q'	*ID begins with 'P'*	*ID begins with 'Q' -OR- is date, number, URL, or other non-item data type*
Marshall McLuhan (Q193871)	employer (P108)	University of Toronto (Q180865)
Marshall McLuhan (Q193871)	date of birth (P569)	21 July 1911 (no item ID)
Marshall McLuhan (Q193871)	Library of Congress authority ID (P244)	n80126329 (no item ID, links to: https://id.loc.gov/authorities/n80126329)

1 row = 1 statement on Wikidata

OpenRefine

- Designed for structured data
- Open-source
- Train station, NOT a database
- 3 major functions:
  1. Exploring data
  2. Transforming data
  3. Reconciling, or matching, data

- Basic building blocks: **G**eneral **R**efine **E**xpression **L**anguage
  - Rearranging strings
  - Concatenating strings
  - Extracting parts of strings
  - Math functions

OpenRefine extensions

- Wikidata extension:
  - Comes pre-installed
  - Connects your instance of OpenRefine to Wikidata
  - Major functions:
    - Reconcile your data to Wikidata and get back additional data based on matches
    - Submit new and edited Wikidata statements

- Other extensions & reconciliation services:
  - Social Networks and Archival Context extension
  - Virtual International Authority File
  - Advanced tools for reconciliation and extension

Reconciliation

- Exists outside of OpenRefine or Wikidata
- Match entities between two datasets
- Allows you to get unique IDs that refer to external databases
- Can "extend" data by pulling other columns or statements from external DBs
- Core function of OpenRefine

Your dataset

1. Item A
2. Item b
3. Item C
4. …

Database or other dataset

Item B (ID: 73)
Item A (ID: 84)
Item F (ID: 62)
…

Your dataset, reconciled

1. Item A (ID: 84)
2. Item B (ID: 73)
3. Item C (no match)
4. …