Wikidata as authority linking hub

Joachim Neubert (ZBW)

Jakob Voß (VZG)

Introduction

Authority files

Consistently refer to entities

- Via identifier ("things, not strings")
- GND, MeSH, STW, ISIL, RePEc-Authors…

Linking hubs

Connect identifiers among authority files

- `owl:sameAs`, `skos:exactMatch`, `skos:closeMatch`…
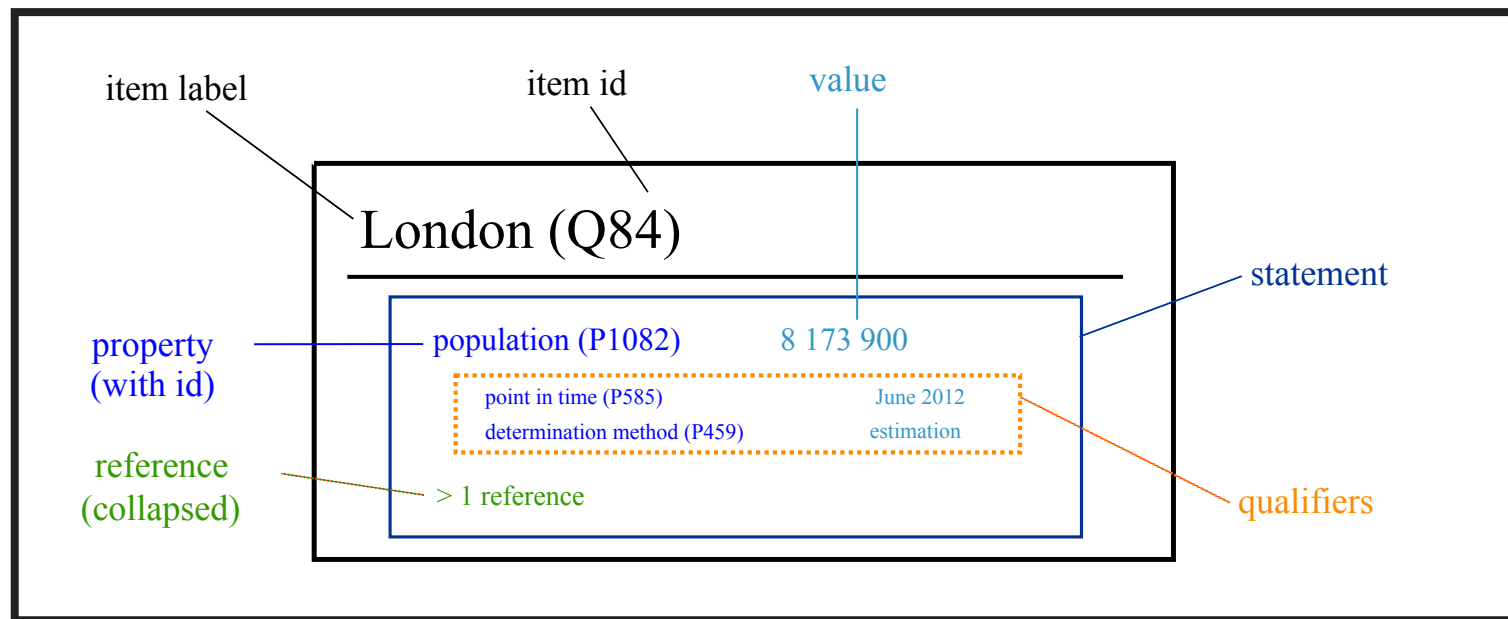- VIAF, sameAs.org, Wikidata…

Wikidata

- Knowledge base of Wikimedia projects
- All kinds of entities
  - concepts, places, people, works…

Wikidata Usage

- Editable by anyone
  - via Website and API
  - via apps that use the API
- Data available
  - http://query.wikidata.org/ (SPARQL)
  - JSON API & database dumps

Wikidata Statements

item label   item id   value

London (Q84)

statement

property (with id)   population (P1082)   8 173 900

point in time (P585)   June 2012
determination method (P459)   estimation

qualifiers

reference (collapsed)   > 1 reference

Wikidata item example

Bina Agarwal (Q4913801)

Indian feminist economist

instance of: Bina Agarwal is a(n) human

Statements	
employer	Harvard University (private research university in Cambridge, Massachusetts, United States)
	University of Michigan (public research university in Ann Arbor, Michigan, United States)
country of citizenship	India (a federal republic in South Asia)
occupation	economist (professional in the social science discipline of economics)
date of birth	1951
educated at	University of Cambridge (collegiate research university in Cambridge, England, United Kingdom)
	University of Delhi (Indian public central university located in Delhi)
sex or gender	female (human who is female (use with Property:P21 sex or gender). For groups of females use with "subclass of (P279)")
award received	Padma Shri (award of the state of India)
described by source	Encyclopedia of Global Justice (2011 ed.) (2011 edition of the reference work published by Springer)

Media	
image	Bina Agarwal at the World Economic Forum on India 2012.jpg
Commons category	Bina Agarwal

Wikimedia Categories and Portals	
Commons category	Bina Agarwal

Links

Wikidata page

Wikipedia article

Reasonator

Identifiers	
ISNI	0000 0...8 7046
SUDOC authorities	033750807
LCAuth ID	n83133496
BnF ID	124586640
National Thesaurus for	071124128

Author Names ID		
GND ID	113900171 ☑	>
VIAF ID	69026738 ☑	>
ORCID iD	0000-0...6-6877 ☑	>

Authority file identifiers in Wikipedia

More than half of all Wikidata properties

- Datatype external identifier (~1,750)
- Properties for authority control (~1,500)
- Properties with corresponding KOS (~220)

Wikidata—ISIL (organizations)

Example:

Neuschwanstein Castle (Q4152)
ISIL (P791): *DE-MUS-051612*

Current state:

- lobid.org: ~30,000 ISIL (DACH only)
- Wikidata: ~6,500 ISIL

Tool: *Mix'n'match*

- Web application mapping tool
- Helps to add 1-to-1-mappings

https://tools.wmflabs.org/mix-n-match/

Step 1: Upload ISIL list with names

Here, you can import a new catalog into Mix'n'match.

Just paste your tab-delimited text (copy from Excel should be fine) into the textbox, and fill out the form!

This import function is experimental! Please be careful, and contact Magnus if something goes wrong!

Catalog name	lobid-museums	A short name for your catalog (like "VIAF")
Catalog description	Museums with an ISIL as	A brief description of your catalog
Catalog URL	https://lobid.org/organisa	The URL (main page) of your catalog
Wikidata property	P791	The Wikidata property for this catalog; *optional*
Main language	de	Which Wikipedia to search for this catalog (language code)
URL pattern	http://lobid.org/organisati	A URL pattern to get to entries from their $id (e.g "http://dummy.org/$id"); *optional*
Default type	Q7075	The default type of the entries (e.g "person"); *optional* If you use a Q-ID (e.g. Q11424) here, automatches will be limited to items linking to this item and its subclasses, resulting in fewer matches of higher quality

The following (tab-delimited,un-quoted) columns are required, *in this order*:

1. **Entry ID** (your alphanumeric identifier; must be unique within the catalog)
2. **Entry name** (will also be used for the search in mix'n'match later)
3. **Entry description**

Step 2: Confirm match candidates

lobid-museums

☑ Load next entry on empty search results

Museums with an ISIL as derived from http://lobid.org/organisations/search?
q=type:Museum&size=10000&format=csv%3Aisil%2Cname%2Clocation%5B0%5D.address.addressLocality

Haus der Wannsee-Konferenz

Berlin

---

lobid
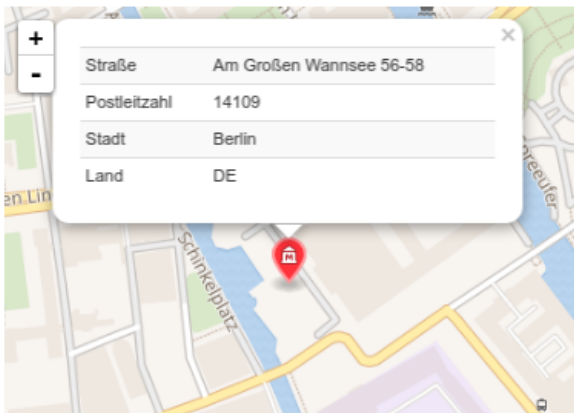
| Suchen in lobid-organisations | 🔍 |

Haus der Wannsee-Konferenz

{<}

ISIL	DE-MUS-429218
Typ	Museum
Klassifikation	Museum

Straße	Am Großen Wannsee 56-58
Postleitzahl	14109
Stadt	Berlin
Land	DE

Haus der Wannsee-Konferenz

Find

Q28144141 [Q28144141]
Building in Berlin

≡ Wikidata 🔍

No label defined (Q28144141)

文A ☆

No description defined ✏

Statements

instance of

🎙 building

0 references

image

🎙 Haus der Wannsee-Konferenz 02-2014.jpg

0 references

located in the administrative territorial entity

🎙 Berlin

0 references

GND—RePEc Authors

- In EconBiz economics search portal authors are identified differently:
  - by **GND ID** in data from ZBW's Econis catalog (and from others)
  - by **RePEc Author ID** in data from *Research Papers for Economics*
- Large volumes: 450,000 vs. 50,000 distinct persons
- ~3,000 pairs of IDs discovered in a previous project

Utilizing Wikidata as Linking Hub

- Wikidata-Properties for both identifier systems

  - GND ID (P227): ~375,000 items which are humans
  - RePEc Short-ID (P2428): ~2,200 items

- Since every identifier should identify exactly one person, we can derive

  - GND ID ⟶ Wikidata ID ⟶ RePEc ID
  - RePEc ID ⟶ Wikidata ID ⟶ GND ID

  where both properties have values (~760 items)

Step 1: Supplement WD items with RePEc Short-IDs

- 77 WD items with GND ID without RePEc Short-ID

- Transform to quickstatements input file (SPARQL query, script)

- Copy & paste to *QuickStatements2*

Bulk editing with Quickstatements2
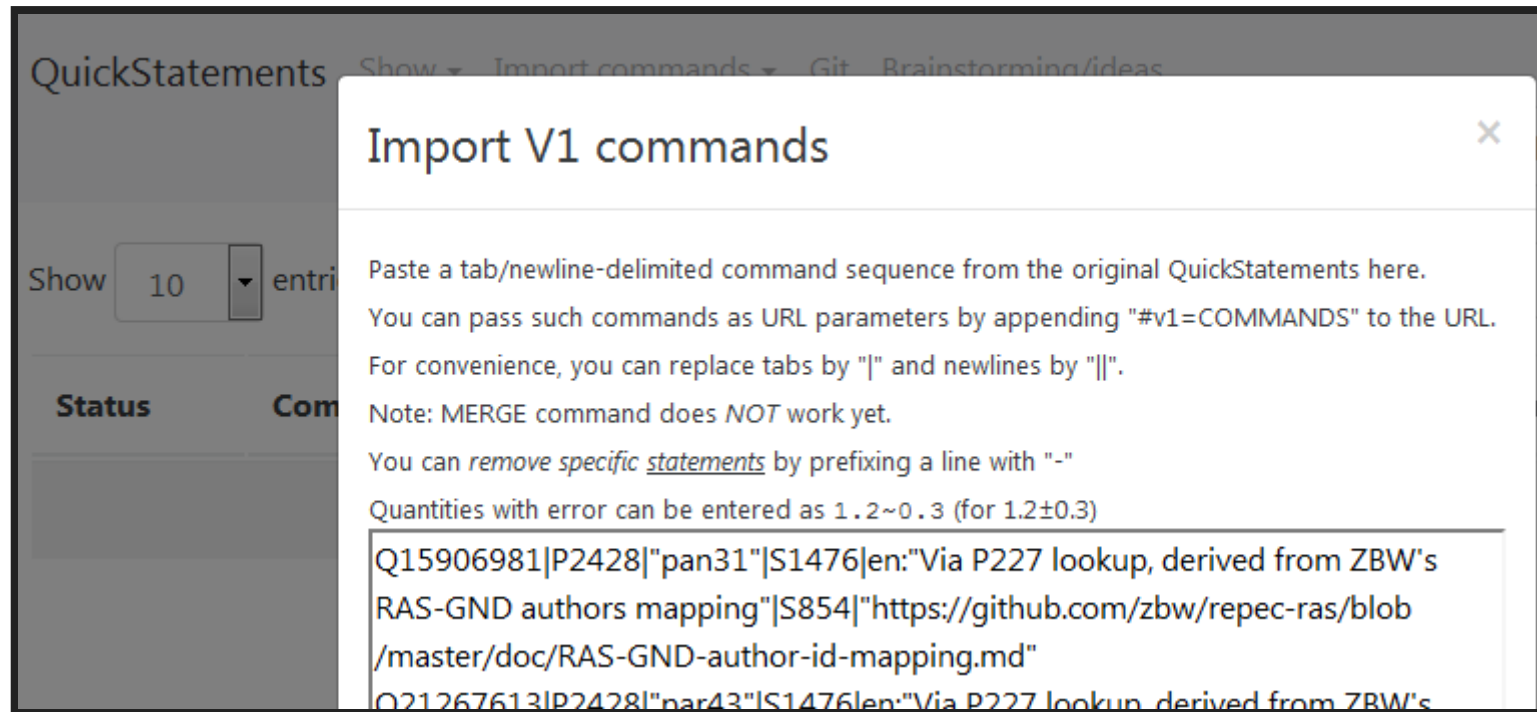
Further simplification with upcoming release of *wdmapper* command line tool

Step 2: Supplement WD items with GND IDs

- 384 WD items with RePEc Short-ID without GND ID
- *same process as other direction*

Step 3: Add "most important" authors with RePEc identifiers

- Scraped from ranking pages (Top 10% economists, Top 10% female economists)
- Transform and load into *Mix'n'match*
- *same process as ISIL use case*
    - Confirm match candidates (1,600 of 4,600)

Step 4: Add "most important" authors with GND identifiers

- 18,000 authors with >30 publications in EconBiz
- loaded as *Mix'n'match* set *GND economists (de)*
- order by publication count (descending)
- 25% matched automatically with Wikidata items

⇒ *Work to do*

Step 5: Rinse and repeat

- Repeat *Mix'n'match* "sync" operation before starting to work manually
  - often, people are adding data at fast rate!
- Repeat bulk adding of missing identifiers to make use of identifiers added meanwhile

Step 6: Add missing Wikidata items

- Verify missing authors indeed are not in Wikidata
- Generate Wikidata items from from existing mappings or lists, e.g. top female economists

QuickStatements

This tool can add statements (with optional qualifiers and sources) to Wikidata items.

First column are articles from   [e.g. enwiki]   **Do it**

```
CREATE
LAST    Len "Ana Rute Cardoso"
LAST    Den "Economist (Barcelona Graduate School of Economics (Barcelona GSE) -> Institute of Economic Analysis)"
LAST    P2428   "pca97"
LAST    P31 Q5
LAST    P21 Q6581072
LAST    P106    Q188094
CREATE
LAST    Len "Maria De Paola"
LAST    Den "Economist (Institute of Labor Economics (IZA); University of Calabria -> Department of Economics, Statistics
and Finance)"
```

You are logged into WiDaR as Jneubert.   Show/hide HOWTO

1. Processing Q29570517 (Q29570517 Len "Ana Rute Cardoso")
2. Processing Q29570517 (Q29570517 Den "Economist (Barcelona Graduate School of Economics (Barcelona GSE) ->
   Institute of Economic Analysis)")
3. Processing Q29570517 (Q29570517 P2428 "pca97")
4. Processing Q29570517 (Q29570517 P31 Q5)
5. Processing Q29570517 (Q29570517 P21 Q6581072)
6. Processing Q29570517 (Q29570517 P106 Q188094)
7. Processing Q29570519 (Q29570519 Len "Maria De Paola")

Result

The mapping, currently (2017-05-02) consisting of

- 1233 matching GND - RePEc short IDs
  - 769 matches from ZBW's mapping
  - 464 matches contributed by non-ZBW staff
- Finally all 3,000 pairs from ZBW's mapping

Further Results

- Identifiers and items added by individual Wikidata contributors add up continuously
- Mapping steps can be repeated with additional input data (e.g., top economists from Latin America, "all authors affiliated to Leibniz institutions in economics")…
- Further identifiers (VIAF, ORCID, …) provide more opportunities for indirect matching

Results from *every step in the mapping process* and all indiviual efforts immediately available and preserved

Tools

- *Mix'n'match* (intellectual matching)
- *QuickStatements2* (addition of generated properties and items)
- *wdmapper* (harvest, diff & add mappings)
  - Support of indirect mappings (e.g., GND-WD-RePEc) in one step
  - Work in progress (no adding by now)
  - Daily harvested mappings in multiple formats: http://coli-conc.gbv.de/concordances/wikidata/

Limitations

- Mapping algorithms to find mapping candidates
- Limitation to easy-1-1-relationships
    - part-whole
    - often new Wikidata items required
    - depends on the use case
- Large sets of mappings and results
- Regular review required for maintainance

Benefits

- Outsourced interface, storage, and operation
- Crowdsourced mapping maintenance
- Wikidata has policies and tools for data quality
- Open Data for multiple and unknown uses
- Additional benefits:
  - multilingual Wikipedia links
  - lots of (formatted) data
  - links to multiple other vocabularies
  - nice pictures …