Wikidata as a linking hub for knowledge organization systems?

Integrating an authority mapping into Wikidata and learning lessons for KOS mapping

Joachim Neubert
ZBW – Leibniz Information Centre for Economics, Kiel/Hamburg

NKOS Workshop @ TPDL, Thessaloniki, Greece
21.9.2017
The idea of linking hubs
Agenda

1. Suitability of Wikidata as linking hub

2. Experiences from moving an authority mapping to Wikidata

3. Outlook to knowledge organization systems
Wikidata basics

- Knowledge base for Wikimedia projects
- All kinds of entities: concepts, places, people, works …
- Editable by everyone
- Data available (under CC0)
  - JSON API & database dumps
Wikidata statements

London (Q84)

- item label: London
- item id: Q84
- value: 8,173,900
- property: population (P1082)
- reference (collapsed): > 1 reference
- determination method: estimation
- point in time: June 2012

statement

qualifiers
### Statements

<table>
<thead>
<tr>
<th>Property</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>occupation</td>
<td>economist (professional in the social science discipline of economics)</td>
</tr>
<tr>
<td>country of citizenship</td>
<td>India (federal republic in Asia)</td>
</tr>
<tr>
<td>date of birth</td>
<td>1951</td>
</tr>
</tbody>
</table>
| sex or gender             | female (human who is female (use with Property P21 sex or gender). For groups of females use with "subclass of (P279)"
| educated at               | University of Cambridge (collegiate public research university in Cambridge, England, United Kingdom) |
|                           | University of Delhi (Indian public-central university located in Delhi) |
| employer                  | Harvard University (private research university in Cambridge, Massachusetts, United States) |
|                           | University of Michigan (public research university in Ann Arbor, Michigan, United States) |
|                           | University of Manchester (public research university in Manchester, England) |
| award received            | Padma Shri in literature & education                                  |
|                           | Leontief Prize for Advancing the Frontiers of Economic Thought (award in economics) |
|                           | point in time: 2010                                                   |

### Links

- Wikidata page
- Wikipedia article
- Reasonator

### Identifiers

- SUDOC authorities: 033750807
- Library of Congress authority ID: n83133496
- ISNI: 0000 0...87046
- GND ID: 113900171

### Media

- image: Bina Agarwal at the World Economic Forum on India 2012.jpg
- Commons category: Bina Agarwal
Linking mechanism: external identifiers

- Property value: unique IDs from external database
- + URL stub in the property definition („formatter URL“)
- ~2,000 external identifier properties
- Examples:
  - VIAF
  - proteins
  - African plants
  - Swedish cultural heritage objects
  - TED conference speakers
Integrating GND – RePEc author mapping into Wikidata

In the EconBiz economics search portal, authors are identified differently:

- by GND ID in data from ZBW’s Econis catalog
- 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
Person identifiers at Wikidata and EconBiz

unknown overlap at project start
Linking via Wikidata

Wikidata-Properties for both identifier systems

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

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

- GND ID → Wikidata item → RAS ID
- RAS ID → Wikidata item → GND ID

where both properties have values (~760 items, as of 2017-04-25)
Supplement WD items with missing IDs values from mapping

Federated SPARQL queries revealed:
- 384 WD items identified by RAS ID missing GND ID
- 77 WD items identified by GND ID missing RAS ID

Adding missing IDs in bulk
- Transform to *QuickStatements2* input file ([SPARQL query](#), [script](#))
- Copy & paste to *QuickStatements2*
Bulk editing with *QuickStatements2*

Further simplification with upcoming release of *wdmapper* command line tool
Add identifiers for „most important“ authors from GND and RAS

- 4,600 Top 10% economists scraped from RePEc ranking page
- 18,000 GND authors with more than 30 publications in EconBiz

- Transform and load into Wikidata‘s *Mix’n’match* ([RePEc Top, GND economists (de)])
  - CSV file with ID, name, description
  - Confirm match candidates
  - Repeat adding missing identifiers from the mapping
Checking proposed matches in *Mix’n’match*

### RePEc Top

Top 10% Economists, per "Research Papers in Economics", Feb 17

<table>
<thead>
<tr>
<th>#</th>
<th>Author</th>
<th>Affiliation</th>
<th>Match Status</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Matthias Sutter</td>
<td>University of Gothenburg -&gt; School of Business, Economics and Law -&gt; Department of Economics; Innsbruck University -&gt; Faculty of Economics and Statistics -&gt; Institute for Public Economics, Innsbruck University -&gt; Faculty of Economics and Statistics (rank: 1280, publications: 58)</td>
<td>By Jneubert</td>
</tr>
<tr>
<td>2</td>
<td>Larry E. Jones</td>
<td>National Bureau of Economic Research (NBER); Federal Reserve Bank of Minneapolis -&gt; Research Department; University of Minnesota -&gt; Department of Economics (rank: 944, publications: 99)</td>
<td>Automatically matched</td>
</tr>
<tr>
<td>3</td>
<td>Larry Eugene Jones</td>
<td>US-American historian (*1940) ; American historian</td>
<td>Confirm</td>
</tr>
<tr>
<td>4</td>
<td>David J. Teece</td>
<td>University of California-Berkeley -&gt; Walter A. Haas School of Business (rank: 1109, publications: 62)</td>
<td>Automatically matched</td>
</tr>
<tr>
<td>5</td>
<td>David Teece</td>
<td>New Zealander economist and university teacher (*1945) ; American business academic</td>
<td>Confirm</td>
</tr>
<tr>
<td>6</td>
<td>Udo Ludwig</td>
<td>Halle Institute for Economic Research; University of Leipzig -&gt; Faculty of Economics and Business (rank: Automatically matched 1154, publications: 240)</td>
<td>Automatically matched</td>
</tr>
<tr>
<td>7</td>
<td>Udo Ludwig</td>
<td>German journalist (*1958) ; German journalist</td>
<td>Confirm</td>
</tr>
<tr>
<td>8</td>
<td>Jan Svejnar</td>
<td>Columbia University -&gt; School of International and Public Affairs (SIPA); Center for Economic Research and Graduate Education and Economics Institute (CERIE-ED) (rank: 1178, publications: 156)</td>
<td>Automatically matched</td>
</tr>
</tbody>
</table>
Add missing Wikidata items from mapping

- Verify missing authors indeed are not in Wikidata
- Generate Wikidata item data from existing mapping in QuickStatements2 input format (SPARQL query, script)
- 2179 new Wikidata items created, synthesized with values:
  - name from GND
  - occupation “economist” from occurrence in RePEc Top 10%
  - gender and date of birth/death (if available) from GND
  - description from GND’s info field and RePEc’s “works for”
Recommendations for item creation

- Pay attention to Wikidata’s notability criteria
- Explain your plan and ask for feedback in the Wikidata project chat
- Apply for a bot account to make mass edits (example)
- Source every statement (hints)
- Use QuickStatements2 (more convenient input format, batch mode, sources)
Result: the mapping in Wikidata

4087 Wikidata items with both GND and RAS IDs

• 3081 matches from ZBW’s mapping
• 1006 matches contributed by Wikidata users

(as of 2017-06-04)
RePEc “Top economists”: 60 % coverage

Wikidata economists (est. ~ 40,000)

EconBiz frequent GND

RAS Top

Wikidata economists is a rough estimate of the set of Wikidata persons in the field of economics. (Twice the number of those with the explicit occupation "economist".) – Numbers/sizes as of 2017-06-04
Additional immediate benefits

- Links to multilingual human-readable Wikipedia pages about authors (~1470 English, ~680 German, ~270 Spanish, etc.)
- Additional data from Wikidata
- Mappings to other authorities for free (e.g., ~1,600 RAS ↔ VIAF)
Strategic benefits

- Outsourced interface, storage and operation
- Crowdsourced mapping maintenance
- Wikidata has policies and tools for data quality
- Identifiers and items inserted by individual contributors or systematic efforts add up continuously and are available as Open Data
Mapping knowledge organization systems

External identifier properties for thesauri and classifications exist, e.g.
- GND subject headings
- Art & Architecture Thesaurus (Getty)
- UNESCO Thesaurus
- DDC classes

Upcoming: Mapping of „STW Thesaurus for Economics“ to Wikidata – started with STW sub-thesaurus „Geographical Names“
Beyond sameness – mapping relations

In rare cases (for locations), different mapping relations are required:

- broad or narrow matches – e.g.,
  „Lake Constance“ (Wikidata) < „Lake Constance region“ (STW)
- close matches – e.g.,
  „Overseas territories“ (Wikidata) \(\cong\) „Overseas territories“ (STW)
  („territories which have a special relationship with one of the member states of the EU“ in Wikidata, not defined in STW)

Property proposal under discussion (SKOS mapping relations as optional qualifiers for external id properties)
Workflow considerations

- Existing concordances (e.g., STW ./ GND) can be exploited for the creation of „mapping candidates“, see query based on
  STW descriptor $\langle$skos:exactMatch$\rangle$ GND subject heading $\langle$wdt:P227$\rangle$ Wikidata item
- 2,034 candidates for STW‘s 5,339 non-geographical descriptors
- Evaluation of 50 randomly selected entries revealed only 2 false, 7 more not exactly matching
- Therefore, automatic creation of these STW properties and intellectual check afterwards is an option
- Remainder can be handled by $Mix’n’match$
Conclusions

- Use of Wikidata as a linking hub for KOS looks very promising
- Different from existing “universal” KOS (think DDC), Wikidata is easily extensible
- Tools for matching as well as for consistency checks are in place
- Crowd contribution enabled and embraced
Thanks for listening!

Joachim Neubert
ZBW – Leibniz Information Centre for Economics
j.neubert@zbw.eu

http://zbw.eu/labs
https://github.com/zbw/repec-ras
https://github.com/zbw/sparql-queries/tree/master/wikidata