Linking Knowledge Organization Systems via Wikidata

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

Dublin Core Metadata Initiative Conference, Porto, 10.09.2018
The idea of linking hubs
Agenda

1. Suitability of Wikidata as linking hub  
   a. ZBW‘s experiences with mappings to Wikidata  
   b. Extending the model with mapping relations  
2. Tools used  
3. Indirect mappings to other vocabularies  
4. Outlook
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)

population (P1082) 8 173 900

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

> 1 reference
Linking mechanism: external identifiers

- Property value: unique IDs from external database
- + URL stub in the property definition ("formatter URL")

- More than 3,000 external identifier properties
- Examples:
  - VIAF
  - proteins
  - African plants
  - Swedish cultural heritage objects
  - TED conference speakers
ZBW’s experiences with mappings to Wikidata

1. **Moved a mapping of personal name authorities to Wikidata**
   (Research Paper for Economics author ID ./ GND ID)
   successfully done in 2017 (3,081 crosslinks, in Sept. 2018: 5,434)

2. **Now: STW Thesaurus for Economics**
   bilingual (German/English) thesaurus on economics, business economics and neighbouring field
   ~ 6,000 concepts
   started in mid-2017 with sub-thesaurus Geographic names (392)
   now on-going: sub-thesaurus Economic sectors (1520)
Industrialized countries

Industrieländer (german)

used for: High-income countries, Developed countries, Advanced countries

Narrower Terms

- Donor countries
- G7 countries
- OECD countries

Related Terms

- Industrial society

Subject Categories

- G.06 Political and economic regions

Links to other Thesauri and Vocabularies

= developed country (from Wikidata)
= Industriestaat (from DBpedia)
= industrial nation (from TheSoz)
Wikidata item about an economic concept

developed country (Q132453)
country with a developed industry and infrastructure

subclass of: every developed country is also a(n) country
instance of: developed country is not an instance of any other class

Classification

Direct superclasses
country

Direct subclasses
none

All subclasses
0

Statements

Own statements

number of out of school children
6026936
point in time: 2015

6670645
point in time: 2014

7147148
point in time: 2013

17 statements

opposite of
developing country (nation with a low living standard relative to other countries)

facet of
economic development (process and policies to improve the economic, political, and social well-being of people)

Identifiers

STW Thesaurus for Economics ID
10499.3

GACS ID
10630

UNESCO Thesaurus ID
concept939

MeSH ID
D019049

BNCF Thesaurus ID
32372

Item display in Sqid browser
Beyond sameness – mapping relations

- Wikidata external ids imply „sameness“ of linked concepts
- Even with geographic names, other mapping relations are required in some cases. Examples:
  - close matches – e.g.,
    „Yugoslavia“ (1918-1992) (Wikidata) ≅ „Yugoslavia (until 1990)“ (STW)
  - broad or narrow matches – e.g.,
    „Appenzell Innerrhoden“ (Wikidata) < „Appenzell“ (STW)
    „Appenzell Ausserrhoden“ (Wikidata) < „Appenzell“ (STW)
Introducing „mapping relation type“ (P4390)

- Introduced after a community discussion in October 2017
- To be used as a qualifier, with a fix set of values, at the closest item:
  - „exact match“
  - „close match“
  - „broad match“
  - „narrow match“
  - „related match“
  - strictly in line with the according SKOS mapping relations
- Applicable to any external-id property, for which the community agrees
# STW/Wikidata-Mapping in SKOS

<table>
<thead>
<tr>
<th>wd</th>
<th>skosRelation</th>
<th>stw</th>
</tr>
</thead>
<tbody>
<tr>
<td>Energy forecasting</td>
<td>skos:exactMatch</td>
<td>Energieprognose</td>
</tr>
<tr>
<td>Plantation</td>
<td>skos:closeMatch</td>
<td>Plantage</td>
</tr>
<tr>
<td>nuclear industry</td>
<td>skos:exactMatch</td>
<td>Nuklearindustrie</td>
</tr>
<tr>
<td>AKP-Gruppe</td>
<td>African, Caribbean and Pacific Group of States</td>
<td>skos:exactMatch</td>
</tr>
<tr>
<td>ASEAN</td>
<td>ASEAN</td>
<td>skos:exactMatch</td>
</tr>
<tr>
<td>Afghanistan</td>
<td>Afghanistan</td>
<td>skos:exactMatch</td>
</tr>
</tbody>
</table>

Extracted by a federated SPARQL query from STW and Wikidata endpoints

Usage of „mapping relation type“

<table>
<thead>
<tr>
<th>property</th>
<th>propertyLabel</th>
<th>items</th>
<th>statements</th>
</tr>
</thead>
<tbody>
<tr>
<td>wd:P2892</td>
<td>UMLS CUI</td>
<td>13339</td>
<td>15430</td>
</tr>
<tr>
<td>wd:P1550</td>
<td>Orphanet ID</td>
<td>9197</td>
<td>9794</td>
</tr>
<tr>
<td>wd:P699</td>
<td>Disease Ontology ID</td>
<td>7582</td>
<td>7594</td>
</tr>
<tr>
<td>wd:P492</td>
<td>OMIM ID</td>
<td>7183</td>
<td>8611</td>
</tr>
<tr>
<td>wd:P486</td>
<td>MeSH ID</td>
<td>6348</td>
<td>6471</td>
</tr>
<tr>
<td>wd:P1748</td>
<td>NCI Thesaurus ID</td>
<td>5391</td>
<td>5694</td>
</tr>
<tr>
<td>wd:P4342</td>
<td>Store norske leksikon ID</td>
<td>2653</td>
<td>2654</td>
</tr>
<tr>
<td>wd:P3911</td>
<td>STW Thesaurus for Economics ID</td>
<td>788</td>
<td>792</td>
</tr>
</tbody>
</table>
Wikidata as a universal linking hub

To sum up so far: Three characteristics make Wikidata suitable as an universal linking hub for the vast diversity of knowledge organization systems:

- easy extensibility with new properties for external identifiers
- immense fund of existing items, with the full set of SKOS mapping relations for more or less exact mappings to these
- immediate extensibility with new items
Tools used
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>Name</th>
<th>Institution</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: 944, publications: 99)</td>
<td>By Jneubert</td>
</tr>
<tr>
<td></td>
<td>Matthias Sutter</td>
<td>Austrian economist and university teacher (*1968)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>[Q13028348]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Larry E. Jones</td>
<td>National Bureau of Economic Research (NEBR); 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></td>
<td>Larry Eugene Jones</td>
<td>US-American historian (*1940)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>[Q1309051]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</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></td>
<td>David Teece</td>
<td>New Zealander economist and university teacher (*1948)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>[Q884277]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</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></td>
</tr>
<tr>
<td></td>
<td>Udo Ludwig</td>
<td>German journalist (*1958)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>[Q1471124]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</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 (CERGE-EI) (rank: 1178, publications: 156)</td>
<td>Automatically matched</td>
</tr>
</tbody>
</table>
Revealing quality problems

- minor issues, like missing labels in a particular language, can be fixed on the go
- duplicates (on both sides)
  - e.g., GND economists – solvable only in the long run
  - in Wikidata - easy to solve immediately by merging items
- clusters of overlapping concepts in Wikidata
  - e.g., for STW „Fisheries“, in Wikidata:
    - „fishing“ – as an activity
    - „fishery“ – as an economic branch
    - „commercial fishing“ as both an economic activity and sector
New item creation via *Quickstatements*

More information:
https://www.wikidata.org/wiki/Wikidata:WikiProject_Authority_control#Item_creation_from_a_thesaurus_concept_via_Quickstatements

**QS: Create item from STW**

**Offshore-Industrie | Offshore industry**

CREATE
LAST|Lde|"Offshore-Industrie"
LAST|Len|"offshore industry"
LAST|Ade|"Offshore-Technik"
LAST|Ade|"Offshore-Anlage"
LAST|Ade|"Offshore-Ausrüstung"
LAST|Aen|"offshore equipment"
LAST|P31|Q29028649|S248|Q26903352|S248|Q26903352
LAST|P227|"125537-9"|S248|Q26903352|S3911|"18394-5"|P4390|Q39893449

**Wasserstofftechnik | Hydrogen technology**

CREATE
LAST|Lde|"Wasserstofftechnik"
LAST|Len|"hydrogen technology"
LAST|P31|Q29028649|S248|Q26903352|S3911|"14634-3"|P4390|Q39893449

Leibniz-Informationssysteme
Wirtschaftsgeschichte
Leibniz Information Centre
for Economics
Excursus: Recommendations for item creation

- Pay attention to Wikidata’s notability criteria
- Do not pollute Wikidata with new items very close to existing ones – better link to the latter with an appropriate mapping relation
- When you start a larger endeavour, 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)
Quality control tools and procedures

- vandalism prevention and monitoring of suspect edits (e.g., new editor deleting statements)
- constraint definitions for properties
  - warnings during data input, when e.g. a supposedly unique identifier is added to more than one item
  - generated lists of constraint violations (e.g., for GND)
- when „mapping relation types“ are defined, modified constraints apply – see Maintenance reports for STW
- additional reports can be created via SPARQL queries
Earning links to other vocabularies
Knowledge organization systems linked to WD

External identifier properties for thesauri and classifications exist, e.g.

- GND subject headings
- Art & Architecture Thesaurus
- UNESCO Thesaurus
- DDC classes
- US National Cancer Institute Thesaurus
- Medical Subject Headings
- PATCOLS Archeology Thesaurus
- UK Parliament Thesaurus
- Hornborstel-Sachs class. of musical instruments
Some large vocabularies with high coverage

- 46,000 Gene ontology IDs, 740,000 NCBI Entrez Gene IDs
- 14,000 MeSH IDs (ca. 51 %)
- 15,000 AAT descriptors (ca. 40 %)
- 20,000 GND subject headings (ca. 15 %)

Vocabularies (aligned to BARTOC) and timelines:
http://coli-conc.gbv.de/concordances/wikidata/
## Indirect mapping STW – UNESCO thesaurus

Showing 1 to 50 of 283 entries

<table>
<thead>
<tr>
<th>stw</th>
<th>rel</th>
<th>unesco</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abortion</td>
<td>skos:exactMatch</td>
<td>Abortion</td>
</tr>
<tr>
<td>Afghanistan</td>
<td>skos:exactMatch</td>
<td>Afghanistan</td>
</tr>
<tr>
<td>Africa</td>
<td>skos:exactMatch</td>
<td>Africa</td>
</tr>
<tr>
<td>Agricultural cooperative</td>
<td>skos:exactMatch</td>
<td>Agricultural cooperatives</td>
</tr>
<tr>
<td>Agricultural policy</td>
<td>skos:exactMatch</td>
<td>Agricultural policy</td>
</tr>
<tr>
<td>Agricultural technology</td>
<td>skos:exactMatch</td>
<td>Agricultural engineering</td>
</tr>
</tbody>
</table>

Derived dynamically through a query against Wikidata, STW and UNESCO endpoints, restricted to exact matches for STW and presuming exact matches for UNESCO thesaurus

Future work

- extending and evaluating indirect mappings
- monitoring a mapping in regard to community changes (wdmapper tool)
- mechanisms for exception lists: adding or removing triples from an extracted or indirectly generated mapping, to adapt it to a particular custom use
Thanks for listening!

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

http://zbw.eu/labs
https://hackmd.io/2bfSBXtjQim8Ega4OQhwwQ# (GND/RePEc)
https://github.com/zbw/stw-mappings