Linking Knowledge Organization Systems via Wikidata

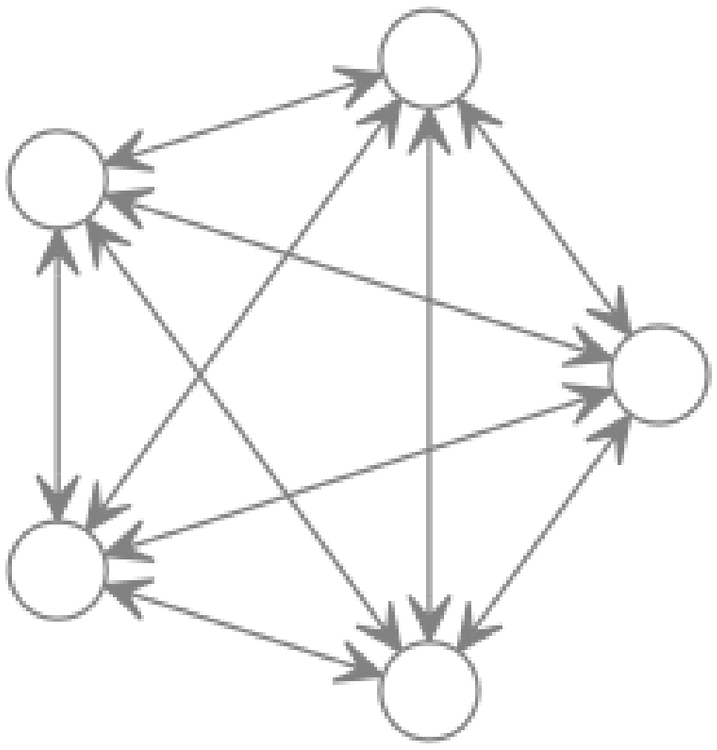
Joachim Neubert

ZBW – Leibniz Information Centre for Economics, Kiel/Hamburg

Dublin Core Metadata Initiative Conference, Porto, 10.09.2018

Leibniz-Informationszentrum
Wirtschaft
Leibniz Information Centre
for Economics

The idea of linking hubs



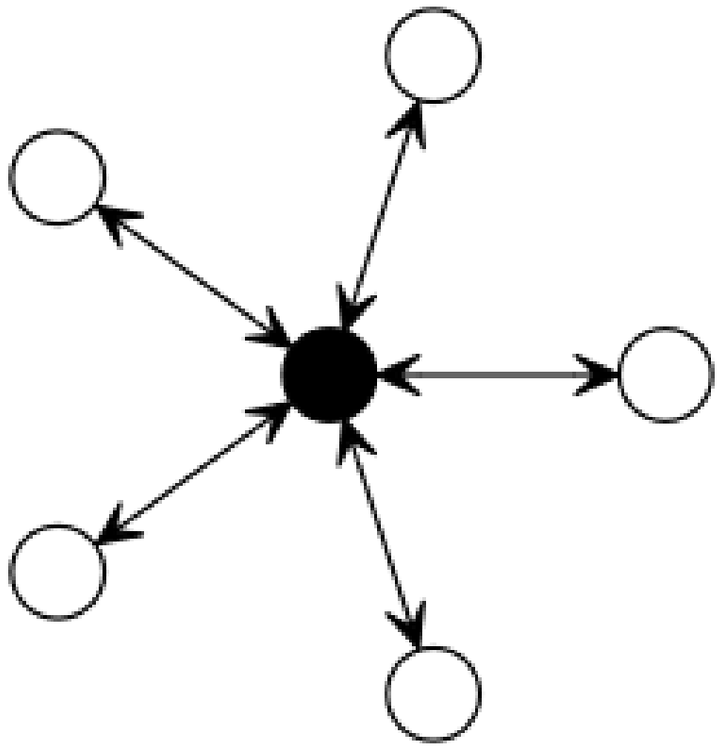


Image by Jakob Voß (ELAG 2017)

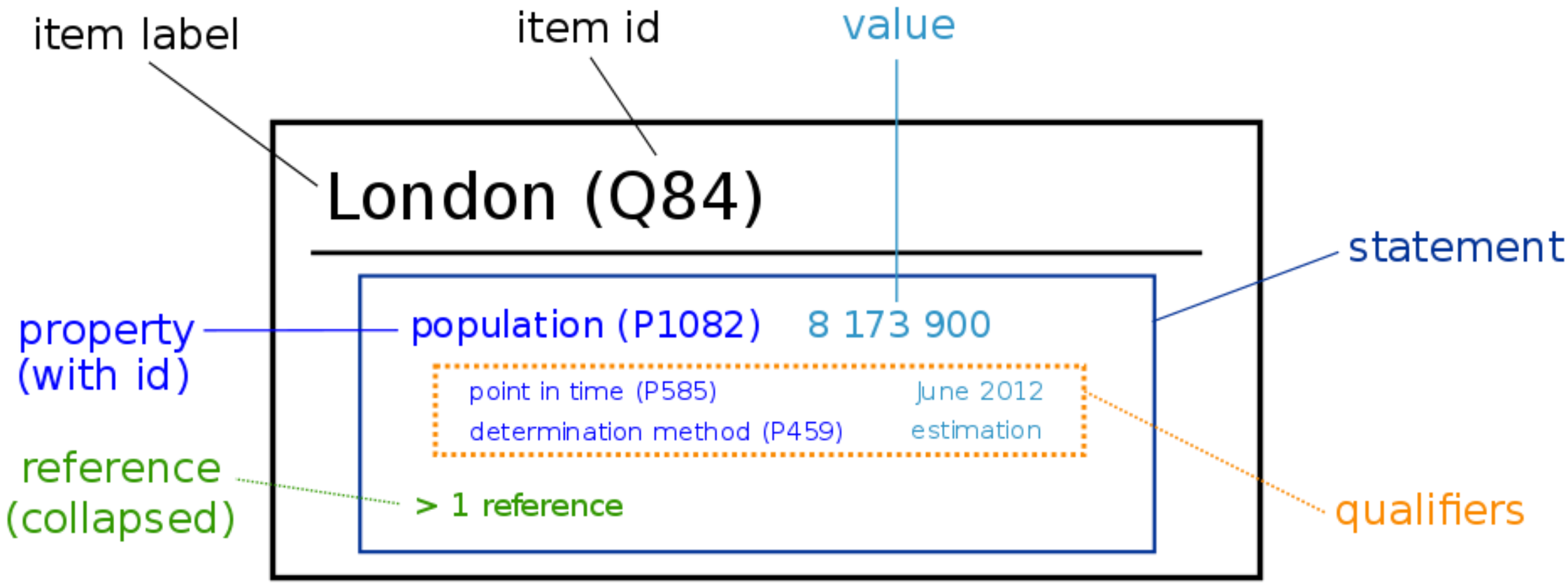
Agenda

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)
  - http://query.wikidata.org/ (SPARQL)
  - JSON API & database dumps

Wikidata statements



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 Econmics author ID ./. GND ID)
   sucessfully 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)

Home
STW Relaunch
Alphabetical descriptor list
Mappings
Versions
Web Services
Downloads
About

- V Economics
- B Business economics
- W Economic sectors
- P Commodities
- N Related subject areas
- G Geographic names
- A General descriptors

Industrialized countries EB

Industrieländer (german)

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

Narrower Terms

- Donor countries EB
- G7 countries EB
- OECD countries EB

Related Terms

- Industrial society EB

Subject Categories

- G.06 Political and economic regions

Links to other Thesauri and Vocabularies

= developed country (from Wikidata) W
≅ Industriestaat (from DBpedia)
= industrial nation (from TheSoz)

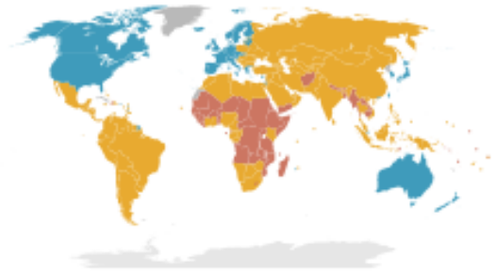
Wikipedia: developed country

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 2860
Direct subclasses	*none*
All subclasses	0

Statements

Own statements | From related entities

number of out of school children (17 statements)	6026936 ★ point in time : 2015
	6679645 point in time : 2014
	7147148 point in time : 2013
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)

Links

Wikidata page
Wikipedia article
Reasonator

Identifiers

STW Thesaurus for Economics ID	10499-3 mapping relation type : exact match (SKOS mapping relation (for use with P4390): two concepts have equivalent meaning, and the link can be exploited across a wide range of applications and schemes. The link is meant to be transitive.)
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

Showing 1 to 50 of 794 entries Search: Show 50

wd	skosRelation	stw
| Energy forecasting	skos:exactMatch	Energieprognose | Energy forecast
| Plantation	skos:closeMatch	Plantage | Plantation
| nuclear industry	skos:exactMatch	Nuklearindustrie | Nuclear industry
AKP-Gruppe | African, Caribbean and Pacific Group of States	skos:exactMatch	AKP-Staaten | ACP countries
ASEAN | ASEAN	skos:exactMatch	ASEAN-Staaten | ASEAN countries
Afghanistan | Afghanistan	skos:exactMatch	Afghanistan | Afghanistan

Extracted by a federated SPARQL query from STW and Wikidata endpoints

http://zbw.eu/beta/sparql-lab/?endpoint=http://zbw.eu/beta/sparql/stw/query&queryRef=https://api.github.com/repos/zbw/sparql-queries/contents/stw/wikidata_mapping.rq

Usage of „mapping relation type“

25 Ergebnisse in 2904 ms </> Code Herunterladen Link

property	propertyLabel	items	statements
wd:P2892	UMLS CUI	13339	15430
wd:P1550	Orphanet ID	9197	9794
wd:P699	Disease Ontology ID	7582	7594
wd:P492	OMIM ID	7183	8611
wd:P486	MeSH ID	6348	6471
wd:P1748	NCI Thesaurus ID	5391	5694
wd:P4342	Store norske leksikon ID	2653	2654
wd:P3911	STW Thesaurus for Economics ID	788	792

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*

Mix'n'match | English | Welcome, Jneubert | Search | Search

RePEc Top

Action

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

1 2 3 4 5 6 7 8

# Matthias Sutter	University of Gothenburg -> School of Business, Economics and Law -> Department of Economics; Innsbruck University -> Faculty of Economics and Statistics -> Institute for Public Economics; Innsbruck University -> Faculty of Economics and Statistics (rank:	By Jneubert
Matthias Sutter [Q1910346]	Austrian economist and university teacher (*1968) ♂; Austrian economist	Remove
# Larry E. Jones	National Bureau of Economic Research (NBER); Federal Reserve Bank of Minneapolis -> Research Department; University of Minnesota -> Department of Economics (rank: 944, publications: 99)	*Automatically matched*
Larry Eugene Jones [Q1806051]	US-American historian (*1940) ♂; American historian	Confirm | Remove
# David J. Teece	University of California-Berkeley -> Walter A. Haas School of Business (rank: 1109, publications: 62)	*Automatically matched*
David Teece [Q984277]	New Zealander economist and university teacher (*1948) ♂; American business academic	Confirm | Remove
# Udo Ludwig	Halle Institute for Economic Research; University of Leipzig -> Faculty of Economics and Business (rank: 1154, publications: 240)	*Automatically matched*
Udo Ludwig [Q1471124]	German journalist (*1958) ♂; German journalist	Confirm | Remove
# Jan Svejnar	Columbia University -> School of International and Public Affairs (SIPA); Center for Economic Research and Graduate Education and Economics Institute (CERGE-EI) (rank: 1178, publications: 156)	*Automatically matched*
Jan Švejnar [Q1682473]	US-American-czech republic economist, educationist, and university teacher (*1952) ♂; IZA Prize in Labor Economics; Czech Czech president candidate (2008) and economist	Confirm | Remove

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*

QuickStatements | English | New batch | Last batches | Chat | Git | Help | Jneubert | Your last batches

Create new command batch for Wikidata as batch name

```
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|S
LAST|P227|"4125537-9"|S248|Q2690335
LAST|P3911|"18394-5"|P4390|Q3989344
```

Import V1 commands | Import CSV comman

More information:

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

ZBW Leibniz-Informationszentrum Wirtschaft
Leibniz Information Centre for Economics

QS: Create item from STW

Offshore-Industrie | Offshore industry

```
CREATE
LAST|Lde|"Offshore-Industrie"
LAST|Len|"offshore industry"

LAST|Ade|"Offshore-Technik"
LAST|Ade|"Bohrinsel"
LAST|Ade|"Bohrplattform"
LAST|Ade|"Offshore-Anlage"
LAST|Ade|"Offshore-Ausrüstung"
LAST|Aen|"offshore equipment"

LAST|P31|Q29028649|S248|Q26903352|S3911|"18394-5"|S1476|en:"Offshore industry"|S813|+2018-09-05T0
LAST|P227|"4125537-9"|S248|Q26903352|S3911|"18394-5"|S1476|en:"Offshore industry"|S813|+2018-09-0
LAST|P3911|"18394-5"|P4390|Q39893449
```

Wasserstofftechnik | Hydrogen technology

```
CREATE
LAST|Lde|"Wasserstofftechnik"
LAST|Len|"hydrogen technology"

LAST|P31|Q29028649|S248|Q26903352|S3911|"14634-3"|S1476|en:"Hydrogen technology"|S813|+2018-09-05
LAST|P227|"4064784-5"|S248|Q26903352|S3911|"14634-3"|S1476|en:"Hydrogen technology"|S813|+2018-09
LAST|P3911|"14634-3"|P4390|Q39893449
```

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 muical 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 Search:

stw	rel	unesco
Abortion	skos:exactMatch	Abortion
Afghanistan	skos:exactMatch	Afghanistan
Africa	skos:exactMatch	Africa
Agricultural cooperative	skos:exactMatch	Agricultural cooperatives
Agricultural policy	skos:exactMatch	Agricultural policy
Agricultural technology	skos:exactMatch	Agricultural engineering

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

http://zbw.eu/beta/sparql-lab/?endpoint=http://zbw.eu/beta/sparql/stw/query&queryRef=https://api.github.com/repos/zbw/sparql-queries/contents/stw/indirect_mapping_via_wd.rq

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