Wikidata as a linking hub for knowledge organization systems?

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

Joachim Neubert[0000-0002-8086-185X]

ZBW – Leibniz Information Centre for Economics, Kiel/Hamburg, Germany

Abstract. Wikidata has been created in order to support all of the roughly 300 Wikipedia projects. Besides interlinking all Wikipedia pages about a specific item – e.g., a person - in different languages, it also connects to more than 1500 different sources of authority information.
We will present lessons learned from using Wikidata as a linking hub for two personal name authorities in economics (GND and RePEc author identifiers) and demonstrate the benefits of moving a mapping from a closed environment to Wikidata as a public and community-curated linking hub. We will further ask to what extent these experiences can be transferred to knowledge organization systems and how the limitation to simple 1:1 relationships (as for authorities) can be overcome. Using STW Thesaurus for Economics as an example, we will investigate how we can make use of existing cross-concordances to "seed" Wikidata with external identifiers, and how transitive mappings to yet unmapped vocabularies can be earned.

Keywords: Linked Open Data, Knowledge Organization System, Authority File, Alignment, Mapping, Cross-Concordance, Wikidata, GND, RePEc, STW.

1 Wikidata – structure and suitability as an authority linking hub

Wikidata was launched by the Wikimedia foundation in 2012, in order to create a shared knowledge base and provide structured data for the Wikipedias in different languages, Wikimedia Commons and other projects of the foundation. Like Wikipedia itself, it can be enhanced and edited by everybody. Different from Wikipedia, the underlying data structures are constantly and quickly evolving and can be easily enhanced in consensus-oriented community processes.

1.1 Items, properties and external identifiers

The basic building blocks of Wikidata are items, which are identified by an abstract identifier ("Q value") and can be named and further specified by labels and descriptions in different languages. Properties, also identified by abstract identifiers ("P val-

ues") can be added to the items, in order to specify e.g. the date of birth for a person or the surface area of a country in their values.[1] The available data can be exported as RDF and queried in a public SPARQL endpoint[1]. The labels and values of items and properties can be directly accessed in Wikipedia projects through template mechanisms.[2]

In the context of this paper, properties of type "external identifier"[3] are of particular interest. Their values uniquely identify the item in an external database. If a URL stub (called "formatter URL") is defined for the property, they are displayed as links in Wikidata. This simple mechanism is extensively used to connect Wikidata items to authority files for people, works, places, organizations and various other types of entities. [2] The total number of external id properties has reached almost 2,000, with more than 1,500 classified as "properties for authority control"[4]. This includes widely used resources such as VIAF or GeoNames, but also very domain-specific identifiers for proteins, Swedish cultural heritage, African plants or speakers at TED conferences.

Implicitly, Wikidata serves as an organically growing hub, linking all these authorities. While often the external id properties are populated sparsely, for certain properties the numbers of occurrences are high (GeoNames with ~3 million of ~6.5m "location" items, VIAF with ~1 million, mostly persons, in relation to ~3.5m "humans") and constantly increasing. When considering the use of Wikidata as a linking hub in a systematic approach, it is however crucial if Wikidata can be extended in ways that allow to map external authority files or knowledge organization systems completely.

1.2 Policies and community communication structures

New properties for authority control can be added in a "property proposal" process. Based on a template, properties can be suggested by everybody and are discussed for at least a week[5]. If the author is responsive to comments, some members of the community support the proposal, and there are no or only a few opponents, the property is created and can be used immediately.

In Wikipedia, only "notable topics" of general encyclopedic interest are allowed as pages – what would prohibit the mapping for large portions of library authority files. In Wikidata however, notability criteria are much more relaxed. Besides everything which has a page in any of the language-specific Wikipedias, an item may be added if "it refers to an instance of a clearly identifiable conceptual or material entity. The entity must be notable, in the sense that it can be described using serious and publicly available references."[6] The community seems to agree on the interpretation that authority files provide such "serious and publicly available references".[7]

---

1 https://query.wikidata.org/
2 https://www.wikidata.org/wiki/Wikidata:How_to_use_data_on_Wikimedia_projects
3 https://www.wikidata.org/wiki/Help:Data_type#External_identifier
4 numbers in this section as of 2017-08-31
5 e.g., https://www.wikidata.org/wiki/Wikidata:Property_proposal/Australian_Women%27s_Register_ID
6 https://www.wikidata.org/wiki/Wikidata:Notability
7 see e.g. statement by Wikidata admin ChristianKl in the discussion referenced in footnote 23

Wikidata's "Project chat"[8] is an appropriate place to discuss, e.g., a larger mapping project which would include the creation of multiple new items. Feedback there often includes valuable hints, especially for new actors in the community. If tool-supported mass edits are planned, in addition a "bot flag" should be a requested.[9]

During the mapping projects described in the remainder of this paper, the communication structures of Wikidata as described here proved both helpful and effective.

2 The GND – RePEc Author mapping project

The *EconBiz*[10] search portal comprises publications in economics from different sources. In some of these sources, in total 460,000 authors are unambiguously identified by identifiers of the Integrated Authority Files (GND). In another source, Research Papers for Economics (RePEc), a comprehensive database of working papers and articles in economics, about 50,000 authors are identified by persistent IDs of the RePEc Author Service (RAS). The service allows authors to claim their papers and is used to create rankings of economists and their institutions, which is an incentive for high data quality and current updates. While GND is well known and linked to many other authorities, RAS had no links to any other personal identifier systems.

Since it is highly desirable to be able to show *all* papers of a certain author in EconBiz across different sources, some years ago a mapping of GND-RAS author IDs was created at ZBW, automatically derived with a high degree of trustworthiness, which yielded 3081 pairs of identifiers.[11] Though this set covered only a small fraction of the supposed overlap in both identifier systems, it could serve as a starting point for further intellectual or automatic mapping efforts (which in itself are not subject of this paper). The maintenance and possible future extension of the existing mapping however stayed as an unresolved issue. Creating a custom application for that purpose, particularly when it comes to additions or amendments by non-technical staff or users outside of ZBW, would not only have required some programming effort, but also have overstretched the available operating capacities. Wikidata as a publicly available database offered another opportunity.

2.1 Initial situation in Wikidata

Economists are already well represented among the 3.4 million persons in Wikidata, though the precise extent is difficult to estimate.[3] Although, properties for linking GND and RePEc author identifiers to Wikidata items were already in place:

- P227 "GND ID", in ~375,000 items
- P2428 "RePEc Short-ID" (further-on: RAS ID), in ~2,200 items

[8] https://www.wikidata.org/wiki/Wikidata:Project_chat
[9] https://www.wikidata.org/wiki/Wikidata:Bots#Approval_process
[10] http://www.econbiz.de
[11] https://github.com/zbw/repec-ras/blob/nkos2017/doc/RAS-GND-author-id-mapping.md

- both properties in ~760 items[12]

The relative amounts of IDs in EconBiz and Wikidata is illustrated by Fig. 1. For both properties, the "single value" and "distinct values" constraints are defined, so that (with rare exceptions) a 1:1 relation between the authority entry and the Wikidata item should exist. That, in turn, means that a 1:1 relation between both authority entries can be assumed.

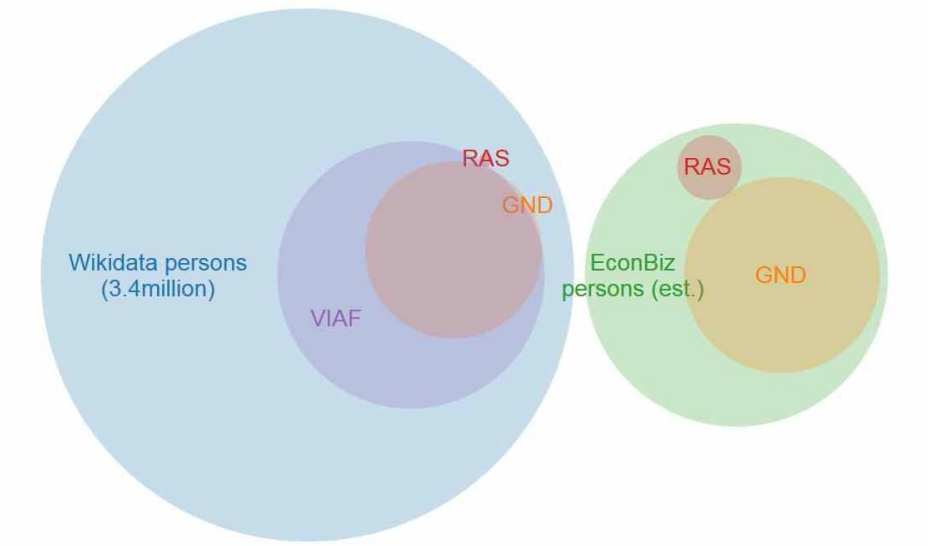

Fig. 1. Person identifiers in Wikidata and EconBiz, with unknown overlap at the beginning of the project (the number of persons in EconBiz is a very rough estimate, because most names – outside GND and RAS – are not disambiguated)

Since many economists have Wikipedia pages, what means that according Wikidata items have been created routinely, the first task was finding these items and adding GND and/or RAS identifiers to them. The second task was adding items for those persons which did not already exist in Wikidata.

2.2 Adding mapping-derived identifiers to Wikidata items

For items already identified by either GND or RAS, the reciprocal identifiers where added automatically: A federated SPARQL query[13] on the mapping and the public Wikidata endpoint retrieved the items and the missing IDs. A script[14] transformed that

[12] all numbers as of 2017-04-25

[13] https://github.com/zbw/repec-ras/blob/nkos2017/sparql/missing_ids_in_wikidata_from_mapping.rq

[14] https://github.com/zbw/repec-ras/blob/nkos2017/bin/create_missing_ids_in_wikidata_from_mapping.pl

into input for Wikidata's *QuickStatements2*[15] tool, which allows adding statements (as well as new items) to Wikidata. The tool takes csv-formatted input via a web form and applies it in batch to the live dataset (see Fig. 2).

That step resulted in 384 added GND IDs to items identified by RAS ID, and, in the reverse direction, 77 added RAS IDs to items identified by GND ID. For the future, it is expected that tools like *wdmapper*[16] will facilitate such operations.

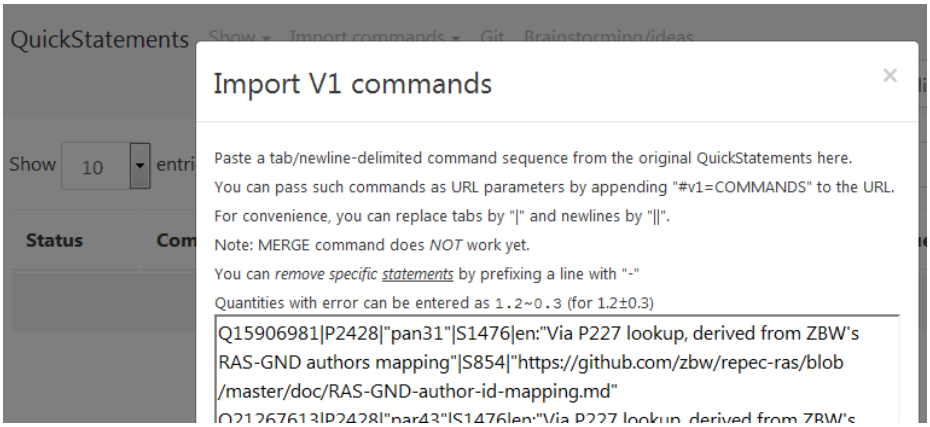

Fig. 2. Import statements for *QuickStatements2*. The first input line adds the RAS ID "pan31" to the item for the economist James Andreoni. The rest of the input line creates a reference to ZBWs mapping for this statement and so allows tracking its provenance in Wikidata

2.3 Identifying more Wikidata items

Obviously, the previous step left out the already existing economists in Wikidata, which up to then had neither a GND nor a RAS ID. Therefore, these items had to be identified by adding one of the identifiers. A semi-automatic approach was applied to that end, starting with the "most important" persons from both datasets. That was extended in an automatic step, taking advantage of existing VIAF identifiers (a step which could have been also the first one).

For RePEc, the "Top economists" ranking page[17] (~4,600 authors) was scraped and cross-linked to a custom-created basic RDF dataset of the RePEc authors.[18] The result was transformed to an input file for Wikidata's *Mix'n'match*[19] tool, which had been developed for the alignment of external catalogs with Wikidata. The tool takes a simple CSV file, consisting of a name, a description and an identifier, and tries to automatically match against Wikidata labels. In a subsequent interactive step, it allows to

[15] https://tools.wmflabs.org/quickstatements/
[16] https://github.com/gbv/wdmapper
[17] https://ideas.repec.org/top/top.person.all.html
[18] For details, see https://github.com/zbw/repec-ras
[19] https://tools.wmflabs.org/mix-n-match/

confirm or remove every match. If confirmed, the identifier is automatically added as value to the according property of the matched Wikidata item.

For GND, all authors with more than 30 publications in EconBiz where selected in a custom SPARQL endpoint. Just as the "RePEc Top" dataset[20], a "GND economists (de)" dataset[21] with ~18,000 GND IDs, names and descriptions was loaded into *Mix'n'match* and aligned to Wikidata.

Becoming more familiar with the Wikidata-related tools, policies and procedures, existing VIAF property values were exploited as another opportunity for seeding GND IDs in Wikidata. In a federated SPARQL query on a custom VIAF and the public Wikidata endpoint, about 12,000 missing GND IDs were determined and added to Wikidata items which had been identified by VIAF ID.

After each of these steps, the first task – adding mapping-derived GND or RAS identifiers – was repeated. That resulted in 1908 Wikidata items carrying both IDs. Since ZBWs author mapping based on at least 10 matching publications, the alignment of high-frequency resp. highly-ranked GND and RePEc authors made it highly probable that authors already present in Wikidata were identified in the previous steps. That reduced the danger of creating duplicates in the following task.

2.4 Creating new Wikidata items from the mapped authorities

For the rest of the authors in the mapping, 2179 new Wikidata items were created. This task was carried out again by the *QuickStatements2* tool, for which the input statements were created by a script[22], based on a SPARQL query on the aforementioned endpoints for RePEc authors and GND entries. The input statements were derived from both authorities, in the following fashion:

- the label (name of the person) was taken from GND
- the occupation "economist" was derived from RePEc (and in particular from the occurrence in its "Top Economists" list)
- gender and date of birth/death were taken from GND (if available)
- the English description was a concatenated string "economist" plus the affiliations from RePEc
- the German description was a concatenated string "Wirtschaftswissenschaftler/in" plus the affiliations from GND

The use of Wikidata's description field for affiliations was a makeshift: In the absence of an existing mapping of RePEc (and mostly also GND) organizations to Wikidata, it allows for better identification of the individual researchers. In a later step, when according organization/institute items exist in Wikidata and mappings are in place, the items for authors can be supplemented step-by-step by formal "affiliation" (P1416) statements.

---

[20] https://tools.wmflabs.org/mix-n-match/#/catalog/406
[21] https://tools.wmflabs.org/mix-n-match/#/catalog/431
[22] https://github.com/zbw/repec-ras/blob/nkos2017/bin/create_missing_wikidata.pl

According to Wikidata's policy, an extensive reference to the source for each statement in the synthesized new Wikidata item was added.[23]

The creation of items in an automated fashion involves the danger of duplicates. However, such duplicates turned up only in very few cases. They have been solved by merging items, which technically is very easy in Wikidata[24]. Interestingly, a number of "fake duplicates" indeed revealed multifarious quality issues, in Wikidata and in both of the authority files, which, too, have been subsequently resolved.[25]

2.5 Results

The immediate result of the project was:

- all of the 3081 pairs of identifiers from the initial mapping by ZBW is incorporated now in Wikidata items
- 1006 Wikidata items in addition to these also have both identifiers (created by individual Wikidata editors, or the efforts described above)

While that still is only a beginning, given the total amount of authors represented in EconBiz, it is a significant share of the "most important" ones (Fig. 3):

[23] for details, see https://www.wikidata.org/wiki/Wikidata:Project_chat/Archive/2017/05#Source_statements_for_items_syntesized_from_authorities_-_recommendations.3F

[24] documented extensively in https://www.wikidata.org/wiki/Help:Merge

[25] see details (in German) https://www.wikidata.org/wiki/User_talk:Jneubert#Dubletten

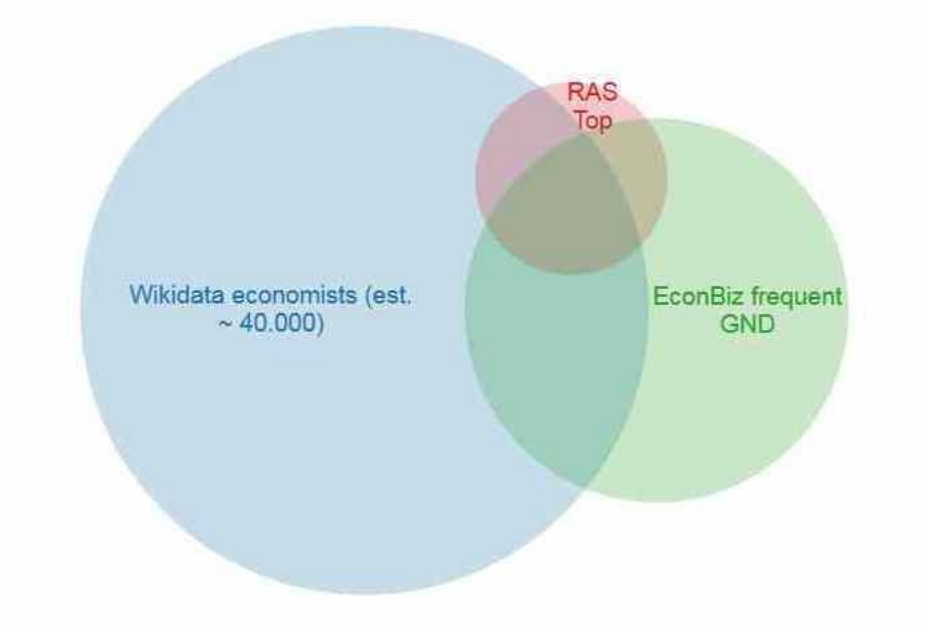

Fig. 3. Top 10 % RAS and frequent GND in EconBiz (> 30 publications). "Wikidata economists" is a rough estimate of the amount of persons in the field of economics (twice the number of those with the explicit occupation "economist")

More than 60 % of the "Top 10 %" RePEc authors are covered now by Wikidata and mapped to GND.[26]

The mapping data can be retrieved by everyone, via SPARQL queries, by specialized tools such as *wdmapper*, or as part of the Wikidata dumps. What is more, it can be extended by everybody – either as a by-product of individual edits adding identifiers to persons in Wikidata, or by a directed approach. For directed extensions, any subset can be used as a starting point: Either a new version of the above mentioned ranking, or other rankings also published by RePEc, covering in particular female, or economists from e.g. Latin America; or all identifiers from a particular institution, either derived from GND or RAS. The results of all such efforts are available at once and add up continuously.

Yet, the benefits of using Wikidata cannot be reduced to the publication and maintenance of mapping itself. In many cases it offers much more than just a linking point for two identifiers:

- links to Wikipedia pages about the authors, possibly in multiple languages
- rich data about the authors in defined formats, sometimes with explicit provenance information
- access to pictures etc. from Wikimedia Commons, or quotations from Wikiquote
- links to multiple other authorities

[26] numbers as of 2017-06-04

As an example for the latter, the in total 4560 RAS identifiers in Wikidata are already mapped to 1628 VIAF and 1282 LoC authority IDs (while ORCID with 62 IDs is still remarkably low). At the same time, these RePEc-connected items were linked to 1467 English, 681German and 269 Spanish Wikipedia pages which provide rich human-readable information.[27]

For ZBW, "releasing" the dataset into Wikidata as a trustworthy and sustainable public database not only spares the "technical" costs of data ownership (programming, storage, operating, for access and for maintenance). Responsibility for extending, amending and keeping the dataset current can be shared with many other interested parties and individuals.

3 Considerations for mapping a thesaurus to Wikidata

Wikidata covers not only individual material entities, like persons, organizations or places, but also abstract concepts, and in doing so overlaps with knowledge organization systems in general. External identifiers are in place for descriptors from thesauri (such as the AAT or UNESCO thesaurus) or classes from classifications (e.g., DDC).

STW Thesaurus for economics[28] is used at ZBW for indexing publications and for search support in different scenarios [4]. It has been made available in SKOS [5] under an Open Database License. A mapping of STW to Wikidata is desirable for two reasons: Firstly, in order to provide on the descriptor pages of the web representation links to Wikipedia pages in English and German, which can help users with extended explanations and context of concepts. Secondly, in order to exploit links to other knowledge organization systems which have already been mapped to Wikidata.

Therefore, a property proposal for the STW descriptor ID as external identifier has been submitted to Wikidata, discussed and accepted.[29] The external identifier properties of Wikidata, however, can only cover the case of an "equivalent" match (without explicit ontological meaning). That is straightforward for humans, already less so for organizations, but not sufficient for knowledge organization systems in general.

3.1 Beyond sameness - mapping properties in Wikidata

The limitations became clearly visible as a result of mapping a first section of STW to Wikidata – namely the geographic names sub-thesaurus. Since Wikidata is comprehensive in respect to locations, most of the 392 descriptors could be matched straightforward (after having been loaded into the Mix-n-match tool described above). However, different kinds of non-exact matches were found, which are not covered well by the use of plain external identifier properties:

- broader or narrower matches – e.g., STW has "Lake Constance region", whereas Wikidata has "Lake Constance".

[27] numbers as of 2017-08-31

[28] http://zbw.eu/stw

[29] https://www.wikidata.org/wiki/Wikidata:Property_proposal/STW_Thesaurus_for_Economics_ID

- close matches – e.g., STW and Wikidata both have "overseas territories", but in Wikidata these are defined as "territories that have a special relationship with one of the member states of the EU", whereas in STW no explicit definition is given (and the publications indexed with this descriptor may or may not cover, e.g., the Russian colonization in America). The missing or vague definitions (what means "special relationship"?) don't even allow to state that one concept is broader than the other.

Exact matches in SKOS, or exact equivalences in ISO-25964-2, are meant to be transitive, and usable across a broad range of applications, so unintended consequences may occur when used carelessly or in absence of alternatives.

In online discussions, experienced Wikidata users suggested to create additional Wikidata items, which would match the external concept exactly. While Wikidata makes that very easy, and it could be a solution in some cases, it cannot solve the problem completely:

1) Sometimes, the differences in meaning are only slight (see the "overseas territories" example above). Adding another item and trying to define it more rigidly would in such cases proliferate new items (at times, only used with that exact meaning in some scholarly sub-community), which would not be linked to anything but to the external identifier.
2) Even if the introduction of a new item could make sense for Wikidata, because it would describe a clearly definable entity, it often would
    a. require a considerable amount of research (what is meant exactly by "Lake Constance region" in Germany, in Swiss and in Austria?) – which would be valuable when amended with further information such as surface area or population, but normally is out of reach for a vocabulary mapping project; and
    b. betray the original purpose of the mapping, namely connecting a KOS to existing concepts elsewhere – in the particular in case of Wikidata and "Lake Constance", connecting it to dozens of Wikipedia pages about the lake and its surroundings, and also to more than 15 other external identifiers. With a new item created, that value would be lost for links from the originating KOS and only be retrievable by additional navigation steps in Wikidata.

For these reasons, another property proposal in Wikidata was submitted, which would allow modifying each assignment of an external identifier to a Wikidata item individually: A relationship could be qualified differently as "exact match", "close match", "narrow match", "broad match" or "related match", when appropriate.[30] The definitions of these qualifier values refer to the according SKOS mapping relations[31]. When adopted, the proposal would make Wikidata formally fit as a universal linking hub for knowledge organization systems.

[30] https://www.wikidata.org/wiki/Wikidata:Property_proposal/mapping_relation_type
[31] https://www.w3.org/TR/skos-reference/#mapping

3.2 Exploiting existing mappings for mapping candidates

When a KOS new to Wikidata is already mapped to another KOS by prior efforts, and both have external identifier properties in Wikidata defined, this prior mapping can be exploited for seeding a mapping of the newly connected KOS to Wikidata. STW descriptors, for example, are already mapped to the subject headings in the Integrated Authority File (GND), with currently more than 4,700 skos:exactMatch relations. Using these mappings in a query against Wikidata[32], it turns out that 2,034 of the 5,339 non-geographical STW descriptors are already transitively linked to Wikidata items via GND ID.

Out of this set of "mapping candidates", entries from 50 randomly selected GND IDs were evaluated intellectually. 42 of the entries were correct and represented an exact match. For 7 entries, the link would have to be modified with another mapping relation (2 close, 4 broad, 1 related matches).

Problematic cases often unveil only on second sight. The STW descriptor "documentation", e.g., covers what in German is called "Dokumentationswesen" – the application of information science to practical use. In Wikidata, "documentation" is defined as "set of documents providing knowledge" (in English), while in German it's intended to mean "Nutzbarmachung von Informationen zur weiteren Verwendung" (utilization of information for further use) – which is much closer to the meaning in STW. The obvious issue – differing definitions of concepts in different languages – may be spotted more often in Wikidata – which is under heavy development by many users – than in thesauri developed by a small team. The underlying problem, that certain concepts may be not easily mapped across languages at all, is an issue for all multilingual KOS, and even more for mapping concepts across such KOS.

Only 2 of the 51 reviewed mappings[33] were completely wrong, i.e. would not make sense even with another mapping relation type. So it seems economical – but is not yet decided – to add all the derived mappings automatically, and then check them one by one, modifying or deleting the relations which don't fit.

In several cases, the causes for inexact or plain wrong mappings were not found in the assignment of GNDs to Wikidata items, but in problematic relations in the pre-existing GND/STW cross concordance. So checking all derived relations can be also seen as a quite effective measure of quality control for the pre-existing mapping.

4 Future work

For creating a complete mapping of STW with Wikidata, ZBW plans to use *Mix-n-match* again. A limitation of the tool however is, that it does not take advantage of

[32] 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_candidates_via_gnd.rq, as run at 2017-08-31

[33] Because one STW descriptor wrongly had two targets in GND, the total number of reviewed mappings was 51.

multilingualism of external vocabularies. So for STW, either German or English preferred labels can be loaded and matched. Additionally, there seems no given way to exploit synonyms defined in the external vocabulary to improve matching results. While the first limitation perhaps can be worked around by loading a German and English version of the thesaurus as formally independent catalogs with shared identifiers, for the second limitation no workaround is in sight.

After finalizing a complete mapping of STW to Wikidata, this mapping could be exploited to connect to identifiers of further KOS. A join from the above mentioned 2034 STW descriptors to Wikidata items (via GND, without removing non-exact matches) reveals that, e.g., 279 Library of Congress authority IDs (P244), 195 Encyclopædia Britannica Online IDs (P1417) or 137 Le Monde diplomatique subject IDs (P3612) are connected to the matching Wikidata items. Extending that to the whole of STW and removing inexact matches may double the numbers. These mappings are sparse currently – the potential overlap is probably much larger. However, they can be gained as a windfall profit from mapping STW singly to Wikidata, and can be enhanced any time by the owner of STW as well as by any interested third party. As with the authority mapping described in the first part of this paper, every small addition or coordinated approach will add up to enrich an interlinked set of knowledge organization systems. Due to the nature of Wikidata, this network of concepts from many domains is accessible and maintainable as a single open and collaborative knowledge base.

5 Acknowledgements

With thanks to Kim Plassmeier and Sascha Junker, who created ZBW's GND - Repec authors mapping, to Jeanne Deillon, who mapped the STW Geo descriptors, to Manfred Faden, who reviewed the STW-GND-WD test set, and to Jakob Voß, who introduced me to Wikidata.

6 References

1. Krötzsch, M., Vrandečić, D.: Wikidata: A Free Collaborative Knowledgebase. Communications of the ACM. 57, 78–85 (2014).
2. Voß, J., Bausch, S., Bogner, J., Schmitt, J., Berkelmann, V., Ludemann, F., Löffel, O., Kitroschat, J., Bartoshevska, M., Seljuzki, K.: Normdaten in Wikidata. Lulu. com (2014).
3. Neubert, J.: Economists in Wikidata: Opportunities of Authority Linking, http://zbw.eu/labs/en/economists-in-wikidata-opportunities-of-authority-linking, (2017).
4. Kempf, A.O., Neubert, J.: The Role of Thesauri in an Open Web: A Case Study of the STW Thesaurus for Economics. Knowledge Organization. 43, (2016).
5. Borst, T., Neubert, J.: Case Study: Publishing STW Thesaurus for Economics as Linked Open Data. W3C Semantic Web Use Cases and Case Studies. (2009).