

Classification of Knowledge Organization Systems with Wikidata

Jakob Voß

Verbundzentrale des GBV (VZG), Göttingen, Germany
`jakob.voss@gbv.de`

Abstract. This paper presents a crowd-sourced classification of knowledge organization systems based on open knowledge base Wikidata. The focus is less on the current result in its rather preliminary form but on the environment and process of categorization in Wikidata and the extraction of KOS from the collaborative database. Benefits and disadvantages are summarized and discussed for application to knowledge organization of other subject areas with Wikidata.

Keywords: Knowledge Organization Systems · Wikidata

1 Classification of Knowledge Organization Systems

Since introduction of the term knowledge organization system (KOS), several attempts have been made to classify KOSs by types such as glossaries, thesauri, classifications, and ontologies [6, 4, 11, 2, 10]. The set of KOS types and each of their extent varies depending on both domain or context of application (for instance different academic fields) and criteria of classification (for instance classification by purpose or structure). In many cases, KOS types are arranged according to the degree of controls, ranging from simple term lists to complex ontologies.

The most elaborated classification of KOS types in the field of knowledge organization is the the NKOS KOS Types Vocabulary [3]. It was developed by the DCMi NKOS Task Group with definitions mainly based on the KOS Taxonomy by Zeng and Hodge [17] and on ISO 25964 [7]. The KOS types vocabulary differentiates between 14 types of KOS (categorization scheme, classification scheme, dictionary, gazetteer, glossary, list, name authority list, ontology, semantic network, subject heading scheme, synonym ring, taxonomy, terminology, and thesaurus). One of the rare applications of these KOS types is their use in the Basel Register of Thesauri, Ontologies & Classifications (BARTOC)¹ where more than 1.800 KOSs have been classified so far [8].

¹ <http://bartoc.org/>

2 Wikidata

Wikidata² is the most recent project of the Wikimedia Foundation. In short, it is a collaboratively edited, free knowledge database that can be read and edited by both humans and machines. A good overview is given by two of Wikidata’s main creators Vrandečić and Krötzsch [16]. The primary goals of the project are:

1. Centralize links between Wikipedia language editions and other Wikimedia project sites. For instance all Wikipedia articles about “encyclopedia” (in any language) are linked to one Wikidata *item* with identifier Q5292. These so called *sitelinks* and other data about the concept known as “encyclopedia” can be looked up at <https://www.wikidata.org/wiki/Q5292>.³
2. Centralize Infoboxes. More and more manually edited infoboxes (tables with basic, factual information about a topic) are being extended to use Wikidata as their database backend, so the displayed information will be the same in all Wikipedia editions.
3. Provide an interface for rich queries. The content of Wikidata can be queried via a public SPARQL interface at <https://query.wikidata.org/> (see figure 1 for an example). Query results are planned to be integrated into Wikipedia and other projects as lists, tables, maps and other forms.

The data model of Wikidata is neither relational nor based on RDF (although mappings to RDF exists) but it reflects the strategy of Wikidata to store *statements* instead of facts. Each statement should be sourced by *references* and contradicting statements are not forbidden on purpose. Statements can further be controlled by *qualifiers*, such as domain and date of validity, eventually supporting *n*-ary relations between Wikidata items. The Wikidata ontology consists of Wikidata properties, which are defined by community consensus. For instance P571 identifies the property “inception” to state the date when something was created or founded. Labels and scope notes can be edited independently from statements with support of synonyms and homonyms.

3 Knowledge Organization Systems in Wikidata

To a large degree, Wikidata contains mappings to other KOSs: links between Wikidata items to VIAF and Geonames are among the most used Wikidata properties with around 100.000 Wikidata each⁴ and at least 1051 of 2490 Wikidata properties⁵ refer to external identifier systems of other databases.⁶ With

² <https://www.wikidata.org/>

³ In RDF this URL corresponds to the URI <http://www.wikidata.org/entity/Q5292>.

⁴ https://www.wikidata.org/wiki/Wikidata:Database_reports/List_of_properties

⁵ Measured at <https://www.wikidata.org/wiki/Special:ListProperties>, 2016-05-24

⁶ See also Voß et al. [15] for a German introduction to authority data in Wikidata.

uniquely identified records about virtually anything, Wikidata can also be seen as a KOS in its own right, so it is included in BARTOC.⁷ For the scope of this study, Wikidata items about KOS types and KOS instances are of special interest. KOS instances are Wikidata items linked to KOS type items with property “instance of” (P31).⁸ For instance the Dewey Decimal Classification (Q48460) is an instance of both a “library classification” (Q48473) and a “universal classification scheme” (Q24243801). Both are connected to the item “classification scheme” (Q5962346) with property “subclass of” (P279).

Despite their obvious use for knowledge representation and classification, the subclass and instance properties have no special role in Wikidata. Instead they can freely be used to connect any Wikidata items, only constrained by human intervention of other editors and, hopefully, community consensus in common sense. The lack of stricter rules on use of subclass and instance properties in Wikidata has led to criticism among researchers that try to use it as a formal ontology [1, 12]. Nevertheless, Wikidata can successfully be queried for a hierarchical list of transitive subclasses of “knowledge organization system” (Q6423319) and additional numbers for each class (figure 1).

```
SELECT ?item ?itemLabel ?broader ?parents ?size ?sites {
  {
    # number of additional superclasses
    SELECT ?item (count(distinct ?parent)-1 as ?parents) {
      ?item wdt:P279* wd:Q6423319
      OPTIONAL { ?item wdt:P279 ?parent }
    } GROUP BY ?item
  } {
    # number of instances
    SELECT ?item (count(distinct ?element) as ?size) {
      ?item wdt:P279* wd:Q6423319
      OPTIONAL { ?element wdt:P31 ?item }
    } GROUP BY ?item
  } {
    # number of sitelinks
    SELECT ?item (count(distinct ?site) as ?sites) {
      ?item wdt:P279* wd:Q6423319
      OPTIONAL { ?site schema:about ?item }
    } GROUP BY ?item
  }
  OPTIONAL { ?item wdt:P279 ?broader }
  SERVICE wikibase:label { bd:serviceParam wikibase:language "en" }
}
```

Fig. 1. SPARQL query to extract KOS classification data from Wikidata

⁷ <http://bartoc.org/en/node/1940>

4 A KOS classification extracted from Wikidata

The following classification of KOS types was extracted from Wikidata via SPARQL (figure 1). The command line tool ‘wdtaxonomy’, created especially to extract and display taxonomies, proved to be hugely useful for analysis [14]. The outcome, at the time of its creation at August 23rd 2016, contains 189 Wikidata items of KOS types, grouped in a multi-hierarchy. Classes within one level are sorted by their Wikidata identifier, reflecting the relative time when they were added to the database. Items with multiple superclasses within the KOS type hierarchy are shown in italics at the their second occurrence, for instance a ‘*plant taxonomy*’ is both a ‘biological classification’ and a ‘taxonomy’. Superclasses from other parts of Wikidata are indicated with upwards pointing arrows (↑), for instance ‘datamodel’ is also subclass of ‘model’ and ‘data set’. The numbers right to each class label indicate the current number of instances in Wikidata (if there are any), and after a small plus sign the number of Wikipedia articles or other Wikimedia projects sites linked with the entry (sitelinks). For instance three Wikidata items are marked as instance of ‘semantic network’ and twenty Wikipedia editions include an article about semantic networks (3+20).

knowledge organization system ↑↑↑	2+3	.. defining vocabulary	↑	+1
· encyclopedia ↑	299+183	.. multi-field dictionary		+1
· · internet encyclopedia ↑	29+27	· authority control		36+42
· · · encyclopedic dictionary	42+7	· mind map ↑		+46
· · · · biographical encyclopedia ↑	59+8	· ontology		26+31
· · universal encyclopedia	1+2	· · metamodel		6+11
· · · hypertext encyclopedia	6+1	· · upper ontology		5+2
· dictionary ↑	182+126	· · process ontology		+1
· · lexicon ↑	4+38	· · soft ontology		+1
· · · lexicographic thesaurus	7+48	· · weak ontology		+1
· · · orthographic dictionary	+7	· concept map ↑		+26
· · · etymological dictionary	1+13	· semantic network ↑		3+20
· · · glossary ↑	13+35	· data model ↑↑		5+17
· · · visual dictionary	1+6	· · database model		3+11
· · · explanatory dictionary	4+6	· · data dictionary ↑		+15
· · · · explanatory combinatorial dictionary	+2	· · data schema		+3
· · · · monolingual learner’s dictionary	+2	· · · XML Schema		1+28
· · <i>encyclopedic dictionary</i>		· · · database schema		+13
· · reverse dictionary	+9	· · · database normalisation		3+32
· · machine-readable dictionary	+10	· · · star schema		+12
· · · online dictionary	9+6	· · · logical data model		+2
· · · · Wiktionary language edition ↑↑	7	· · information model ↑		+8
· · · rime dictionary	4+7	· · · <i>database schema</i>		
· · rhyming dictionary	+11	· · domain model		+6
· · conceptual dictionary	+6	· · conceptual schema		+6
· · bilingual dictionary	2+13	· · canonical model		+1
· · slang dictionary	2+3	· · GIS data model		+1
· · idiom dictionary	+5	· · · ArcGIS data model		+1
· · Anagram dictionary	+1	· · generic data model		+1
· · author dictionary	+1	· · semantic data model		+1
· · language for specific purposes dictionary	+1	· · standard data model		+2
· · medical dictionary	1+2	· · White pages schema		+1
· · phonetic dictionary	+1	· · <i>logical data model</i>		
· · picture dictionary	+2	· conceptual graph ↑		+6
· · single-field dictionary	+1	· controlled vocabulary		10+8
· · specialized dictionary	+2	· · thesaurus		11+5
· · · sub-field dictionary	+1	· synonym ring		+5

· numbering scheme	+3	· · · · zoological nomenclature	2+2
· catalog ↑↑	124+11	· · library classification	26+22
· · <i>authority control</i>		· · decimal classification	6+1
· · directory	4+43	· · faceted classification	1+11
· · telephone directory ↑	1+28	· · universal classification scheme	12
· · civil registry ↑	2+25	· · specialized classification scheme	10
· · · birth registry ↑	1	· · · biological classification	10+112
· · web directory ↑	11+30	· · · <i>plant taxonomy</i>	
· · address book ↑	1+11	· · · syntaxonomy	+9
· · astronomical catalog	26+26	· · · taxonomy of Chordata	+1
· · · star catalogue	19+31	· · · · taxonomic classification of Passeriformes +1	
· · · catalogue of galaxies	2+1	· · · · taxonomy of birds	1+2
· · land registration	4+25	· · · · <i>taxonomic classification of Passeriformes</i>	
· · · Otabumi	+4	· · · linguistic typology	3+42
· · exhibition catalogue	7+5	· · climate classification	1+14
· · inventory	2+36	· · · genetic climate classification	1+1
· · · heritage register	75+1	· · · effective climate classification	2+1
· · · buildings at risk register ↑	+1	· · classification of wine	1+7
· · business directory	+4	· · chemical classification	3+8
· · gazetteer ↑	3+18	· · video game content rating system ↑	8+17
· · urbarium	2+18	· · classification in sports	88+3
· · business register	1+1	· · · age category in soccer	+1
· · · company register	9+13	· · · olympic sailing class ↑	4+6
· · bibliography	610+6	· · · equestrian sports classification	+1
· · · discography	5543+35	· · · weight class	3+7
· · · library catalog ↑	19+24	· · · · wrestling weight classes	+2
· · · union catalog	15+7	· · · · lightweight class	
· · catalogue raisonné ↑	48+10	· · · age category in athletics	5
· · filmography	799+10	· · · competition class	84
· · · videography	19	· · · · glider competition class	9+3
· · subject bibliography	+1	· · · · group	3+4
· · playlist ↑	+14	· · · · rally group	1
· · national bibliography	2+2	· · · musical instrument classification	1+11
· · regional bibliography	10+1	· · medical classification	8+3
· · bibliographic index	+4	· · corporate taxonomy	+1
· · · bibliographic database ↑	41+13	· · economic taxonomy	2+2
· · annotated bibliography	+1	· · · job classification system	7+1
· · catalogue of classical compositions	23+1	· · · patent classification	4+2
· · metabibliography	+1	· · · industry classification	18+3
· · · list of music references	1+1	· · · product classification	+1
· · medal table	132+2	· · military taxonomy	+2
· · Olympic medal table	55+3	· · · military casualty classification	6
· · auction catalog	+3	· · safety taxonomy	+1
· · product catalog	+4	· · astronomy classification	1
· · records catalog	+1	· · · stellar classification	25+59
· · thematic catalog	9+7	· · · asteroid classification	3
· · nuancier	1+2	· · · · asteroid family	82+25
· · · RAL color system ↑	+16	· · · · asteroid spectral type	19+22
· · music catalog	2+1	· · · · minor-planet group	1+7
· · population register ↑	+1	· terminology	3+1
· · recording catalog	+1	· · nomenclature	7+25
· · index ↑	4	· · biological nomenclature	12+12
· · terminology registry	2	· · · phylogenetic nomenclature	+10
· · metadata registry	+4	· · · <i>botanical nomenclature</i>	
· conceptual model ↑	3+14	· · · <i>zoological nomenclature</i>	
· · systems architecture	1+9	· · · nosography	1+3
· · hierarchical temporal memory	+5	· · chemical nomenclature	3+27
· · conceptual model (computer science)	+1	· · · IUPAC chemical nomenclature	3+29
· · <i>domain model</i>			
· · <i>conceptual schema</i>			
· · <i>semantic data model</i>			
· classification scheme	61+23		
· · taxonomy	9+63		
· · · numerical taxonomy	+5		
· · · taxonomy	6+40		
· · · plant taxonomy	+21		
· · · · Taxonomy of Narcissus	+1		
· · · botanical nomenclature	1+9		
		Summary	
		number of classes:	189 (100%)
		with instance:	116 (61%)
		with sitelink:	175 (92%)

5 Discussion

The classification of KOS types extracted from Wikidata is detailed but obviously sketchy in its current form. The system was even more incomplete before large parts of it had been edited by the author, mainly to adjust or add missing English labels and items without any instance or subclass statement. A couple of items were also moved or merged after notification and discussion with the Wikidata community.⁸ Instead of criticizing usage limitations of Wikidata class hierarchies such as Spitz et al. [12], or suggesting methods to better spot classification inconsistencies such as Brasileiro et al. [1], peculiarities of knowledge organization systems based on Wikidata shall be highlighted in the following.

First of all, classes and instances are more or less given by existence or non-existence of Wikidata items with sitelinks: only 14 of 189 classes in the classification above don't have at least one corresponding Wikipedia article. To some degree new Wikidata items can also be added without sitelink, but this is controversial at least for abstract concepts which have no obvious unique identification.⁹ KOS types therefore usually require some Wikipedia article before inclusion. The strong connection to Wikipedia also forbids removal, reinterpretation or merging of concepts that don't easily fit into a classification, unless one engages in editing Wikipedias. Anyway, KOS derived from Wikidata are build by a bottom-up approach from general encyclopedic concepts. As scope and definition of a Wikidata item vary between Wikipedia language editions, the concepts are fuzzy to some degree. For instance at the moment there are two items for classification as process and classification as result but only one item for both metamodeling and metamodel.¹⁰ One should not try to solve all of these cases in structured data, as Graham [5] warned about application of Wikidata: "We just need to ensure that we aren't seduced into codifying, categorizing, and structuring in cases when we should be describing the inherent messiness of a situation." It has already been shown that the category system of Wikipedia is more a thesaurus than a classification [13]. KOS extraction from Wikidata may also result in less strict hierarchies without strong formal logic. Participation in Wikidata differs from collaborative ontology engineering [9]. Although the database is often referred to as knowledge base, its practical purpose in many ways is more knowledge organization than knowledge representation.

The number of sitelinks can be used as indicator how established or widely known a concept is. The number of instances more depends on which KOS instances are relevant for inclusion in Wikimedia projects and have been classified in Wikidata at all. Despite this bias, instances are very helpful to judge the application of a concept for classification. Both new Wikidata instances and new sitelinks are added to KOS types virtually every day. The benefit of this dynamic growth is better coverage of multiple views and domains, so most KOS extracted

⁸ The question about 'taxonomy' vs. 'biological classification' could not be resolved.

⁹ See <https://www.wikidata.org/wiki/Wikidata:Notability>

¹⁰ "classifying" (Q13582682), "classification" (Q5962346), and "metamodel" (Q1925081)

from Wikidata include aspects of a universal classification and less suffer from opinionated knowledge organization. The downside of this crowd-sourcing is lack of a final consensus. Therefore extraction of KOS from Wikidata is an iterative process that requires continuing review and contribution to Wikidata (figure 2).

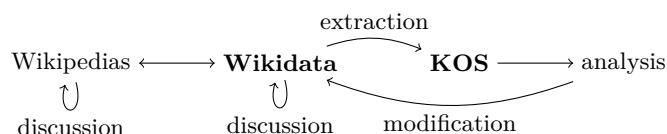


Fig. 2. Iterative process of KOS extraction from Wikidata

6 Summary and outlook

This paper introduced the extraction of knowledge organization systems from Wikidata exemplified by a classification of KOS types. With 181 classes the result is probably the most detailed classification of KOS types so far. Before further evaluation and cleanup, however, it can only serve as starting point.

The strong grounding in Wikipedia is both benefit and challenge. The system is more build bottom-up from instances instead of top-down from theoretical properties like existing classifications and typologies. Further improvement of the particular classification requires involvement of both the Wikidata community and domain experts in knowledge organization. For instance one could express different characteristic of division as Wikidata qualifiers.¹¹ It is also not sure yet whether the hierarchical structures extracted from Wikidata can better be expressed by other types of KOS such as multi-level models [1] or thesauri [13].

Further work on the KOS classification includes alignment with the NKOS KOS Types Vocabulary [3]. KOS instances have been classified in both Wikidata and BARTOC [8], connected via Wikidata property “BARTOC ID” (P2689). More elaborated KOS extraction should also incorporate Wikidata qualifiers. Existing tools such as SQID Wikidata Browser¹² could further be extended to better support management of KOSs extracted from Wikidata.

References

- [1] Freddy Brasileiro et al. “Applying a Multi Level Modeling Theory to Assess Taxonomic Hierarchies in Wikidata”. In: *Proceedings of WWW 2016 Companion*. 2016. DOI: [10.1145/2872518.2891117](https://doi.org/10.1145/2872518.2891117).

¹¹ This has already been started for domain-specific subtypes of classification schemes.

¹² <http://tools.wmflabs.org/sqid/>

- [2] Eva Bratková and Helena Kučerová. “Knowledge Organization Systems and Their Typology”. In: *Revue of Librarianship* 25.2 (2014), pp. 1–25. URL: http://full.nkp.cz/nkk/knihovna142_suppl/1402sup01.htm.
- [3] Dublin Core Metadata Initiative. NKOS Task Group. *KOS Types Vocabulary*. Oct. 2015. URL: http://wiki.dublincore.org/index.php/NKOS%5C_Vocabularies.
- [4] Alan Gilchrist. “Thesauri, taxonomies and ontologies – an etymological note”. In: *Journal of documentation* 1 (2003), pp. 7–18. DOI: [10.1108/00220410310457984](https://doi.org/10.1108/00220410310457984).
- [5] Mark Graham. “The Problem With Wikidata”. In: *The Atlantic* (2012). URL: <http://www.theatlantic.com/technology/archive/2012/04/the-problem-with-wikidata/255564/>.
- [6] Gail Hodge. *Systems of knowledge organization for digital libraries: beyond traditional authority files*. Digital Library Federation, 2000.
- [7] *ISO 25964 Information and documentation – Thesauri and interoperability with other vocabularies*. ISO standard. 2013.
- [8] Andreas Ledl and Jakob Voß. “Describing Knowledge Organization Systems in BARTOC and JSKOS”. In: *Proceedings of International Conference on Terminology and Knowledge Engineering*. 2016, pp. 168–178. URL: <http://hdl.handle.net/10760/29366>.
- [9] Claudia Müller-Birn et al. “Peer-production system or collaborative ontology engineering effort: what is Wikidata?” In: *OpenSym*. Ed. by Dirk Riehle. ACM, 2015, 20:1–20:10. DOI: [10.1145/2788993.2789836](https://doi.org/10.1145/2788993.2789836).
- [10] Vreda Pieterse and Derrick G. Kourie. “Lists, Taxonomies, Lattices, Thesauri and Ontologies: Paving a Pathway Through a Terminological Jungle”. In: *Knowledge Organization* 41.3 (2014), pp. 217–229.
- [11] Renato Rocha Souza, Douglas Tudhope, and Maurício Barcellos Almeida. *The KOS spectra: a tentative typology of Knowledge Organization Systems*. 2010. URL: <http://mba.eci.ufmg.br/downloads/ISKO%5C%20Rome%5C%202010%5C%20submitted.pdf>.
- [12] Andreas Spitz et al. “State of the Union: A Data Consumer’s Perspective on Wikidata and Its Properties for the Classification and Resolution of Entities”. In: *Proceedings of ICWSM 2016*. 2016.
- [13] Jakob Voß. *Collaborative thesaurus tagging the Wikipedia way*. Tech. rep. 2006. URL: <http://arxiv.org/abs/cs/0604036>.
- [14] Jakob Voß. *wikidata-taxonomy 0.2.7*. 2016. DOI: [10.5281/zenodo.60708](https://doi.org/10.5281/zenodo.60708). URL: <https://github.com/nichtich/wikidata-taxonomy>.
- [15] Jakob Voß et al. *Normdaten in Wikidata*. 2014. ISBN: 978-1-291-85658-3. URL: <https://hshdb.github.io/normdaten-in-wikidata/>.
- [16] Denny Vrandečić and Markus Krötzsch. “Wikidata: A Free Collaborative Knowledgebase”. In: *Communications of the ACM* 57.10 (2014), pp. 78–85. DOI: [10.1145/2629489](https://doi.org/10.1145/2629489).
- [17] Marcia Lei Zeng and Gail Hodge. *Taxonomy of Knowledge Organization Sources/Systems*. 2000. URL: http://nkos.slis.kent.edu/KOS_taxonomy.htm.