Kempf, Andreas Oskar

Conference Paper

The Need to Interoperate: Structural Comparison of and Methodological Guidance on Mapping Discipline-Specific Subject Authority Data to Wikidata

Suggested Citation: Kempf, Andreas Oskar (2018) : The Need to Interoperate: Structural Comparison of and Methodological Guidance on Mapping Discipline-Specific Subject Authority Data to Wikidata, In: ISKO '18 – Proceedings of the 15th International ISKO Conference, 9-11 July 2018, Porto, Portugal, ISBN 978-3-95650-420-4, Ergon-Verlag, Baden-Baden, pp. 645-652, https://doi.org/10.5771/9783956504211-644

This Version is available at:
http://hdl.handle.net/11108/366

Andreas Oskar Kempf

The Need to Interoperate: Structural Comparison of and Methodological Guidance on Mapping Discipline-Specific Subject Authority Data to Wikidata

Abstract

The linking paradigm of Linked Data (LD), the RDF-based information architecture of the WWW, and RDA with its underlying Entity-Relationship-Model (ERM) demand an increased entity-based semantic interoperability of subject authority data. Wikidata (WD), the knowledge base and sister project of Wikipedia (WP) seems to be a promising environment for joint efforts to bring authority data into the semantic web and to undertake this task. However, it seems necessary to bear in mind that with regard to subject authority data WD interlinks different forms of knowledge organization (KO) that are characterized not only by different functional purposes but also by distinct structural principles of modeling concepts clearly linked to the diverse information systems (IS) from which they derive. On these grounds this paper provides methodological guidance on how to approach a mapping process between topical thesaurus concepts and WD items.

1. Introduction

As the web has expanded in syntax and scope, it has evolved from what was originally a web of documents into an all-encompassing medium (see Gradmann, 2013) and has become *the* place to search for information. At the same time, the so-called 'Web of Things' enables the representation of everything that could become part of an RDF statement. This has enormously extended the scope of representation in the web and "finding and selecting moves steadily toward connecting and relating every record with every other record in an all-embracing web" (Buckland, 2017, 176).

In the spirit of the LD paradigm new web environments evolve to interlink different types of entities. While they seem suitable to interlink individual concepts, such as author names (Neubert, 2017), with regard to subject authority data they run the risk of juxtaposing entities with different purposes of representation and frames of meaning.

By referring to topical concepts of the STW Thesaurus for Economics as mapping candidates to WD items, this paper demonstrates how one might approach a mapping process between topical thesaurus concepts and WD. Section 2 states the research problem by hinting at the interconnectedness between functional and structural principles of KO entities and the respective IS to which they are tailored. Section 3 presents core functional requirements and structures of documentation languages (DL) in general, before referring to the concrete example of the STW. Section 4 refers to the function and structure of concept modeling practices in the WD backbone WP. Chapter 5 contrasts the concept modeling principles of the two forms of KO by referring to their corresponding IS purposes and parameters. On this basis, section 6 takes stock of the ways in which one might select potential thesaurus concepts for a systematic mapping process between the two by referring to the latest ISO standard. The conclusion resumes the role of WD as a linking infrastructure in this context.

2. Problem statement

The library field disposes of a long standing expertise in the development of subject authority data. For a long time, the application of these vocabularies was on a library collection scale. As the web becomes the place to search for information, the scope of thesaurus-indexed collections is getting more and more relative. In recent years, controlled vocabularies were increasingly published on the web to extend their coverage beyond mere bibliographic metadata, but their inherent structural characteristics which derived from the IS they were initially developed for remained. However, LD and the RDF-based architecture of the semantic web demand an increased semantic interoperability of subject authority data to "get rid of" vocabularies that are "deeply rooted in the focus on information containers" (Gradmann, 2013, 255).

Linking environments promise to bring together the same entities of different authority files, this way serving as a linking hub. However, while they seem suitable for individual authority data, a closer look at matching subject authority data in a crosswalk service like WD reveals clear differences in modeling concepts between the different forms of KO they relate to. By taking the example of the STW, we provide guidance on best practice on how to approach a mapping process between subject-specific authority data taken from a thesaurus and topical items from WD, paying special attention to their distinct functional and structural principles.

3. Controlled vocabularies for indexing and search

3.1 Structural characteristics and representation functions

According to the latest ISO standard, a thesaurus is a formally structured vocabulary in which "concepts are represented by terms, organized so that relationships between concepts are made explicit, and preferred terms are accompanied by lead-in entries for synonyms or quasi-synonyms" (ISO, 2013, 14). Conceptual and terminological control to "guide both the indexer and the searcher to select the same preferred term or combination of preferred terms to represent a given subject" (ibid.) are thus two key and interlinked principles underlying thesaurus design.

This concept navigation function is always related to specific IS parameters. According to Wersig (1985), these parameters (further expanded in section 3.2 with reference to the STW) include the degree of stability of the domain-specific terminology and the terminological richness (i.e. the amount of descriptors and non-descriptors). Additionally, he emphasizes different representation functions in the modeling process of a DL element – resp. a thesaurus concept represented by a descriptor – and distinguishes between four different levels. On a first level it could represent a class of concepts as for example the STW descriptor "Oil and gas services"[1]. On a second level, the DL element could represent a particular class of

[1] http://zbw.eu/stw/descriptor/19047-0 Accessed 30 January 2018.

objects from a set of object classes formed by the class of concepts. Which of these objects are specifically meant, in turn, does not derive from the verbal description of the class itself, but only from the context of the element. For example the STW descriptor "Institutional infrastructure"[2], could represent legal as well as culture specific conditions of social and economic life. A third level represents the entire system-specific context, e.g. what kind of documents need to be indexed and according to which indexing rules. This could lead to a system-specific narrowing or broadening of the meaning of the DL element. On a fourth level, in practice the DL element represents the amount of concrete documents it was assigned to (ibid., 74pp.).

Moreover, also referring to Wersig (ibid., 77p.), when conducting conceptual control, two aspects are of particular practical importance with regard to the semantic frame of a concept: first, the implicit and explicit meaning of a term, second, the natural language meaning and the DL meaning of a term. With regard to the first aspect: the implicit meaning(s) of a term is/are learned in the socialization process and come(s) out of its/their usage. (The) explicit meaning(s) of a term is/are contained in authoritative lexical, resp. encyclopedic works. In the case of a DL they could be made explicit in scope notes. Both meanings do not need to be identical to the current use of a term for indexing. With regard to the second aspect, DL recruit or build their vocabulary from the vocabulary of currently used natural language, incorporating its communicative meaning (i.e., the set of all uses of the term realized in current or past communication situations) and its lexical meaning usually being a subset of it. Communicative/lexical meanings are not necessarily always fully accepted in the DL or taken over (e.g. polysemy control). Equally conceivable is the reverse case, in which a meaning could be assigned to a term, although it does not correspond to its lexical or communicative meaning (e.g. quasi-synonym definitions).[3]

In sum, according to Wersig (ibid., 79) conceptual control during thesaurus development and maintenance should take into account the following aspects: identification of the communicative meaning, whereby the lexical meaning cannot always be adopted into the DL; it often needs to be supplemented with meanings of current use in the relevant domain (conceptual analysis), elimination of (components of) meanings that are not relevant to the DL, for example by exclusions (conceptual adjustment); extension of further (components of) meanings that are not communicatively defined, but are to be subsumed here for purposes of the DL and made explicit (conceptual extension).[4]

Integration of new concepts or their elimination could affect the definitions of meaning of other conceptual units, therefore monitoring of meaning as well as of

[2] http://zbw.eu/stw/descriptor/12009-1 Accessed 30 January 2018. See also the corresponding scope note.

[3] The fact that a thesaurus could be used in more than one IS (ibid.) should be excluded here.

[4] Wersig (ibid., 79) also mentions system-related specifications when using a cross-system thesaurus.

assignment needs to be done. Moreover, the use of a term in the context of common language and discipline-specific terminology is not static (see ISO, 2011, 96).

Structures for conceptual order in a thesaurus which give orientation and assist discovery are its classical semantic relations and an existing category system.

3.2 The STW Thesaurus for Economics

The STW, whose nearly 6,000 descriptors and about 20,000 non-descriptors cover all economics-related and on a broader level the most important related subject areas, comprises different types of concepts. In its geographical subthesaurus, comparable to a name authority list, it mainly contains individual concepts, i.e. country descriptors, while the other six subthesauri include general concepts, characterized by different levels of abstractedness, culminating in the subthesaurus 'general descriptors'.[5]

With regard to its IS context, merging four different DL, ranging from a simple keyword list to a fully-fledged thesaurus and reflecting specific collection focuses, marks the birth of the STW, which today is used in the IS of the resp. institutions originally involved in its development, as well as in other IS. The large amount of source terminology has further promoted the terminological richness already present in this social science domain. This is reflected in extensive equivalence classes in some of its branches. Regarding key principles for concept modeling, both methods, precombination and post-coordination, are used. The overall thesaurus structure is polyhierarchical.

Right from the start, the scope of the STW, now available in German and in English has been constantly expanded. Similar to the situation in other disciplines, scientific discourses in many sub-disciplines of the economic sciences became more and more international. The ZBW – Leibniz Information Centre for Economics, which has taken over STW development, consequently directed its information services towards an international, English-speaking scientific community (Kempf/Neubert, 2016).

Comparable with other social science thesauri, the terminology is rather fuzzy and the degree of terminological stability is rather low. Some terms could stand for different theoretical models which cannot be further differentiated. Moreover, the concepts are often deeply embedded into scientific discussion processes. In addition, far-reaching discourse shifts and new subject fields can be regularly observed. In sum, developments in this domain concern concepts, terminology semantics, and the structure of the vocabulary. For this reason, after more than 15 years of permanent, though rather isolated ad-hoc updating, the STW has been completely revised, descriptor by descriptor, over a period of several years. In the end, more than 750 thesaurus concepts have been added, and nearly 1,100 (out of ca. 6,000) concepts have been eliminated. In addition, entry terms have been adapted and the systematic structure has been improved (Gastmeyer et al., 2016). Detailed change reports, which

[5] http://zbw.eu/stw Accessed 30 January 2018.

make changes of the STW traceable, are provided using the published SKOS files of its latest versions (Neubert, 2015).

Enhanced interoperability has already been achieved thanks to vocabulary mappings established in the past, allowing STW content descriptions to be translated into subject information using other vocabularies (and vice versa). An example is the continuous intellectual mapping between STW and subject authority data of the German Integrated Authority File (GND) (Kempf/Neubert 2016).

4. Wikidata & Wikipedia – information system parameters and concept structure

Launched by the Wikimedia foundation in the year 2012, WD serves as a shared knowledge base to provide structured data for the nearly 300 different language versions of WP. Hence, the collaboratively edited free online encyclopedia WP being the starting point for WD, its concept modeling refers to an encyclopedic design. In the first place, its entities serve as reference in the search for definitions and factual information of encyclopedic nature. Its structural form of KO is for learning; for user guidance connections between entries are established through the use of categories.

Taking the open production model of WP and its fairly vague notability criteria into account, the level of specificity of an entry term can vary enormously. Which conceptual unit receives article status is up to the participating contributors. The conceptual scope of theme or topic units, resp. their entry lemmas, which decisively shape the structure of an encyclopedia, can be broad or narrow. Linked to this, the depth of explanation can also vary greatly, ranging from rough overviews to explanations which place their subject into larger context. Related to this and depending on the content, the lifecycle and degree of dynamics of an entry varies, too. Although its claim is universal and it contains all types of information, there are big gaps in the list of lemmas which are edited by the Wikipedians in a list of article requests. Since it is not designed for indexing, vocabulary control is done rather rudimentarily in the form of redirects. Finally, similar to other universal encyclopedias, it recruits its vocabulary from more or less currently used natural language capturing consensual meanings of a term that have become relatively accepted.

WD, as a central storage repository, for the most part tries to manage WP's data on a global scale (Vrandečić/Krötzsch, 2014, 78) this way bringing encyclopedic and DL forms of KO together. Authority data also used in bibliographic databases are assigned to various types of entities which are part of WP articles. WD serves as a linking hub, allowing for relations of equivalence, and, only recently, for those of closeness and hierarchy between the WD item identified by an abstract identifier and the item in an external database identified for its part by a so called external identifier. Special sitelinks connect an item to corresponding content on client wikis, such as WP. By mapping external entities, WD connects to more than 1,500 different sources of authority information. Moreover, if a source of authority information which is registered as external property has already been mapped to a source linked to WD, this prior mapping can be exploited for creating a preliminary mapping of the newly connected source to WD.

By providing structured data for the different WP language versions, variations of meaning between different languages are of particular relevance to WD. As other languages often come along with cultural specificities embedded in particular political, economic and social structures, a concept and the meaning of a term used to express it are language specific and cannot always be translated one-to-one into another language. These cultural foundations of KO that underline the fact that the web is not only a technical but also a social and cultural project (see Garciá-Marco, 2016) also need to be properly addressed.

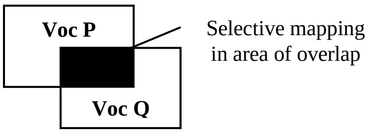
5. Comparison

Clearly common to both forms of KO is that they stand for systematically organized forms of KO and deal with natural language. However, with regard to their most specific aims there are considerable differences between both. While WD items are grounded in an encyclopedic information architecture providing targeted access to factual information helping to understand the meaning of a term, the prime motivation of a DL is knowledge representation and to enable information retrieval. Due to their relatedness to different kinds of IS with diverse inherent scopes, both forms of KO are characterized by different construction principles for concept modeling. The frames of concepts can vary enormously due to differences between lexical meaning in natural language on the one hand and their DL meaning on the other hand. Moreover, thesaurus concepts are characterized by representation functions on various levels.

6. Recommended mapping model and approach

Taking the differences in modeling topical concepts between both forms of KO into account, we would like to argue for a mapping model as depicted in ISO 25964-2 (2013, 78pp.) for mappings between thesaurus and terminology: i.e. a selective mapping (see Fig. 1).

Figure 1: Selective mapping (ISO, 2013, 19)



At the working level, however, "establishing concept mappings (…) should follow the same general methodology and practices as between two thesauri" (ibid., 81). Against this background, the extension of mapping relation types in WD by accepting 'close match', 'broader match' and 'narrower match' as additional properties (Neubert, 2017) is clearly approved. Taking the stated differences in concept modeling into account, we would like to present criteria which could provide guidance on approaching a mapping process between topical thesaurus concepts and WD items and serve as a starting point for identifying suitable concept candidates of a thesaurus by paying special attention to scope, representation function and the semantic stableness of a thesaurus concept.

- As depicted above, the detection of subject-specific authority data from other DL in WD, which had already been mapped to the respective thesaurus, could be useful to identify promising candidates to be mapped.[6][7]
- The allocation of a concept to a certain subthesaurus or, more general, to a specific top term could hint at the degree of abstractedness resp. concreteness of a concept.
- With regard to the degree of specificity of a concept, its location within different subthesauri or different branches of a single subthesaurus could provide information to what extent several meaning components could be included in a concept. Also referring to the degree of specificity of a concept, the problem of hypernyms and hyponyms could be dealt with by identifying underspecified synonyms within an equivalence class.
- Referring to the stableness of subject-specific vocabulary, it was mentioned that huge differences exist between different domains. With regard to the extent of changes within one thesaurus, dynamics in concept modeling in the past could be approached on various levels: on the subject category level (e.g. adding or merging two subject categories into one), the concept level (e.g. adding a new concept, represented by a preferred term), and on the terminological level (e.g. reversing the preference between a preferred and a non-preferred term). Hereby, it could be possible to identify particularly stable or unstable areas below the subthesaurus level within a thesaurus. The SKOS publication of the thesaurus could be used to track these changes (Neubert, 2015).

The procedure depicted, meant as an initial stocktaking exercise, could help to select a core set of fairly stable topical thesaurus concepts whose degree of overlap between documentational and lexical meaning is rather high. The topical thesaurus concepts receive a fingerprint, so to speak, which expresses their suitability for mapping. Armed with this corpus of concepts, the editor is then able to begin with a systematic mapping process. Special attention must still be paid to the different language versions. The mapping in WD could be given an additional predicate by assigning a preferred rank, compared to levels of determinations as used in other mapping contexts.

[6] In WD the interactive mapping tool Mix’n’match could be used for this purpose. http://meta.wikimedia.org/wiki/Mix%27n%27match Accessed 30 January 2018.

[7] Already being mapped to the subject headings of the GND it turned out that 2,034 of the 5,339 topical concepts of the STW are already transitively linked to Wikidata items via GND ID (Neubert, 2017, 11). The intellectual evaluation of a random sample, however, reveals differences in concept modeling between both forms of KO as depicted here.

7. Conclusion

To conclude: in a process which has already gained momentum, WD could serve as a core reference and linking infrastructure helping subject-specific authority data to break through the ‘container’ of closed collections and to “improve their relevance in the digital age” (Garciá-Marco, 2016, 195). With regard to thesaurus maintenance and development, WD could serve as a core reference hub for concept modeling, helping to reconcile collection and web-scale areas of reference and thus leading to a greater harmonization of subject authority data. Thesaurus editors and contributors could immediately benefit from WP’s definitions and encyclopedic information; and WD’s data storage could “assist editorial tasks such as proposing and assessing intralingual and multilingual, preferred terms and equivalences” (ibid.).

Employing WD as a central hub for mapping could lead to a stronger convergence of conceptual modeling within the different sources of authority information, which should be undertaken as a constant endeavor. By bringing authority data from various sources together, it could assist interoperability and vocabulary sharing, because it helps to scrutinize the degree of conceptual overlapping. In a step-by-step process this could foster a new paradigm of standardization already on the rise which favors compatibility instead of homogeneity.

References

Buckland, Michael. 2017. Information and Society. MIT Press, Cambridge, Mass.

Garciá-Marco, Francisco-Javier. 2016. Enhancing the Visibility and Relevance of Thesauri in the Web. KO, 43 (3), 193-202.

Gastmeyer, Manuela; Wannags, Max-Michael; Neubert, Joachim. 2016. Relaunch des STW. Information. Wissenschaft & Praxis, 67 (4), 217-240.

Gradmann, Stefan. 2014. From containers to content to context. Journal of Documentation, 70 (2), 241-260.

ISO 25964-1/-2. 2011, 2013. Information and documentation – Thesauri and interoperability with other vocabularies – Part 1, 2, Geneva: ISO.

Kempf, Andreas; Neubert, Joachim. 2016. The Role of Thesauri in an Open Web. Knowledge Organization, 43 (3), 160-173.

Neubert, Joachim. 2015. Leveraging SKOS to Trace the Overhaul of the STW. Proc. of the Int. Conference on Dublin Core and Metadata Applications, 170-180.

Neubert, Joachim. 2017. Wikidata as a linking hub for KO systems? 17th NKOS workshop at TPDL, Thessaloniki, Greece, http://ceur-ws.org/Vol-1937/paper2.pdf Accessed 30 January 2018.

Vrandečić, Denny; Krötzsch, Markus. 2014. Wikidata: A Free Collaborative Knowledgebase. Communications of the ACM, 57 (10), 78-84.

Wersig, Gernot. 1985^2. Thesaurus-Leitfaden. K. G. Saur: u.a. München.