Cleaning up a legacy thesaurus to make it fit for transformation into a Semantic Web KOS

Anna Kasprzik[0000−0002−1019−3606]

Technische Informationsbibliothek, Welfengarten 1B, 30167 Hannover, Germany
anna.kasprzik@tib.eu

Abstract. Legacy knowledge organization systems (KOS) such as thesauri for specific domains that have been created and maintained as print versions before being digitized typically suffer from a rather heterogeneous quality in terms of their structural consistence and interconnectivity. We take the domain thesaurus "Technik und Management" (TEMA) which has its origins in the assembly of six print thesauri, highlight some exemplary structural challenges resulting from this assembly and from various additions over the years since its creation, and make suggestions of how to align it with the principles of the Semantic Web via the use of corresponding standards such as SKOS, with a special focus on the existing and the potential relations between its concepts and terms. More specifically, we made an attempt to transform a subset of the thesaurus into an ontology, and then realized that we would have to improve the overall structure of the thesaurus first before we could proceed. During the process, we also took into account the concerns of the subject experts at WTI who are in charge of maintaining the thesaurus on a daily basis.

Keywords: knowledge organization systems · legacy thesauri · structural enhancement · semantic relations · Semantic Web standards · SKOS

1 Point of departure

1.1 The domain thesaurus "Technik und Management" (TEMA)

The domain thesaurus "Technik und Management" (technical and management topics; TEMA) started out as a product of the cooperative WTI-Frankfurt eG (formerly "Fachinformationszentrum (FIZ) Technik", founded based on an initiative of the German government), and since 2013 is developed further in collaboration with the Leibniz Information Centre for Science and Technology (TIB).

The thesaurus was created by manually joining five print thesauri (on topics such as mechanical, electrical, medical, materials and textile engineering, electronics, information technology), the contents of a sixth (naval architecture) was added later on. The first printed version with about 34 000 concepts and 80 000 terms was issued in 1998, the fourth and last with 34 900 concepts and 97 800 terms in 2003/4 when it was digitized. In 2018, the thesaurus comprises about 57 000 concepts and 197 000 terms, both German and English, and is one of

the largest for the topics it covers. The terminological material is curated and increased by the subject experts working at WTI on a day-to-day basis.

The TEMA thesaurus is maintained on the Averbis Terminology Platform (ATP) in a proprietary format and its contents can be exported as a .txt file (minus a few fields that are deemed relevant to the subject experts of WTI only). The ATP features a layered system of access rights (read; propose; edit proposals; edit, accept proposals, import/export, create subterminologies; admin) in order to be able to control the editing. By the clients of WTI, the thesaurus is mainly used in combination with the domain-specific document databases provided by WTI, as an indexing base for the associated search engine (TecFinder).

1.2 Project "Fachontologie Technik" (2013–2017)

The project "Fachontologie Technik" ("domain ontology for technical subjects") was started in 2013 as a joint project of TIB and WTI with the goal to enrich the TEMA thesaurus in various ways in order to take it to the next level of the digital development – for example, by transferring it to the ATP, by introducing English terms, thus making the thesaurus bilingual throughout, by evaluating term extraction tools that would help add more content, and by aligning it with concepts from the German authority file ("Gemeinsame Normdatei"; GND) [1]. An additional goal was to increase the interoperability of the TEMA thesaurus with other knowledge organization systems (KOS) by transforming it into a KOS that complies with the established Semantic Web standards, models and description languages such as RDF, RDF Schema, SKOS and/or OWL (depending on the use case, and consequently, the required level of formalization).

2 Steps towards a sandbox ontology for electric mobility

In line with the project goal formulated in the previous section, the author implemented a script (both in Perl and Python) that transforms the .txt output of the ATP into a hybrid OWL/SKOS ontology (with SKOS focussing on the classical thesaurus relations between concepts and with their natural language labels, such as hyperonymy and synonymy, whereas OWL allows for a more formal, set-theoretic approach and additional logical rules and constraints), serialized in Turtle syntax. Since on one hand a domain ontology should be rather concise if it is to have potential for reasoning applications (also see Section 4) and on the other both TIB and WTI had ongoing projects on mobility and transportation, we chose electric mobility as an exemplary domain, and a corresponding subset of TEMA concepts was extracted by subject experts of WTI and provided in a separate .txt file. We also integrated the "TEMA Fachordnung" into the ontology, which is a system of classification codes for the TEMA subjects that can be assigned to concepts on the ATP via a special relation.

The rationale for creating a hybrid ontology was the following: On the one hand, we wanted to keep and enhance the characteristics of a high-quality thesaurus, such as the distinction between concepts and terms, which had been only

recently performed for the TEMA thesaurus by assigning separate IDs to concepts and to each of their labelling terms (*preferred*, *alternative*, *hidden*). Such a distinction can be implemented by using SKOS-XL [3], a standardized extension of SKOS which conceptualizes terms as separate entities instead of mere strings, so that they can be described by metadata and stand in relationships with other entities themselves. Another SKOS extension we considered using is *iso-thes* [4], which aims at ensuring compliance with ISO 25964 [5] on the ideal design of and ways to ensure interoperability between thesauri, for example by allowing more complex relations between concepts such as compound equivalence.[1] We wanted to comply as far as possible with those two standards in order to achieve a maximal potential towards a high-quality knowledge organization system.

On the other hand, since according to the thesaurus manager at WTI the hyperonym relation in the TEMA thesaurus (unlike in the majority of legacy thesauri) is in large parts a proper subclass relation, in this first experiment we wanted to make the hierarchy formed by the TEMA concepts visible by interpreting it as an OWL class hierarchy and visualizing it in the Protégé editor.

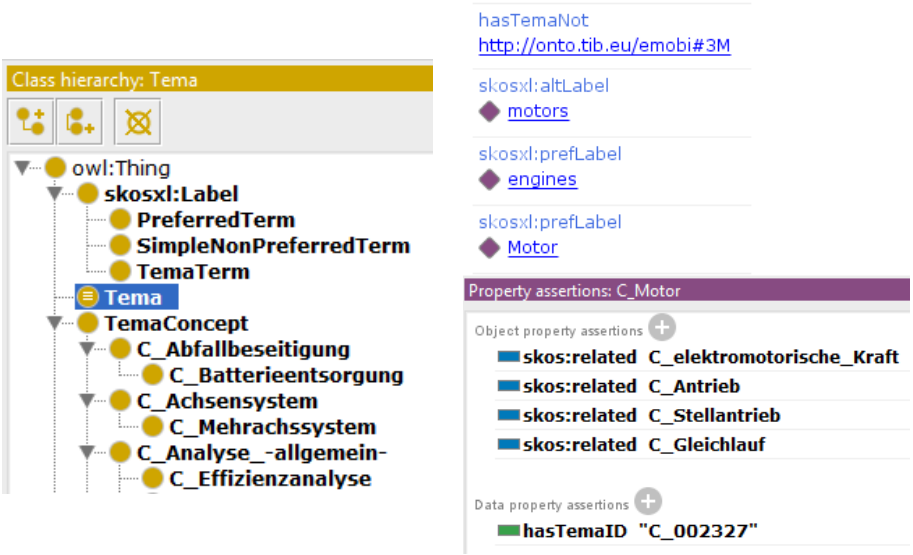

Fig. 1. TEMA hierarchy in Protégé and relations for the concept "Motor"

Work on this first sandbox ontology provided various insights on the structure of the thesaurus – for example the fact that although the ATP visualizes the terminology net in tree form, some concepts do have more than one superconcept, which then leads to the fact that identical subtrees are displayed multiple times

[1] As an example, the English concept labeled by the term "pollution" is equivalent to a concatenation of German concepts, labeled by "Umwelt" and "Verschmutzung".

(Fig. 2). Polyhierarchies in thesauri are not prohibited as such but it is important to be aware of them when trying to visualize or improve their structure.

However, the analysis also showed minor and major flaws in the structure of the thesaurus. Some minor flaws were probably introduced during the manual merge of the six thesauri that served as sources for the TEMA thesaurus, for example the one in Fig. 2 on the right where two concepts both have a subconcept that logically should be their superconcept because it subsumes them both. Therefore, it was decided to dedicate a subproject to cleaning the thesaurus up, with a focus on the (poly-)hierarchy but also with some preliminary considerations towards more expressive relations between concepts.

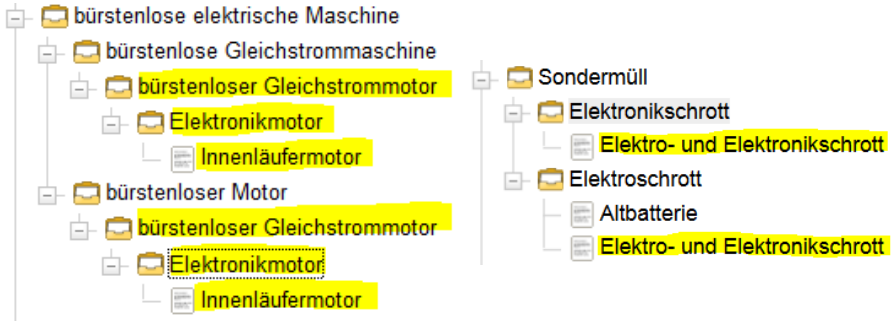

Fig. 2. Polyhierarchy and hierarchical glitches in the TEMA thesaurus

3 A top-level structure for the TEMA thesaurus

The structural analysis of the TEMA thesaurus had also shown that there were approximately 2 200 concepts without a superterm. Since the ATP visualizes the terminology net in tree form, all those terms are shown as the direct children of an artificial mother node. Consequently, we decided to provide the thesaurus with a top-level ("roof") structure with no more than 40–50 branches per level and between 3 and 10 levels that would allow a subject-driven access to its contents. The long-term objective would be to enable a visual explorative navigation in search portals for resources that have been indexed with the thesaurus.

Initially, we identified several candidates that could serve as a primary source for such a top-level structure, among them the International Patent Classification (IPC), the Dewey Decimal Classification (DDC), (a subset of) the GND classification, and the "TEMA Fachordnung" (see Section 2). These were visualized by the author as collapsible trees using JSON and the JavaScript library D3, and provided to the subject experts of WTI for inspiration. In a next step, each subject expert of WTI came up with their own proposal, and the resulting structures were then merged manually by the author and again visualized.[2]

[2] The merge can be explored at https://ontologie.tib.eu/toplevelkand/sturmerge.html.

Currently, the subject experts of WTI are working on refining and expanding the latest candidate structure collaboratively.

Throughout the project TIB and WTI are having regular workshops in order to discuss issues that come up while working on the top-level structure, i.e., questions such as how to integrate this new structure into the actual thesaurus (physically or virtually), which system best to use to maintain the thesaurus and its top-level structure in the future, but also very fundamental questions on thesaurus and ontology design, the proper division into categories, the degree of formality to choose, and the set of relations to use for our specific purposes (for example, to make the thesaurus a suitable source of concepts for smaller, more formalized domain ontologies). We will discuss the latter in the next section.

4 Relations for the TEMA thesaurus

A major issue for the subject experts of WTI when creating the top-level structure was the question of which rationale to use to build the underlying hierarchy. Since for the top-level structure we had decided on a strict monohierarchy, each concept would have one superconcept only. As a consequence, should subareas of chemistry such as petrochemistry be subsumed under chemistry or under the areas where they are applied, e.g., in the oil industry – given the applied character of TEMA? And since the top-level structure was meant to facilitate access by subjects, how would users with different approaches find their respective areas of interest? This leads to the question of how to implement what in the knowledge organization area is known as 'facets'. As of now, such facets are implemented in the TEMA thesaurus by additional cross-sectional concepts such as "part of a machine", "application (general)", etc. which is not ideal since it suggests an underlying hierarchical principle rather than the intention of a facet. A future solution should allow the extraction of subsets of concepts from the thesaurus that share a common aspect but cannot be obtained by selecting a single subtree.

One approach that we propose is to transfer the thesaurus as a whole into the SKOS format and to exploit the resulting possibilities. Although the ATP does have a SKOS export functionality, the output is rather unsatisfactory (missing fields, no SKOS-XL support). The author has programmed an alternative to fix this but it only transfers concepts one by one from .txt into SKOS-XL. However, SKOS provides constructs that can introduce more structure into a Linked Data thesaurus: SKOS concept schemes and SKOS collections [2].

A SKOS concept scheme is an aggregation of concepts sharing a common topic – also called a "microthesaurus" [5] – and via the relation *skos:hasTopConcept* one or more upper concepts can be specified within the scheme. Thus, this would allow us to present the TEMA thesaurus as a cohesive SKOS vocabulary with one or more hierarchical structures in it. SKOS collections, on the other hand, are intended to represent groupings of concepts that share a certain (ideally non-trivial) aspect and are thus additional organizational features orthogonal to the concept hierarchy. An example for such a collection given in the SKOS Primer [2] is "milk by source animal" – an aspect which is not likely to be

chosen as a top concept for a hierarchy but might nevertheless be relevant in some search scenario. SKOS collections are flexible enough to represent both subdivisions of concepts with the same superconcept, e.g., "milk" (also called "thesaurus array" [5]), and facets, i.e., groupings of concepts sharing some aspect which may not necessarily be their superconcept (e.g., "everything related to photography" – "photographer", "tripod", "camera"). Accordingly, we could define SKOS collections for "theoretical foundations that are applied in the oil industry", "theoretical foundations that are applied in the mining industry", etc., and make them accessible as facets to search engines in portals for resources that are indexed with the TEMA thesaurus. Since these measures towards a better structuring of the thesaurus have not been undertaken yet it is an obvious next step.

Another goal of the project "Fachontologie Technik" which we also carry on is to evaluate if the TEMA thesaurus can serve as a base for the development of more formalized domain ontologies that allow for scientific applications such as reasoning and question answering (also see Section 2). It was soon clear that we cannot transform the thesaurus as a whole into an ontology since reasoning in big ontologies generally involves too many steps and relies on the exactness of too many relations so that the result is likely to be unreliable. Moreover, polyhierarchical ontologies are more difficult to maintain and more care has to be devoted to make sure that all inheritances remain correct when changing it. However, it is possible to use thesauri as "quarries" from which to extract subsets of concepts, enrich them with more expressive relations, axioms, and logical rules, and thus transform them into concise ontologies for a certain domain (see for example [6,7]). In order to transform the TEMA thesaurus into such a source, we would have to identify relations that can be established in a meaningful way between the concepts (such as *isApplicationOf*, *isMachinePartOf*), and the subject experts of WTI are collecting such relations during their work on the structure of the thesaurus, for future reference and integration.

5 Summary and Outlook

It is desirable to transfer as much domain knowledge as possible from legacy vocabularies into well-structured KOS in compliance with current Semantic Web standards so that instead of getting lost it can be used both in indexing / search applications (in the form of a thesaurus) and in inferencing / question answering scenarios (ontologies). However, such a transformation is also a challenge which requires a thorough analysis of the terminological material so that historically grown structural peculiarities can be eliminated in order not to become an obstacle for the target system and its applications. A next step would be to continue purifying and enriching the TEMA thesaurus using state of the art thesaurus engineering tools (e.g., *VocBench3* [8]) in order to exploit its maximal potential with respect to Semantic Web applications and alignment with other KOS.

References

1. Bernauer, E., Mehlberg, M., Runnwerth, M., Schmidt, G.: Towards a comprehensive knowledge organisation system for the engineering domain. Slide presentation at the Workshop on Classification and Subject Indexing in Library and Information Science (LIS'2015), in conjunction with the European Conference on Data Analysis (ECDA) (2015). Available from https://publikationen.bibliothek.kit.edu/1000049929
2. SKOS Simple Knowledge Organization System Primer, https://www.w3.org/TR/skos-primer/. Last accessed 30 May 2018
3. SKOS Simple Knowledge Organization System eXtension for Labels (SKOS-XL) Namespace Document, https://www.w3.org/TR/skos-reference/skos-xl.html. Last accessed 30 May 2018
4. ISO 25964 SKOS extension (iso-thes), https://lov.okfn.org/dataset/lov/vocabs/iso-thes. Last accessed 30 May 2018
5. ISO 25964 the international standard for thesauri and interoperability with other vocabularies, https://www.niso.org/schemas/iso25964. Last accessed 30 May 2018
6. Kless, D., Jansen, L., Lindenthal, J., Wiebensohn, J.: A method for re-engineering a thesaurus into an ontology. In: Donnelly, M., Guizzardi, G. (eds.) Proceedings of the Seventh International Conference (FOIS 2012), pp. 133–146, IOS Press (2012)
7. Nowroozi, M., Mirzabeigi, M., Sotudeh, H.: The comparison of thesaurus and ontology: Case of ASIS&T web-based thesaurus and designed ontology, Library Hi Tech (2018). https://doi.org/10.1108/LHT-03-2017-0060
8. VocBench homepage, http://vocbench.uniroma2.it/. Last accessed 30 May 2018