Crafting a System for Knowledge Discovery and Organisation
A Case-Study on KOS for a Non-Standard German Legacy Dataset

Abstract:
This paper describes a case-study developing a knowledge organisation system (facet thesaurus) on the example of a non-standard German language legacy dataset, DBÖ [Datenbank der bairischen Mundarten in Österreich / Database of Bavarian Dialects in Austria]). A particular focus is placed on the 109 original data collection questionnaires contained in the collection, which are understood as an entry point to the entire collection. Here they serve as a case-study to demonstrate the process, which may be extended to the remainder of the collection. Ranganathan (1933, 1967) created the first faceted scheme - Colon Classification - to classify books in libraries. Faceted classification has also been used to assist automated search and retrieval of information (Prieto-Diaz 1991). According to Mills (2004), facet analysis has a very vital role for information retrieval and in the design of classificatory structures by the application of logical division to all forms of the content of records, subject and imaginative. The natural product of such division is a faceted classification.

Building on these previous endeavours, we here introduce a facet thesaurus for eliciting and promoting access and navigability for the items in this collection, in order to make cross cutting topics accessible.

1.0 The aim and scope of the study & introduction
The aim of this study is to introduce a first approach towards creating a knowledge organization system (facet thesaurus) for a non-standard German language legacy resource (DBÖ)\(^1\), enabling transversal knowledge discovery. Here we present the rationale to our approach, the applied methodologies, and a concrete example on technology related terms in the form of a case-study that, in a next step, can be readily applied to other thematic areas within and beyond the collection.

Our undertaking is realised within the project exploreAT!\(^2\) and the wider framework of exploration space\(^3\) at the Austrian Centre for Digital Humanities and Cultural Heritage (ACDH-OeAW). The exploration space is a digital and physical space offering opportunities for experimentation and innovation in the networked Humanities, and has since its establishment in 2017 also been listed as a best practice example for Open Innovation in the Humanities\(^4\). The project exploreAT! is a multidisciplinary endeavour with international collaboration partners (semantic technologies: Adapt Centre, Dublin City University, IE; visualisation tools: VisUSAL, Universidad de Salamanca, ES), with the general aim of opening the DBÖ collection for thematic exploration and exploitation

---


\(^2\) https://www.oeaw.ac.at/acdh/projects/exploreat/

\(^3\) https://www.oeaw.ac.at/acdh/about/core-units/core-unit-4/

\(^4\) http://openinnovation.gv.at/portfolio/oeaw-exploration-space/
by means of semantic technologies and visual prototyping, as well as linking to other resources.

The DBÖ collection is large and rich (~3.5 million entries) and is composed of digitised data collection questionnaires, answers as well as excerpts of vernacular dictionaries and folklore literature. The data, having undergone several stages of digitisation, is to-date available in TEI/XML format and partly as a MySQL database. The questionnaire data (109 thematic questionnaires comprising around 17,000 individual questions) thus constitutes only a fraction of the DBÖ. The questionnaires originally pertained to a dictionary project aimed at capturing the German language spoken by local population from the early 20th century onwards in the area of the former Austro-Hungarian empire. Therefore, apart from being a rich linguistic, non-standard resource, the collection also captures a wealth of historic cultural information of everyday life, e.g. customs, religious festivities, food, traditional medicine, professions, songs, among others. The answers to the questionnaires’ questions thus follow a lexicographic structuring and are composed of headwords (lemmas), senses/meanings, sources, geographic location information (GIS), person information (authors, collectors, data typists), etc. In the context of the exploreAT! project, opening up the collection and questionnaires to new ways of exploration was initiated via lexical concepts, enabling the linking to other resources, such as Linked Open Data (LOD) (cf. Abgaz et al. 2018; Dorn et al. 2019). With this came the necessity for transversal knowledge searching across questionnaires and access to knowledge otherwise inaccessible or hidden. The main topics of the questionnaires reflect more or less detailed aspects of everyday life, which is also reflected in the quantity of questionnaires dedicated to a certain topic. Topics such as “movement” (Bewegung) (11 questionnaires), “wedding” (Hochzeit) (5 questionnaires), “tailoring” (Schneiderei) (4 questionnaires) or “baking bread” (Brotbacken) (3 questionnaires) have been queried extensively, while other topics (e.g. body parts, time, animals, school and education, plants, brewing) were covered with fewer questionnaires. Based on this content and the overall topics represented, we have chosen use-cases that would allow us to explore the data from perspectives relevant for cultural exploration as well as from the Digital Humanities perspective in general. Exploring the basic subject of “technology” across the collection, on the one hand, would allow us to provide a “historical” view of technology related terms and concepts from the time when the questionnaires were conceptualised (1912)⁵, while, at the same time, providing a noticeable contrast to what we understand by technology today. The other selected use-case deals with the topic of “food”, which is covered by specific questionnaires, but also transversally with specific, food related questions occurring across several questionnaires. In addition, food is one of the key aspects of mankind’s culture and thus particularly relevant not only from a historical perspective but also nowadays.

Navigating through such a vast collection is hard when dealing with purely term based search interfaces, and that is why we have chosen the faceted approach to organize the main concepts of the collection in a thesaurus, that would serve as the entry point for navigating through the questions, each one linked to one or (many) more answers. Relevant terms were chosen using its raw frequency in the questions. However, word

⁵ https://vawadioe.acdh.oeaw.ac.at/projekte/wboe/geschichte-des-wboe/
embedding models (Mikolov 2013a) were also used to explore semantic vicinities and to find similar concepts among the whole collection. Ultimately, we have achieved our goal to build a navigable interface providing access to the data fields in the collection of cultural heritage.

2.0 Theoretical background

In the scope of this project, two main knowledge representation techniques were used: faceted analysis and word embeddings. A brief overview of these is provided in this section below.

Faceted classification is a concept and a technique introduced by Ranganathan (1933, 1967) and later developed by the Classification Research Group (CRG) (Vickery 1960). A faceted scheme has several facets and each facet may have several terms, or possible values. A faceted classification scheme for wine, using (Broughton 2006) example, might include the facets (and terms) “grape varietal” (riesling, cabernet sauvignon, etc.), “region” (Napa Valley, Rhine, Bordeaux, etc.), and “year” (2001, 2002, etc.). According to Ranganathan, the process of choosing the facets is analytic-co-synthetic, that is, we first analyze the subject domain and then shape their compounding facets to adequately describe its characteristics and provide room for the concepts. That makes it a very powerful resource to describe and organize information. The facets need not be ordered, nor be of the same type, although they should be clearly defined and mutually exclusive (Broughton 2006). It was first devised for the classification of books in libraries (Ranganathan 1933), but was subsequently adopted in information retrieval systems and search interfaces on the web (Prieto-Diaz 1991; Broughton and Lane 2000; Tudhope et al. 2006). Facet analysis has been used in the construction of information retrieval (IR) thesauri since the publication of the Information Retrieval Thesaurus of Education Terms in 1968 (Spiteri 2000; Barhrydt, Schmidt, and Chang 1968). Spiteri (2000) states that there is no standardized model for the application of facet analysis to information retrieval thesauri and that national and international guidelines for thesaurus construction make minimal mention of the use of facet analysis. She presents, in his study, some critical arguments on how to evaluate the choices of facets for thesauri, and if these choices are coherent to the principles stated by both Ranganathan and the CRG. Ranganathan (1933) has proposed, influenced by the Brahman philosophy (Mazzocchi, 2013), the adoption of basic subjects and 5 characteristics of division used to derive facets: Personality (that we can interpret as “the things”), Matter (its characteristics), Energy (the processes), Space, and Time. His facets system has since the epithet "PMEST". The CRG, alternatively, preferred an ad hoc approach, and proposed that each subject area should be divided into categories that are appropriate to its nature. This latter approach was more suitable for our collection.

Word embeddings are one of the many powerful NLP techniques that have been developed in the past few years. To build semantic models out of large textual collections, we need to represent the semantic units into mathematical vectors. There are mainly two ways to construct this representation: using the simple bag of words model (Zhang et al. 2010), where each word is represented by a specific vector in a huge multidimensional space; and using word embeddings models, such as Word2vec (Mikolov et al. 2013a; Mikolov et al. 2013b), where each word is represented by a linear combination of a smaller set of dimensions or vectors. These “basic” vectors for word
composition are obtained using specific neural network architectures (e.g. “continuous bag-of-words” or “skip-gram”) (Mikolov, Yih, and Zweig 2013). These distributed vector representations of texts impressively capture syntactic and semantic aspects of concepts and their relationships. To generate these contextualized word representation models, it is necessary to feed the underlying neural networks with large corpora of texts of a specific language, so that the transitive relations between the concepts that co-occur in "windows" of contiguous words in a sentence are captured. Word embeddings were used in our project to explore associations between words and to choose possible non-preferred terms on the thesaurus, even if these terms were not as frequent in the raw count.

3.0 The method

The methodologies applied in our study involve both automatic/machine based as well as manual/intellectual processes, in a collaborative setting of technical and domain experts. As a first step, the ~17,000 questions across the 109 questionnaires were tokenized and lemmatized. Abbreviations of words were resolved and stop words (e.g. articles, prepositions, etc.) removed. The non-concepts, i.e., words related to syntax, morphology, semasiology and onomasiology used to build the questions were identified and removed. The remaining "cultural heritage" words were extracted automatically using Python scripts and ranked according to their frequency across all questions (min.=1; max=609). This has yielded a total of 88,883 distinct terms, which we refer to as concepts. As put before, from all the basic subjects pertinent to the collection, we have chosen two for the proof of concept: technology (Technologie) and food (Essen), in which two different sets of domain expert in the team were involved. Technology related terms were identified manually among the most frequent by the domain experts, involving two rounds of evaluation and agreement processes, yielding a total of 186 terms. As we are dealing with a non-standard language collection, automatic identification of technology related terms would have only been partially successful, in spite of the availability of linguistic Knowledge Organization Systems as the german Wordnet (Germanet) and German Thesauri. Then, both the domain and the technical experts determined a suitable set of facets for the thesaurus, based on the concepts that were harvested and the guidelines from examples made by the CRG. The chosen facets were: trade/crafts (Gewerbe); artifacts (Artefakt), processes (Prozess), roles (Rollen) places of application (Anwendungsort), areas of application (Anwendungsbereich) and quality (Eigenschaften). Subsequently, the 186 technology related terms were assigned to these facets by the domain experts, again involving two rounds of evaluation and agreement. Terms that could not be clearly assigned, or when no agreement was reached among the domain experts were temporarily excluded (n=11). For these terms further evaluation is needed in future developments. Finally, a total of 175 technology related terms were distributed as follows: trade/crafts (n=8); artifacts (n=100), processes (n=46), roles (n=4), fields of application (n=9) and places of application (n=8). The food related terms are still being in the process of being harvested and the process has not been completed yet.

6 https://github.com/acdh-oeaw/exploreAT-Concepts
The chosen tools for the management of the vocabularies and data, and the display of the thesaurus hierarchies was the free and open source tool Tematres\(^7\) with the Visual Vocabulary addin\(^8\) installed. The database with the questions was kept as a Pandas Dataframe\(^9\), served by a Pandas REST API\(^10\) bridging the access to the Django REST Framework\(^11\).

After all the technology related terms were chosen, the preferred and non-preferred terms were added to the online thesaurus tool according to the facets in the hierarchies crafted. After all the terms were registered, the connections among the related terms were assigned. The final step was establishing links between each preferred term to the database of questions and answers that contained the term. This was made via the Tematres Multilingual Vocabularies\(^12\) resource ("Relations between vocabularies: RelatedMatch"), that takes the URL and generates a web link in the thesaurus management tool to the dataframe/database kept in the Pandas structure. The main advantage of this architecture is that it is comprised of an autonomous single Docker container that has the LAMP stack (Linux, Apache, MySQL, PHP/Python) tools, besides the Tematres with its enhancements, and it can be deployed quickly in any new environment. We plan to make a Docker image available - devoid of data - in the DockerHub\(^13\) as the initial (and outdated) version of Tematres\(^14\) that was used as the initial image was updated. This will help accelerate the deployment of solutions like this in the future.

### 4.0 Results

The in this section we will show some illustrations of the current state of the solution. Figure 1a below depicts the homepage with both the alphabetic and systematic displays for the thesaurus. The users can also make queries using the query box provided by the Tematres tool.

Figure 1b illustrates a detail on the specific basic subject of technology, showing broader and specific terms and Figure 2 presents the details for the term “Schiff” (ship) under the facet “Artifakts”. We have used bibliographic notes to provide hyperlinks (related match) to the database sources. This kind of coupling between the concept navigation tool and the data can be changed by simply substituting the URIs. In a future version, we plan on implementing a more smooth interface as we can see in the Getty Art & Architecture Thesaurus Hierarchy Display\(^15\). The main advantage of the current architecture of the solution is that it allows for rapid deployment and uses free open source tools.

\(^{7}\) https://www.vocabularyserver.com/
\(^{8}\) https://github.com/tematres/visualVocabulary [last access: 12.12.2019]
\(^{10}\) https://github.com/wq/django-rest-pandas [last access: 12.12.2019]
\(^{11}\) https://www.django-rest-framework.org/ [last access: 12.12.2019]
\(^{13}\) https://hub.docker.com/ [last access: 12.12.2019]
\(^{14}\) https://hub.docker.com/r/systemsector/tematres [last access: 12.12.2019]
Finally, we also offer a hyperbolic geometry navigation interface, provided by the Tematres solution as an optional addin (Figure 3).
5.0 Conclusions and future work

In this paper we have presented a case-study on the development of a knowledge organisation system: a facet thesaurus, for transversal knowledge discovery within a non-standard language legacy dataset. We have demonstrated that a facet approach combined with both manual (human) and automatic (machine) generated concepts are essential for eliciting cultural knowledge. This setting, combined with the online thesaurus and the connection to the database with the original concepts, have provided an effective and intuitive way for navigating and retrieving the information contained in the DBÖ resource in a structured and accessible way. Our approach makes it possible to access, link and visualise thematic knowledge also transversally, which has often proven challenging. As a next step, and as a future extension of the basic subjects being identified. Food is close to completion, and then customs, religious festivities, food, traditional medicine, professions and songs are to follow.

References
Collection Questionnaires on the example of exploreAT!. Budapest: Centre for Digital Humanities, Eötvös Loránd University, 13-15.


© Ergon - ein Verlag in der Nomos Verlagsgesellschaft