Authors
Amelie Dorn[1], Barbara Piringer[1], Yalemisew Abgaz[2], Jose Luis Preza Diaz[1], Eveline Wandl-Vogt[1]

[1] Austrian Centre for Digital Humanities, Austrian Academy of Sciences, Austria
[2] Adapt Centre, Dublin City University, Ireland

Title

Enrichment of Legacy Language Data:
Linking Lexical Concepts in Data Collection Questionnaires on the example of exploreAT!

Keywords: legacy language data, concepts, DBpedia, Linked Open Data, wikification

Abstract

The project *exploreAT!* (Wandl-Vogt et al, 2015) runs as a module within exploration space @ ACDH-OeAW, an Open Innovation Research Infrastructure (OI-RI) for the Humanities.
It aims at revealing and making accessible cultural knowledge contained in a non-standard German language resource (DBÖ [Database of Bavarian Dialects in Austria]), exploiting it from the perspectives of semantic technologies, visual analysis tools and cultural lexicography. Cultural information captured in legacy language collections often remains unaccessed, or requires specific data processing efforts. The DBÖ collection is a large and rich heterogeneous resource (~3.5 mio. entries) from the time of the former Austro-Hungarian monarchy, comprising of digitized data collection questionnaires, answers and excerpts from vernacular dictionaries and folklore literature. With 109 systematic data collection questionnaires, linguistic but also a wealth of cultural information (customs, festivities, food, etc) was captured. By opening up the questionnaires through lexical concepts and their linking to other resources, we aim to enable and foster to understand the “rural world” in the early 20th century as captured in the data (cf. Arbeitsplan, 1912).
In this paper, we describe and demonstrate our approach to linking individual questionnaire topics/lexical concepts to DBpedia concepts and we outline the processes and challenges.
First, individual questionnaire topics, originally in German, were automatically extracted from the questionnaire title which defines the topic of a questionnaire, e.g. *Farben* (DE)/ *colours* (EN), and all questions contained therein. Next, titles were translated into their English equivalent or nearest best fit. To link the questionnaire titles/topics to external resources, e.g. DBPedia, dbpedia spotlight service was employed. For each questionnaire, English DBPedia concepts were identified with a certain degree of confidence, and the corresponding concept generated. Then, experts evaluated the accuracy of the results in a csv file by comparing the DBpedia concepts definitions with the topic of the questionnaire. This knowledge was tapped from experienced experts familiar with the detailed contents of the questionnaires and questions. Where necessary, additional DBpedia concepts were added manually. Once an agreement was reached, these concepts are added permanently to the database, and used as an authoritative source to link the questionnaires with DBpedia concepts.
A number of challenges arose in this process: linking topics to matching German DBpedia concepts was not always straightforward. At times, concepts were only available in English, and sometimes there was no equivalent in DBpedia at all. In these cases, experts opted for the nearest fitting equivalent and noted this uncertainty in the evaluation file. Further, nuanced differences in meaning between German and English words and their corresponding DBpedia

concepts (e.g. German: Schuster; English: shoemaker; cordwainer) also proved challenging. As a result, levels of detail in assigned concepts across questionnaires vary.
Next steps involve the improvement of the concepts assignment as well as adding further concepts to provide a more detailed picture of the conceptualisation of these legacy language data questionnaires, including visualisations of relations and networks (e.g. Apache Superset, see Figure 1).

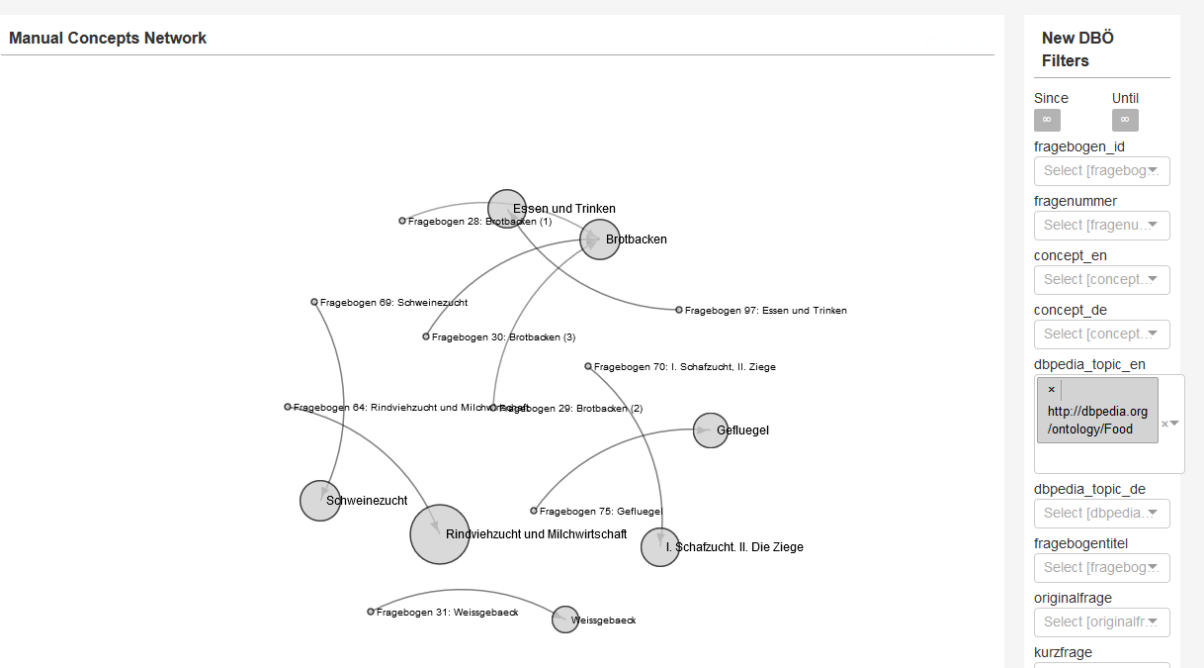

Figure 1: Network visualisation of the DBpedia concept “Food” and linked questionnaires in Apache Superset.

Bibliography

Apache Superset
https://superset.incubator.apache.org/ [accessed: May, 14th 2019]

Arbeitsplan. (1912). Arbeitsplan und Geschäftsordnung für das bayerisch-österreichische Wörterbuch. 16. Juli 1912. Karton 1. Arbeitsplan Bayerisch-Österreichisches Wörterbuch. Archive of the Austrian Academy of Sciences. Wien.

DBpedia
https://wiki.dbpedia.org/ [accessed: May, 14th 2019]

DBpedia Spotlight
https://www.dbpedia-spotlight.org/ [accessed: May, 14th 2019]

[DBÖ] Österreichische Akademie der Wissenschaften. (1993–). Datenbank der bairischen Mundarten in Österreich [Database of Bavarian Dialects in Austria](DBÖ). Wien. [Processing status: 2018.01.]

[dbo@ema] Wandl-Vogt, E. (2010; Ed.). Datenbank der bairischen Mundarten in Österreich electronically mapped [Database of the Bavarian Dialects in Austria electronically mapped] (dbo@ema). Wien. [Processing status: 2018.01.]

Wandl-Vogt, E., Kieslinger, B., O´Connor, A. and Theron, R. (2015). „exploreAT! Perspektiven einer Transformation am Beispiel eines lexikographischen Jahrhundertprojekts“, in: Dhd2015. Von Daten zu

Erkenntnissen. 23. bis 27. Februar 2015, Graz. Book of Abstracts. http://dhd2015.uni-graz.at/de/nachlese/book-of-abstracts/ [accessed: May, 14th 2019].