The Cocoda Mapping Tool

U. Balakrishnan, J.Voß
Head Office of the GBV

Since the 90s, we have seen an explosion of information and with it there is an increase in the need for data and information aggregation systems that store and manage information. However, most of the information sources apply different Knowledge Organizations Systems (KOS) to describe the content of stored data. This heterogeneous mix of KOS in different systems complicate access and seamless sharing of information and knowledge. Concordances also known as cross-concordances or terminology mappings map different (KOS) to each other for improvement of information retrieval in such heterogeneous mix of systems. (Mayr 2010, Keil 2012). Also for coherent indexing with different terminologies, mappings are considered to be a valuable and essential working tool.

However, despite efforts at standardization (e.g. SKOS, ISO 25964-2, Keil 2012, Soergel 2011); there is a significant scarcity of concordances that has led an inability to establish uniform exchange formats as well as methods and tools for maintaining mappings and making them easily accessible. This is particularly true in the field of library classification schemes. In essence, there is a lack of infrastructure for provision/exchange of concordances, their management and quality assessment as well as tools that would enable semi-automatic generation of mappings.

The project "coli-conc" therefore aims to address this gap by creating the necessary infrastructure. This includes the specification of a data format for exchange of concordances (JSKOS), specification and implementation of web APIs to query concordance databases (JSKOS-API), and a modular web application to enable uniform access to knowledge organization systems, concordances and concordance assessments (Cocoda).

The focus of the project "coli-conc" lies in semi-automatic creation of mappings between different KOS in general and the two important library classification schemes in particular - Dewey classification system (DDC) and Regensburg classification system (RVK). In the year 2000, the national libraries of Germany, Austria and Switzerland adopted DDC in an endeavor to develop a nation-wide classification scheme. But historically, in the German speaking regions, the academic libraries have been using their own home-grown systems, the most prominent and popular being the RVK. However, with the launch of DDC, building concordances between DDC and RVK has become an imperative,

although it is still rare. The delay in building comprehensive concordances between these two systems has been because of major challenges posed by the sheer largeness of these two systems (38.000 classes in DDC and ca. 860.000 classes in RVK), the strong disparity in their respective structure, the variation in the perception and representation of the concepts. The challenge is compounded geometrically for any manual attempt in this direction.

Although there have been efforts on automatic mappings (OAEI Library Track 2012 -- 2014 and e.g. Pfeffer 2013) in the recent years; such concordances carry the risks of inaccurate mappings, and the approaches are rather more suitable for mapping suggestions than for automatic generation of concordances (Lauser 2008; Reiner 2010). The project "coli-conc" will facilitate the creation, evaluation, and reuse of mappings with a public collection of concordances and a web application of mapping management.

The proposed presentation will give an introduction to the tools and standards created and planned in the project "coli-conc". This includes preliminary work on DDC concordances (Balakrishnan 2013), an overview of the software concept, technical architecture (Voß 2015) and a demonstration of the Cocoda web application.

References

Balakrishnan, U. (2013): Das Projekt „coli-conc" Ein Bericht zur semi-automatischen Erstellung von Konkordanzen zur Dewey Dezimalklassifikation. In: VZG-Aktuell 1/2013, 12-16.

ISO (2013): ISO 25964-2: Information and documentation – Thesauri and interoperability with other vocabularies – Part 2: Interoperability with other vocabularies.

Keil, Stefan (2012): Terminologie Mapping: Grundlagen und aktuelle Normungsvorhaben. In: Information – Wissenschaft und Praxis. Band 63, Nr. 1, 45-55.

Lauser, Boris; Johannsen, Gudrun; Caracciolo, Caterina, Keizer, Johannes; van Hage, Willem Robert; Mayr, Philipp (2008): Comparing human and automatic thesaurus mapping approaches in the agricultural domain. In: 10th International Conference on Dublin Core and Metadata Applications. http://arxiv.org/abs/0808.2246

Mayr, Philipp & Petras, Vivien (2008): Building a terminology network for search: the KoMoHe project http://arxiv.org/abs/0808.0518

Pfeffer, Magnus (2013): Automatic creation of mappings between classification systems. In: Workshop Klassifikation und Sacherschließung (LIS'2013), University of Luxembourg (Vortrag). urn:nbn:de:swb:90-357776

Reiner, Ulrike (2010): Automatische DDC-Klassifizierung bibliografischer Titeldatensätze der Deutschen Nationalbibliografie. In: Dialog mit Bibliotheken 2010/1, 22. Jahrgang, S. 23-29. URN urn:nbn:de:101-2011012860

Soergel, Dagobert (2011): Conceptual foundations for semantic mapping and semantic search. In: F. Boteram, W. Gödert & J. Hubrich (Hrsg.): Concepts in context, S. 13-35. Würzburg: Ergon.

Voß, J. (2015) Cocoda Technical Report 1. DOI 10.5281/zenodo.16786