Towards integrated systems for KOS management, mapping, and access: Coli-conc and its collaborative computer-assisted KOS mapping tool Cocoda

Abstract
Increased use of data sharing makes interoperability between Knowledge Organization Systems (KOS) ever more important, but concordances between these systems are rather rare. Project Coli-conc aims to address this gap by developing tools, methods and techniques to simplify and accelerate the intellectual creation of concordances. It also aims to ease their use and exchange and at the same time to provide quality monitoring that aids quality management. The project creates a set of reusable software modules to enable uniform access to KOSs, concordances and concordance assessments. These modules are provided as a web application to support effective processing of concordances. In addition, existing software has been evaluated and enhanced with new components for storage of, access to and analysis of different concordances.

1. Introduction
Coli-conc, under development, enables interoperability between Knowledge Organization Systems (KOS) with an initial focus on German library KOS. We describe the functionality and architecture of Coli-conc and its mapping tool Cocoda and develop a broader design vision.

Coli-conc will provide easy access to many KOS and to mappings between KOS. It fills a void by developing a technical infrastructure for managing many KOS and KOS mappings and keeping them current. The specific objectives of the project are:

- Develop a collection of freely combinable open source APIs and provide web-based services for uniform access to KOS and KOS mappings and to support computer-assisted intellectual mapping among many KOS in multiple languages.
- Provide guidance for best practices for quality assurance of KOS and mappings.
- Collect, store, and manage KOS data and mappings in a uniform structure.
- Support using these data for many purposes, including:
  - knowledge base for automatic subject cataloging (e.g. Petrus 2011) and assistance in manual subject cataloging;
  - knowledge base for retrieval through browsing hierarchies and/or automated or computer-assisted query expansion, adding synonyms and related terms;
  - mapping and enriching subject cataloging data and queries between systems.

2. Coli-conc system description (Fig. 1)
The system architecture emphasizes synergy between the components, a common infrastructure in which modules providing a given functionality (such as intuitive hierarchy browsing) as well as data can be shared across, reducing the effort for providing functionalities and data as compared to stand-alone implementation
Fig 1: Coli-conc Functional Architecture
The key elements of the infrastructure are shown here and elaborated in the remainder of Section 2.
2.1. The KOS and KOS mappings database and the JSKOS data format

At the heart of Coli-conc is a database of KOS and KOS mapping data, including

- mappings imported from existing concordances produced by various mapping projects, such as KoMoHe (Mayr and Petras, 2008) and Wikidata (a linking hub to a large number of authority files and other KOS, Neubert 2017), and
- new mappings created in the Coli-conc project.

KOS and mapping data are structured using the JSKOS format to be discussed next. For each piece of data the sources (imported KOS and concordances, editors) are given.

![JSON Representation of KOS and KOS Mapping Data]

```json
{
    "type": ["http://www.w3.org/2004/02/skos/core#narrowMatch"],
    "fromScheme": {"uri":"http://dewey.info/scheme/edition/e22"},
    "toScheme": {"uri":"http://bartoc.org/en/node/454"},
    "from": {
        "memberSet": [
            {"uri": "http://dewey.info/class/387/e22/",
            "notation": ["387"],
            "preflabel": { "en": "Water, air, space transportation" } }
        ]
    },
    "to": {
        "memberSet": [
            {"uri": "http://id.loc.gov/authorities/subjects/sh85121579",
            "preflabel": { "en": "Shipping" } }
        ]
    }
}
```

We developed the JSKOS format for the unified representation of KOS and KOS mapping data from disparate sources (Voß, Leld and Balakrishnan 2016; Voß 2017). JSKOS is based on SKOS and JSON-LD. It allows conversion to and from RDF and to and from the MARC 21 Format for Classification Data (Heggø 2017). JSKOS was developed primarily for web applications, but it can also be used as the internal format of NoSQL databases such as MongoDB. JSKOS combines elements of SKOS and Dublin Core such as concepts, concept schemes, modification times, and publishers and extends these existing standards to cover a richer set of KOS and KOS mapping data, including the following:

- strict definition of how to encode repeatable and non-repeatable fields;
- treating mappings between KOS elements as first-class objects;
- confidence level for mappings;
- elements for concept occurrences and co-occurrences;
- an extensible list of relationship types beyond just USE/UF, NT/BT, and RT;
- mappings with multiple concepts, ordered lists;
- closed world statements.
2.2. The KOS and KOS concordance registry

This registry contains metadata for all KOS and KOS concordances that were or will be ingested into the Coli-conc database. Metadata include indicators for the quality of the data in KOS and KOS concordances; these quality indicators inherit to the specific data items in the Coli-conc database. After a review of KOS registries (Voß, Agne, Balakrishnan and Akter 2016), we focused on BARTOC (Basel Register of Thesauri, Ontologies & Classifications) and used its metadata schema as the basis for the more complete Coli-conc registry schema (Voß, Ledl and Balakrishnan 2016), which also refines the NKOS KOS types used in BARTOC with KOS types from Wikidata (Voß 2016), to be extended to a faceted classification of KOS types. The Coli-conc KOS registry contains a subset of BARTOC. It supports keyword search and filtering by KOS type and downloading KOS metadata.

2.3. Input processes

These include general data checking and cleaning, transformation to the JSKOS format, and ingestion to the database with detection and consolidation of duplicates.

2.4. The Cocoda mapping tool

The core part of Coli-conc is the mapping tool Cocoda, a platform for computer-assisted intellectual mapping. The main modules of Cocoda are:

• The suggestion tool provides mapping suggestions from several sources:
  • mappings from various sources/projects, including Wikidata, stored in the Coli-conc database; a source may create mappings manually and/or automatically;
  • implicit mappings derived from correlating subject descriptors (subject headings or classes) assigned to the same document in a library or union catalog (Buckland et al. 1999).

• The edit and collaboration module supports computer-assisted editing for high-quality mappings. An editor works with mapping suggestions and selects the best one, edits a mapping, or enter an entirely new mapping. The editor can also explore the source and target KOS to better understand concepts to verify a suggested match or find a better one. The collaboration function allows several editors to divide work on mappings for one KOS pair and communicate about specific mappings, and it supports expert review of the mappings.

• The KOS Representation Module supports KOS exploration through searching, browsing, and displaying the hierarchical and other KOS structure (see 2.5).

• The measure module for monitoring quality. It produces statistics on statistics on: the types of KOS, frequency of usage on KOS and terms for search, Information on mappings: per KOS, per topic or subject field, their frequency, new entries, status of on-going mapping work.
Fig 3: User Interface - Mapping Tool Cocoda
2.5. User exploration interface

The KOS exploration interface is provided through a module used throughout the system, in Cocoda, in systems to assist catalogers, and in the end-user interface. It supports searching, meaningful display of hierarchies and concept maps (visual displays of concept relationships) for browsing, display of all information about a concept, and navigation through hyperlinks.

2.6. Terminology services and connection to cataloging systems

Convenient access to actual KOS data notations, concepts, terms, and relationships requires web services. A few systems make KOS data available via APIs (Voß, Agne, Balakrishnan and Akter 2016), but coverage and quality of terminology services differs widely. Queries to a terminology service should support applications such as browsing and searching in KOS in a uniform way. Based on a review of existing KOS APIs, we are developing a JSKOS API both as a wrapper for existing terminology services and as an independent terminology service (Dührkoph 2017). Access honors restrictions, such as those imposed by OCLC for the Dewey Decimal Classification.

As we mentioned above, the KOS and KOS mapping data in the Coli-conc database are very useful to catalogers. Cataloging systems can make these data available by importing them using the terminology service (under consideration by the semi-automatic subject cataloging tool Digitaler Assistant, Hinrichs et al. 2016), or they can arrange with Coli-conc for integrating use of Cocoda into their system.

3. Next steps

In the short term the project will focus on
- ingesting more data and making the system operational;
- improving browsing functionality and usability;
- defining criteria for mapping relationship types;
- moving to a completely integrated DBMS to ensure higher performance;
- automatic enrichment of VZG’s “Gemeinsamer Verbundkatalog” (union catalog).

In the medium term, we are thinking about
- extension to other KOS (museums, archives, all areas);
- support for KOS in multiple languages using existing knowledge sources, including dictionaries, WordNet, and multilingual KOS such as the Library of Congress Subject Headings (LCSH), the Universal Decimal Classification (UDC), the Dewey Decimal Classification (DDC), and the Unified Medical Language System (UMLS), which includes the Medical Subject Headings
(MeSH) among its 150+ KOS;

- making many automated mapping tools available (See Fifty …);
- implement measures for quality assessment.

4. A vision of Knowledge Organization System management utopia

In our utopia the functionality of a system like Coli-conc would be expanded to include the functionality of a full-fledged KOS management tool that combines the best features of existing tools, including deriving concepts, terms, and relationships from a corpus of relevant texts, creating and maintaining a meaningful hierarchical arrangement of concepts, an efficient user interface; and collaborative KOS management. The Coli-conc environment is ideal for creating and maintaining KOS, since an initial collection of concepts and terms and many kinds of relationships can be extracted from the KOS and KOS mappings database.

Both mapping and establishing concept relationships within a KOS can be made easier and better through a concept hub. (Fig. 4) (Soergel 2011). In the hub, concepts are expressed through combinations of elemental concepts, more precisely through description logic (DL) formulas (Bechhofer and Goble 2001). Two terms with the same DL formula, are likely to designate the same concept. If they are not, we need to refine the system of elemental concepts and/or the system of syntagmatic relationships used in the DL. If the two terms come from different KOS, the hub establishes a mapping; if they come from the same KOS, the two terms are synonyms.

The beauty of this approach is that a reasoner can infer relationships between DL formulas and therefore between the corresponding concepts. This can be used to assist in the elaboration of the structure of one KOS and in establishing mapping relationships other than equality between two KOS.

We hope to try out the concept hub approach in a pilot, comparing inferred relationships with verified relationships in the database.
5. Conclusions

We have presented a vision of an integrated systems for KOS management, mapping, and access and pointers towards the incremental and modular implementation of the functionality of such a system. The ideas presented should be helpful to designers and developers. Such an integrated system opens many possibilities for improving information systems to better support users.

Acknowledgements

Coli-conc is funded by the Deutsche Forschungsgemeinschaft (DFG) (German Research Foundation) grant no. GZ-DI 364/8-1 / AOB J 622525, extension GZ-DI 1364/8-3 / AOB J 642034 to the Verbundzentrale des GBV (VZG.), the Head Office of the Gemeinsame Bibliotheksverbund (Common Library Network)

References


Heegø, D. (2017) mc2skos. Python script for converting Marc21 Classification and Authority records to SKOS/RDF. Available at: https://github.com/scriptotek/mc2skos.