Meaningful Concept Displays for KOS Mapping

Xia Lin, Jae-wook Ahn
Drexel University

Dagobert Soergel
The University at Buffalo

There are two observable trends in recent development of Linked Data applications with Knowledge Organization Systems (KOS). One is the growth of Linked Open Vocabularies (LOV) [1]. The other is the conversion or release of traditional KOS systems as Linked Open Data. Currently, there are more than 446 vocabularies in the LOV vocabulary space. Some are significantly large with rich semantic relationships defined; some are relatively small and incomplete. LOV vocabularies are described in RDFS or OWL. They can be queried either at the vocabulary or element level. They are aggregated in one single endpoint and searchable through SPARQL queries.

On the other front, major institutions have released traditional, well-defined controlled vocabularies and classification systems to be used as Linked Open Data (LOD). For example, Getty Research Institute released Art & Architecture Thesaurus (AAT) as LOD February this year [2], and Library of Congress has provides a linked data service to access multiple LC vocabularies such as LC Subject Headings, LC Classification, Thesaurus for Graphics Materials, etc. [3]. OCLC has both made available its Dewey Decimal Classification (DDC) and its 194 millions bibliographic work description as LOD [4, 5].

While a main purpose of linked open data is to facilitate matching terms and concepts across vocabularies or be able to aggregate multiple vocabularies for a single application, it is still a significant challenge to link or complete semantic alignment among linked open data [6]. Making LOD available through SPARQL endpoints and aggregating them through search queries is only the first step to make LOD linkable and mappable semantically. In this presentation, we will discuss our efforts and strategies in mapping semi-structured, uncontrolled LOV vocabularies to an established KOS in the Linked Open Data space.

As a part of the project on meaningful concept displays for searching and browsing [7], we are investigating how to make better use of Getty's AAT thesaurus for searching and browsing on a collection of arts images that are not directly indexed by AAT. The collection, a special sub-collection of arts images from ARTstor, was indexed by a set of indexing terms which forms a semi-structured, uncontrolled vocabulary govern by its own internal rules. When described by RDF, the set of indexing terms is similar to a LOV vocabulary.

First, we developed a pattern-based technique for composition concept mapping. Since concepts in the AAT are primarily elemental and ARTstor indexing terms are pre-combined concept phrases, we need to decompose the concept phrases to elemental concepts and to possible sub-concepts (or conceptual components). Using linguistic

analysis and information extraction techniques [8,9], augmented with rules and some manual intervention, we were able to run a batch process to convert all the ARTstor concept phrases into concept components that can be mapped to AAT elemental concepts.

Second, we created a universal database structure that enables a database to store multiple KOS and maintain concept mapping relationships. An important design consideration of the database is how to distinguish concepts, terms, and strings in the database and during searching. Specialized APIs for the database was also created to facilitate use of pattern-based concept mapping results. During the presentation, we will particularly discuss how to maintain semantic mapping for the "Material" facet in AAT and how to map and store pattern-based mapping results in the database.

Finally, we developed a Web-based search interface that makes use of the concept mapping results. A meaningful concept display (MCD) was created to help the user in searching and browsing: Users can start with a free text query and the query will be mapped to AAT terms showed on MCD; the users then can interact with MCD to enrich the queries through AAT hierarchies and other concept relationships; the system will map these AAT queries into ARTstor terms before conducting the search. Such queries will have high precision and recall since they contain the terms that are actually used to index the collection by the search engine. During the presentation, we will demonstrate the interface prototype and show how thesaurus structures and semantic mapping will work behind-the-scenes to help the user construct and expand the queries for effective searching and browsing. We will discuss how meaningful concept displays can be used for KOS mapping as well.

References

1. LOV -- Linked Open Vocabularies. http://lov.okfn.org/dataset/lov/
2. Getty LOD -- http://www.getty.edu/research/tools/vocabularies/lod/
3. Library of Congress LOD -- http://id.loc.gov/
4. DDC LOD -- http://dewey.info/
5. Tennant, R. (2014). OCLC exposes bibliographic works data as linked open data. http://hangingtogether.org/?p=3614
6. Papadakis, I & Kyprianos, K. (2013). Merging controlled vocabularies through semantic alignment based on linked data. Metadata and Semantics Research, Communications in Computer and Information Science, v390, pp. 330-341.
7. Lin, X., Soergel, D., & Baca, M. (2012). Meaningful Concept Displays: The First Step. Paper presented at the 11th European NKOS Workshop at TPDL 2012, Paphos, Cyprus, Sept. 26-27, 2012.
8. Euzenat, J., & Shvaiko, P. (2013). Ontology Matching. 2nd ed. Heidelberg: Springer.
9. Giunchiglia, F., Yatskevich, M., & Shvaiko. P. (2007). Semantic Matching: Algorithms and Implementation. Journal on Data Semantics IX . Lecture Notes in Computer Science Volume 4601, pp. 1-38, 2007.