**Best Practice Demonstration:**
A Model and Roles of a Common Terminology to Improve Metadata Interoperability

(Boaz) Sunyoung Jin
University of Illinois at Urbana-Champaign, United States
sunjin@illinois.edu

**Keywords:** metadata; interoperability; common terminology; metadata model; MARC; (Q)DC

1. Introduction

Interoperability issues pose a barrier to sharing and exchanging information among digital libraries and repositories. This is due to the use of diverse metadata standards, and their different degrees of generality or specificity. This causes loss of information at all metadata model levels (e.g., schema, schema definition language, record, and repository) (Chan & Zeng, 2006) (Haslhofer & Klas, 2010, p. 19). As a possible solution for a long-term problem, historically argued standardization on a common communications format (Svenonius, 1983, p. 2), and a common command language or vocabulary (Lancaster & Smith, 1983, p. 21) are considered. A Common Terminology (CT), thus, is suggested as a bridge to various degrees’ metadata standards to give uniformity for searching and to achieve metadata interoperability at multiple levels.

2. The Abstract Model and Roles of a Common Terminology (CT)

Based on DCMI abstract model (DCMI, 2013), an abstract model of CT is diagrammed in Figure 1. The definitions for terms in this extended abstract model are as follows:

- A Common Terminology is a set of Common Terms of element names in widely used metadata schemas such as MARC, MODS, DC and QDC.
- A Common Term is a property (element) or class.
- A property (sub-property) can be one kind of common element (field) or attribute (subfield) in two or more metadata schemas.

![FIG. 1. The CT Abstract Model based on DCMI abstract model (DCMI, 2013)]
The core role of CT is to encompass various metadata schemas allowing communities to use their own standards, while providing uniformity to searching. CT is a bridge of existing standards to maintain balance between different degrees of generality or specificity, minimizing loss of information at all metadata model levels. CT is to provide uniformity for search with CT union catalog and Linked Open Data connecting online accessible metadata records on the Web. CT, ultimately, is to provide a common standard way to achieve interoperability at multiple levels in order to share resources readily among many libraries, organizations, and governments.

3. The Developed CT to Improve Metadata Interoperability

Taking commonly used standards (MARC, MODS, DC, and QDC) as bases, CT has developed as a bridge across different generality and specificity levels. CT is selected to improve metadata interoperability at the schema, schema definition language, record, and repository model levels.

3.1. At the Schema Metadata Model Level

The developed CT (Jin, 2014) is a set of 12 Common Terms (properties), and 58 qualifiers (sub-properties) that specify and subdivide 12 properties in detail, with CTScheme. CTScheme is defined as a controlled set of values that are specific to CT. The development bases on crosswalks of Library of Congress (e.g., MARC from/to (Q)DC, etc.) (LC). The development is supported by usages of MARC tags and (Q)DC elements in 5 search interfaces and in actual metadata records of Harvard (MARC, 12 million records), UIUC (MARCXML, 10 million), and MIT (QDC, 20,000) through cooperation of three universities in the USA. The selected CT at the schema level is generalized common terms which maximize lexical and semantic interoperability, used over 50% usage in Harvard, WorldCat and UIUC metadata records; and used in all 5 search interfaces. 12 Common Terms are contributor, date, description, format, identifier, language, publisher, relation, rights, subject, title, and typeGenre. 58 qualifiers are on the project website.

3.2. At the Schema Language Definition Level

The generalized 12 Common Terms and 58 qualifiers are represented with XML schema (ct.xsd) and RDF schema (ct.rdf) with SKOS concepts (ctskos.rdf) to improve semantic interoperability.

3.3. At the Record Level

The performance of CT in achieving and improving metadata interoperability is presented through empirical evaluations with Harvard (MARC), MIT (QDC), and UIUC (MARCXML) records through cooperation of three universities. A conversion with Python language is designed to convert (Q)DC of MIT records to CT, and to measure transfer rate and lexical and semantic match rates. As a result of the conversion of mapping experiments, total transfer rate from (Q)DC of MIT to CT is 99.9%. Lexical and semantic match rates are 98.7% and 100%. Loss of information rate is extremely lower as 0.00463%. CT, thus, maximizes lexical and semantic interoperability reducing significantly the gaps of different degrees of generality or specificity. Finally, CT minimizes considerably loss of information at multiple levels.

3.4. At the Repository Level

As a next step, a prototype is planned to achieve and improve metadata interoperability at repository level. The prototype will build CT union catalog and Linked Open Data connecting 3 million online accessible records of Harvard (MARC), MIT (QDC) and UIUC (MARCXML) libraries providing a portal for them. The prototype will demonstrate a certain solution to build interoperability globally with CT among libraries or Well-Designed Digital Libraries all over the world that will consist of International Open Public Digital Library (Jin, 2014).
Conclusion

The Common Terminology (CT) has developed as a bridge across different generality and specificity levels such as MARC, MODS, DC, and QDC. CT minimizes considerably loss of information reducing the gaps among them. CT increases significantly accuracy in mappings showing high lexical and semantic match rates. The planned prototype will build CT union catalog and Linked Open Data connecting records of three universities on the Web, and provide a portal for Harvard, MIT and UIUC libraries. CT will give an assured solution to achieve and improve interoperability among university libraries and further among libraries and organizations to work together and share information reducing loss of information at multiple metadata levels.

Acknowledgements

This research project has been supported by the Graduate School of Library and Information Science (GSLIS) at the University of Illinois at Urbana-Champaign (UIUC). Thanks to Professors David Dubin and Linda Smith for their guidance on this project. We also appreciate the cooperation of Harvard, MIT, and UIUC in providing their metadata for the project.

References