Encoding Library of Congress Subject Headings in SKOS: Authority Control for the Semantic Web

Corey A Harper
University of Oregon Libraries
Tel: +1 541 346 1854
Fax:+1 541 346 3485
charper@uoregon.edu

Abstract

This paper will explore using XSLT stylesheets to translate LCSH Authority Records from MARC/XML or MADS XML formats into RDF documents according to the SKOS project's Quick Guide to Publishing a Thesaurus on the Semantic Web. Creating an RDF Data Store that represents the content of LCSH will have tremendous long-term benefits in allowing a greater breadth of applications to make full use of the relationships between concepts provided by LC Subject Headings.

Keywords:
Controlled Vocabularies, Authority Control, Thesauri, SKOS, RDF

1. Introduction

For a number of years, members of the Semantic Web development community have been calling on librarians to become involved in Semantic Web activities. In his keynote presentation at Dublin Core 2004, Eric Miller, W3C Semantic Web Activity Lead, reiterated his request that the library community bring their rich experience to the table. The roles Miller envisions for libraries include exposing collections, sharing past experience, and "web'ifying" thesauri (1).

Ontologies and thesauri, published in web-friendly, RDF-based formats, are an integral part of the tool-kit that will be used to bring about the Semantic Web.

The recent W3C development of the Simple Knowledge Organization System (SKOS) RDF vocabulary provides a means for expressing concept schemes in RDF (2). As SKOS continues to gain momentum as the encoding of choice for controlled vocabularies, thesauri, subject heading lists, and classification schemes, it behooves the library community to begin providing its rich controlled vocabularies encoded in SKOS.

2. Library Controlled Vocabularies

Libraries have been engaged in the process of developing controlled lists of terms to be applied to resources during subject analysis for over a century. Library of Congress Subject Headings (LCSH), Dewey Decimal Classification (DDC), and Library of Congress Classification (LCC) are among the oldest and most well known of these systems. Others include The Art and Architecture Thesaurus (AAT) and the Ethnographic Thesaurus.

These classification schemes, thesauri, and lists of subject headings have been developed, refined, and added to by professional librarians over the course of many decades. The collective knowledge organization experience of thousands of librarians is represented by these 'concept schemes'.

LCSH is an example of a particularly mature and advanced concept scheme used in the library community and represents an area where Semantic Web development could be greatly enhanced by drawing on existing library vocabularies. Providing a machine-readable encoding of LCSH, using an emerging Web standard such as SKOS, will improve search engine results, expose relationships between and hierarchies among terms, and improve the effectiveness of metadata generation tools.

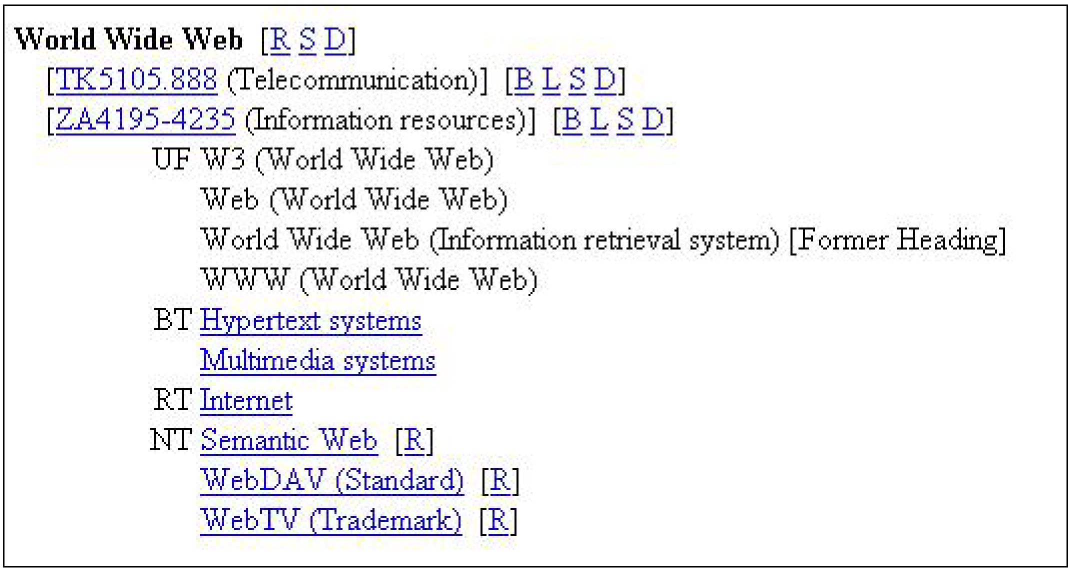
2.1 LCSH and Dublin Core Metadata

The Dublin Core Metadata Element Set provides for the use of Library of Congress Subject Headings as an encoding-scheme qualifier for the Subject element. However, most Dublin Core applications, aside from library bibliographic tools and utilities, are unable to make use of the rich syndetic structure provided by LCSH headings and cross-references.

As more Dublin Core metadata is represented in RDF encoding, it becomes increasingly important to provide an RDF-encoded version of LCSH to allow a wider array of tools to tap into the structure of LCSH headings. SKOS provides a model and syntax for encoding vocabularies like LCSH and is becoming widely recognized in the Semantic Web community as the best tool for this purpose. The SKOS community also provides guidance for how to publish a thesaurus on the Semantic Web (3).

3. LCSH and SKOS

While conversion to SKOS format will likely result in the loss of some of the finer details of LCSH that are represented in MARC 21 format, the conversion will still have a positive impact on the utility of Dublin Core records that make use of LCSH.



World Wide Web [R S D]
[TK5105.888 (Telecommunication)] [B L S D]
[ZA4195-4235 (Information resources)] [B L S D]
UF W3 (World Wide Web)
Web (World Wide Web)
World Wide Web (Information retrieval system) [Former Heading]
WWW (World Wide Web)
BT Hypertext systems
Multimedia systems
RT Internet
NT Semantic Web [R]
WebDAV (Standard) [R]
WebTV (Trademark) [R]

Figure 1: LCSH Record

Figure 1 shows an LCSH Record captured from the Library of Congress tool Classification Web. LCSH entries include 'Use For' type cross-references (UF), are associated with one another by means of Broader, Narrower and Related term entries (BT, NT, RT), and can be connected to classification schemes. Most of these structures are easily represented using SKOS. In this example, conversion to SKOS would result in minimal loss of information. The connection to an LC Classification number cannot be represented in SKOS because it draws on a separate concept scheme. Additionally, the concept of former heading, or former preferred label in SKOS parlance, cannot be easily represented using SKOS Core but could be represented using a SKOS extension.

3.1 Example Record in SKOS

The following code fragment shows how the record in Figure 1 can be encoded in SKOS.

```
<skos:Concept rdf:about="http://example.com/lcsh#95000541">
      <skos:prefLabel>World Wide Web</skos:prefLabel>
      <skos:altLabel>W3 (World Wide Web)</skos:altLabel>
      <skos:altLabel>Web (World Wide Web)</skos:altLabel>
      <skos:altLabel>World Wide Web (Information Retrieval System)</skos:altLabel>
      <skos:broader rdf:about="http://example.com/lcsh#88002671" />
      <skos:broader rdf:about="http://example.com/lcsh#92002381" />
       <skos:related rdf:about="http://example.com/lcsh#92002816"/>
       <skos:narrower rdf:about="http://example.com/lcsh#2002000569"/>
       <skos:narrower rdf:about="http://example.com/lcsh#2003001415"/>
       <skos:narrower rdf:about="http://example.com/lcsh#97003254"/>
</skos:Concept>
```

4. Transforming Records to SKOS

Subsets of LCSH exist in MARC 21 format in libraries around the world, and MARC records can easily be converted to MARCXML. Once in XML, one promising way to convert to SKOS is using an XSLT style-sheet.

4.1 XSLT Transformations

XSLT provides a convenient mechanism for transforming XML documents from one XML format to another. Since MARC and SKOS both have a defined XML syntax, XSLT appears to provide an excellent mechanism for generating a SKOS representation of Library of Congress Subject Headings.

A prototype style-sheet can be found at:
http://darkwing.uoregon.edu/~charper/MARC21slim2SKOS.xsl. The development of this style-sheet and testing against a set of 60,000 LCSH Authority Records has illustrated a number of problems with the XSLT approach.

The most significant of these problems is the fact that MARC Authority Records for LCSH do not include the narrower terms associated with a heading. This is done intentionally, to reduce duplication of effort, since references between records are established through the broader term information in the related record. This poses a problem for an XSLT based transformation: there is not a one-to-one correspondence between a MARCXML record and the target SKOS record. Instead, information must be culled from other records throughout the file that reference a particular record in order to supply the narrower term information. This cannot be easily accomplished using XSLT due to its reliance on XPATH to perform a transformation during a single navigation of the source XML tree.

A second problem arises from the way related, broader and narrower terms are referenced from a given concept's MARC record. Rather than using the identifier to reference related concepts, it uses the related concept's preferred label. This is done to maximize the human readability of the records, but does not translate well to SKOS.

The prototype style-sheet described above does not address the lack of narrower terms. It deals with the identifier versus label problem by using a url encoded form of preferred label to build URIs. While not an ideal approach, this does allow the relationships between concepts to be coded explicitly.

4.2 A Proposed Alternative Approach

Both of these two problems could be resolved through a non-XSLT based approach to record transformation. Rather than trying to translate a large, flat MARCXML file containing tens, or hundreds, of thousands of records directly to SKOS, a relational database could be used as an intermediate format.

Figure 2 shows a draft database schema for the proposed alternative approach. While this relational structure does not currently capture the full range of data contained in MARC authority records, it does include the data elements that are most useful in the context of SKOS records. A database developed to enable the translation of MARC authority data into SKOS could utilize a schema more complicated to the one diagramed here. The repeatable nature of many MARC data fields requires adding data tables for each major type of element included. In addition to non-repeatable MARC record data, this schema tracks LC class numbers, cross-references, use-for terms and notes.

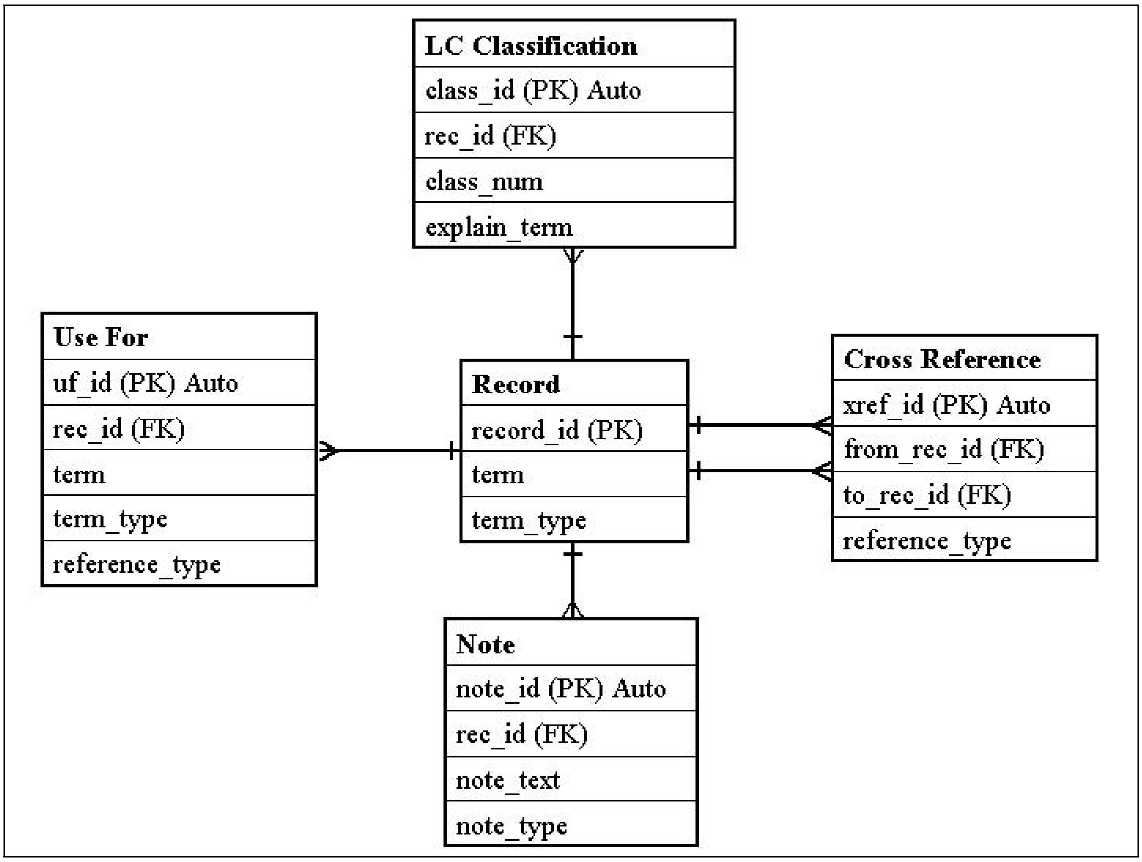


Figure 2: Sample Database Schema

The schema could easily be enhanced to include tables for complex subject references, series treatment information, and complex name references.

Multiple classification numbers, notes and use-for references can be represented using one-to-many relationships from the auxiliary tables to the Record table at the schema's core. Hierarchical and associative relationships between authorized terms are modeled using a recursive many-to-many relationship, which is represented in the cross-reference table. Additional pieces of information, such as term type (geographic, personal name, corporate name, topical subject, etc), reference type (broader, narrower, related) and note type (public, biographical, etc), are represented as attributes of the relevant table. While not all of this information can currently be expressed in SKOS, storing the data now will enable future enrichments using SKOS extensions or an alternate vocabulary representation.

Using this database schema, source records could be parsed and stored in a relational structure. A second process could build the SKOS records, querying for the record identifier of any concept referenced as a broader or related term in the source record. Additionally, the preferred label of a given source record could be searched against the database's broader term field to find other records that have the source record as a broader term and should be included as narrower terms in the target SKOS record. A more practical approach, that would improve the efficiency of database queries, would involve creating rows in the cross-reference table for narrower terms when records are parsed and ingested into the database. Other query processes could be developed for future vocabulary markup languages.

5. Summary

Library of Congress Subject Headings have the potential to help make the Semantic Web a reality and improve the usefulness of a variety of Semantic Web tools as well as Web based search and retrieval systems. Encoding LCSH using RDF vocabularies is a significant first step toward the realization of this potential, and SKOS is emerging as a prime candidate for the target syntax to convert to. While XSLT holds some promise for performing this transformation, initial testing has identified a number of possibly insurmountable challenges for an XSLT based approach. An alternative approach using a database as an intermediate storage mechanism to allow more complete conversion is being explored.

6. References

1. E. Miller. "The Semantic Web and Digital Libraries." From The International Conference on Dublin Core and Metadata Applications, 2004, Shanghai, China, 11-14 October 2004.
2. A. J. Miles and D. Brickley eds. SKOS Core Guide, W3C Working Draft, 15 February 2005. World Wide Web Consortium, 2005.
3. A. J. Miles, et al. Quick Guide to Publishing a Thesaurus on the Semantic Web: W3C Working Draft 17, May 2005. World Wibe Web Consortium, 2005.