Web Services for Genre Vocabularies

Diane Vizine-Goetz
OCLC Research
OCLC Online Computer Library Center, Inc.
Tel: +1-614-764-6000
Fax: +1-614-764-2344
vizine@oclc.org

Eric Childress
OCLC Research
OCLC Online Computer Library Center, Inc
Tel: +1-614-764-6000
Fax: +1-614-764-2344
eric_childress@oclc.org

Andrew Houghton
OCLC Research
OCLC Online Computer Library Center, Inc
Tel: +1-614-764-6000
Fax: +1-614-764-2344
houghton@oclc.org

Abstract:

This paper presents an approach for providing terminology Web services for controlled vocabulary terms. Services are implemented within a service oriented framework. A set of experimental services for genre vocabularies are provided through the MS Office Research pane, a built-in feature of Internet Explorer (IE) when users have loaded MS Office 2003. Web browsers, such as Mozilla Firefox and Opera, also provide side-bars which could be used to deliver loosely-coupled Web services.

Keywords:

Controlled vocabularies, DCMI Type Vocabulary, genre terms, OCLC Terminology Services Project, MS Office Research pane, Service Oriented Architecture (SOA), Web services.

1. Introduction

Along with subject, the genre of a creative work is one of the most important factors in determining if a resource will be used. In simple terms, genre means type or class. In the Dublin Core metadata element set, the Type element is used to describe the "nature or genre of the content of a resource" (1).

Popular book, movie, and music web sites routinely categorize resources by genre (2-4). At the highest level, literary works are commonly classified as fiction or nonfiction; musical works as classical or popular; and films as documentary or fiction. These categories are further divided into sub-genres which are often subdivided into deeper-level categories.

1.1. Genre Vocabularies

For resource description, there are many controlled vocabularies that can be used to describe the genre of a work. The DCMI Type Vocabulary is one example. It includes values such as, Image, Sound, and Text which can be used to describe the nature, or, mode of expression for a work. Catalogers of library materials can choose from more than 60 lists of controlled terms that have been approved for this purpose (5).

Recognizing the value of genre terms for resource discovery, OCLC researchers have prototyped a series of web services for genre vocabularies as part of the OCLC Terminology Services Project (6).

2.Terminology Services Project

The goal of the Terminology Services Project is to make the concepts in knowledge organization schemes more accessible to people and computer applications. Terminology services are Web services for various types of knowledge organization schemes, including subject heading systems, thesauri, and controlled lists of genre terms.

2.2 Encoding Genre Vocabularies

Before a Web service can be developed for a given knowledge organization scheme, it is often necessary to convert the concept data from word processing documents or HTML pages to more structured data formats, for example, the MARC 21 Format for Authority Data. MARC 21 was chosen for the project

because it includes mechanisms for coding common controlled vocabulary elements, such as preferred and non-preferred terms, term relationships, term mappings, and the source of the content and origin of changes (7). A snippet of the DCMI Type value 'Image' encoded in MARC XML is shown in Figure 1. Data field tag "040" subfield code "a" contains the MARC organization code for DCMI, the originator of the content; subfield code "c" contains the code for OCLC Research, the party responsible for converting the content to the MARC format. The genre term 'Image' is coded in tag "155" and the associated genre term 'Still Image' is coded in tag "555".

The SKOS core, an RDF schema for thesauri and related knowledge organization schemes, and the Zthes 0.5 schema are also suitable formats for encoding vocabulary resources. The Zthes 0.5 encoding for 'Image' is shown in Figure 2. The end products are XML files that can be used as the basis for terminology Web services

```
<datafield tag="040" ind1="" ind2="">
 <subfield code="a">OhDuDCMI</subfield>
 <subfield code="b">eng</subfield>
 <subfield code="c">OCoLC-O</subfield>
 <subfield code="d">OCoLC-O</subfield>
 <subfield code="f">dct</subfield>
</datafield>
<datafield tag="155" ind1="" ind2="">
 <subfield code="a">Image</subfield>
</datafield>
<datafield tag="555" ind1="" ind2="">
 <subfield code="w">h</subfield>
 <subfield code="a">Still Image</subfield>
   </datafield>
```

Figure 1 MARC encoding of DCMI Type value 'Image'

```
<term>
 <termId>DCT000004</termId>
 <termName>Image</termName>
 <termType>PT</termType>
 <termCreatedDate>
  2004-08-20
 </termCreatedDate>
 <termCreatedBy>
  OhDuDCMI
 </termCreatedBy>
 <termModifiedDate>
  2005-04-18T13:06:00.0
 </termModifiedDate>
 <termModifiedBy>
  OCoLC-O
 </termModifiedBy>
<relation>
  <relationType>NT</relationType>
  <termId>DCT000011</termId>
  <termName>Still Image</termName>
  <termType>PT</termType>
</relation>
```

Figure 2 Zthes 0.5 encoding of DCMI Type value 'Image'

3. Terminology Web Services

The implementation of Web services support in several widely-adopted platforms and applications presents an opportunity to offer terminology Web services in a variety of modular arrangements. OCLC Research is making a set of services for genre vocabularies available on an experimental basis. Services are provided through the MS Office Research pane, a built-in feature of Internet Explorer when users have loaded MS Office 2003. An illustration of the service architecture is shown in Figure 3.

Vocabularies can be stored as full text databases, SQL databases or XML files, and can be accessed with SRW/U, REST or SOAP based interfaces. Using MARC XML as input, OCLC's experimental implementation uses the OCLC Pears full text database software along with an SRW Web service interface to access the vocabularies. The terminology Web service acts as a proxy to the vocabularies providing query and markup translation along with authentication and authorization, when necessary.

The MS Office Research pane provides a Service Oriented Architecture (SOA) to Web services and integrated services to the application hosting it. The MS Office Research pane appears as a side-bar in most MS Office 2003 applications and Internet Explorer. OCLC's experimental terminology Web service takes advantage of the provided infrastructure to interact with Web based metadata editing applications.

OCLC's experimental terminology Web service is a service provider that is loosely-coupled to the MS Office Research pane service consumer. The functionality of the MS Office Research pane could be duplicated in other applications and environments. Web browsers, such as Mozilla Firefox and Opera, also provide side-bars which could be used to deliver loosely-coupled terminology Web services that interact with Web based metadata editing applications.

As a side-bar application, the MS Office Research Pane provides users the ability to conveniently interact with remote databases without interrupting their interaction with their main application (e.g., a metadata editing application in the main window in IE). Information retrieved in the Research Pane (e.g., a genre term) can be easily transferred into the main application.

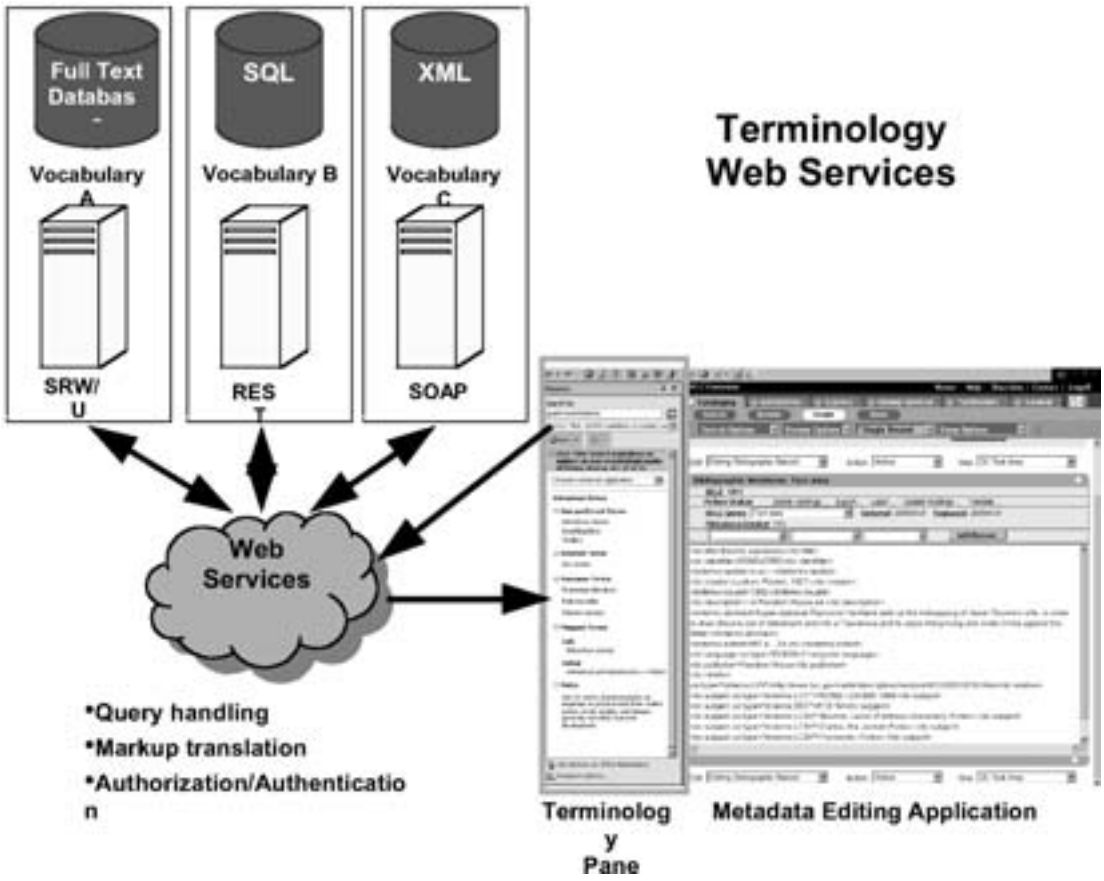

Figure 3 Terminology Services Architecture

4. Next Steps

Using a phased approach, OCLC is providing access to several genre vocabularies, beginning with GSAFD, a list of genre terms for fiction and drama, the DCMI Type Vocabulary, and the Newspaper Genre List. Additional genre vocabularies and thesauri will be offered in later stages of the project.

References

1. Dublin Core Metadata Element Set, Version 1.1: Reference Description. 20 Dec. 2004. Dublin Core Metadata Initiative. 18 Apr. 2005 <http://dublincore.org/documents/dces/>.
2. Book Browser. 18 Apr. 2005 <http://www.barnesandnoble.com/subjects/subjects.asp>.
3. allmusic explore by genre. 2005. All Media Guide. 18 Apr. 2005 <http://www.allmusic.com/cg/amg.dll?p=amg&sql=73:>.
4. Dirks, Tim . Film Genres. 1996-2005. 18 Apr. 2005 <http://www.filmsite.org/genres.html>.
5. MARC Code List: Part IV: Term, Name, Title Sources. 09 Feb. 2005. Library of Congress. 18 Apr. 2005 <http://www.loc.gov/marc/relators/relasour.html#rela655b>.
6. Terminology Services. 09 Mar. 2005. OCLC Research. 18 Apr. 2005 <http://www.oclc.org/research/projects/termservices/>.
7. Vizine-Goetz, Diane, et al. "Vocabulary Mapping for Terminology Services." Journal of Digital Information 4 (2004). 18 Apr. 2005 <http://jodi.ecs.soton.ac.uk/Articles/v04/i04/Vizine-Goetz/>.

roject-Related Standards

DCMI Type Vocabulary
http://dublincore.org/documents/dcmi-terms/#H5

GSAFD — Guidelines on Subject Access to Individual Works of Fiction, Drama, etc.

http://www.oclc.org/research/projects/termservices/resources/gsafd.htm

MARC 21 Format for Authority Data
http://www.loc.gov/marc/authority/ecadhome.html

MARC 21 XML Schema
http://www.loc.gov/standards/marcxml/

Newspaper Genre List
http://www.lib.washington.edu/mcnews/ngl/

OCLC Pears Database Software
http://www.oclc.org/research/software/pears/

SKOS Core Guide
http://www.w3.org/2004/02/skos/core/guide/

SRW/U: Search/Retrieve for the Web
http://www.loc.gov/z3950/agency/zing/srw/

Zthes: a Z39.50 Profile for Thesaurus Navigation, version 0.5 http://zthes.z3950.org/profile/zthes-05.html