Building a MARC-to-OLAC Crosswalk: Repurposing Library Catalog Data for the Language Resources Community

Christopher Hirt, Gary Simons, and Joan Spanne
SIL International and Graduate Institute of Applied Linguistics
7500 W. Camp Wisdom Rd.
Dallas, TX 75236

{chris_hirt, gary_simons, joan_spanne}@sil.org

ABSTRACT

The Open Language Archives Community (OLAC) is an international partnership of institutions which are building a network of interoperating repositories and services to create a worldwide virtual library of language resources (that is, resources that document, describe, or develop the more than 7,000 known languages of the world). OLAC uses a community-specific refinement of qualified Dublin Core [http://www.language-archives.org/OLAC/metadata.htm] along with a community-specific refinement of the OAI Protocol for Metadata Harvesting [http://www.language-archives.org/OLAC/repositories.htm] to maintain an aggregated catalog of the holdings of the 35 participating archives. OLAC recognizes that the language resources of interest to the community come not only from sources within the community but also from many sources outside the community. This poster describes one approach we have developed for addressing this issue, namely, a crosswalk that transforms the MARC21 catalog for a library or archive into an OAI static repository that holds an OLAC metadata record for each MARC record identified as describing a language resource.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Information filtering; H.3.7 [**Digital Libraries**]: Standards.

General Terms

Design, Experimentation, Standardization.

1. THE DESIGN OF THE CROSSWALK

The MARC-to-OLAC crosswalk adds two tasks to a general MARC-to-DC crosswalk. First, we require a filtering step to determine whether a record describes a language resource. Second, if it does, we must extract the specialized metadata that facilitates language resource discovery. A fairly narrow subset of Library of Congress Subject Headings (LCSH) is used in cataloging language resources. The simple existence of the word "language" in subfield $a of an entry element is the main clue; institution-specific cataloging policies may also prove relevant. If it contains one or more of these clues, a record passes through the filter stage. Then a transformation stage applies specialized mappings for identifying languages and the type of a language resource. OLAC metadata uses ISO 639-2/3 [http://www.sil.org/iso639-3/] as an encoding scheme for precise identification of languages We have thus mapped LCSH "language names" to their corresponding ISO 639 identifiers for use with DC:Language and OLAC's "language as subject" extension of DC:Subject. Specific LCSH terms are also mapped to OLAC's linguistic resource type vocabulary applied as an extension to DC:Type.

2. IMPLEMENTING THE CROSSWALK

The crosswalk is implemented with a series of XSL transformations driven by a Python wrapper script. The Python script applies the transformations in batches and is thereby able to process gigabytes of MARC data with minimal memory requirements. The preliminary filtering step is configured by specifying tests on any MARC field or subfield with simple string comparison. We use a two-stage filtering process in which the first stage (the "select stage") uses tests to select the maximal set of MARC records to be considered for inclusion in the OLAC repository, while the second stage (the "reject stage") uses tests to reject records that are in fact not wanted. The resulting record set is then transformed from MARC XML to OLAC's XML-based format by means of an XSL stylesheet that defines an <xsl:template> for each MARC field. An optional stylesheet is called to apply overriding institution-specific mappings. The complete source code and data tables for the crosswalk (including the LCSH to ISO 639-3 mappings) are being shared on an open-source basis.

3. RESULTS

The poster will report the results of applying the crosswalk to the two major data sets we have used during the development of the system: The first is the 32,000-record catalog of the Graduate Institute of Applied Linguistics library; roughly 20% of those records are being identified as in scope for OLAC. The second is a 5,000-record collection supplied by the National Anthropological Archives of the Smithsonian Institution; most, but not all, of those records are in scope for OLAC.

Copyright is held by the author/owner(s).
JCDL'09, June 15–19, 2009, Austin, Texas, USA.
ACM 978-1-60558-322-8/09/06.