Leveraging language codes in a stylesheet transformation: OLAC (ISO 639-3) into MARC

John Huck, University of Alberta Libraries (john.huck@ualberta.ca)

ABSTRACT

At **University of Alberta Libraries (UAL)**, the ability to generate MARC records from existing metadata using a stylesheet transformation was a prerequisite for a recent plan to improve discovery and access for a collection of data resources. This poster shows how ISO 639-3 language codes in the source metadata were transformed into LC subject headings and MARC language encodings for approximately 750 records.

PROJECT BACKGROUND

A long-held subscription to datasets published by the **Linguistic Data Consortium (LDC)** presented a cataloguing problem for staff and access difficulties for users: most datasets were available as a download and on a physical format, but the download access was difficult to catalogue and some datasets were only available in physical format. Consequently, there was no single place to discover which datasets the library had purchased, and the steps users had to take to get the data were not straightforward.

The development of a new access model for these resources was driven by three factors:

- a desire to simplify the processes of discovery and access for the end-user;
- a determination that the download platform (LDC Catalog) was not suitable for end-users, in part because it included administrative functions, like ordering and license signing; and
- a new commitment to purchase any LDC dataset the libraries did not already own, when a user requested it, like a purchase-on-demand plan.

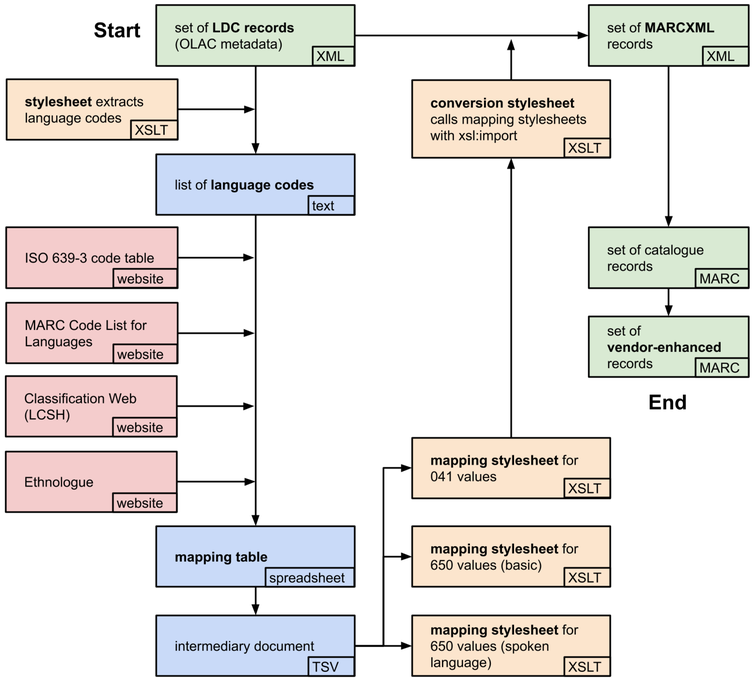
In the new, **mediated access model**, a user could:

- search across all published LDC datasets in the library catalogue;
- request access to a dataset through a web form linked from the catalogue record; and
- have it delivered to them via Google Drive.

Library staff processing the request would determine whether to:

- download the data from the online platform,
- copy it from a physical carrier, or,
- if necessary, initiate a purchase order for it.

Underpinning this new model would be a set of catalogue (MARC) records representing all published LDC datasets, generated from metadata that LDC shares through the **Open Language Archives Community (OLAC)**.



OLAC METADATA

```
<olac:olac xmlns:olac="http://www.language-archives.org/OLAC/1.1/"
        xmlns:dc="http://purl.org/dc/elements/1.1/"
        xmlns:dcterms="http://purl.org/dc/terms/"
        xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance">
        <dc:contributor>Du Bois, John W.</dc:contributor>
        <dc:contributor>Chafe, Wallace L.</dc:contributor>
        <dc:contributor>Meyer, Charles</dc:contributor>
        <dc:contributor>Thompson, Sandra A.</dc:contributor>
        <dc:date xsi:type="dcterms:W3CDTF">2000</dc:date>
        <dcterms:issued xsi:type="dcterms:W3CDTF"
                >2000-01-01</dcterms:issued>
        <dc:description>*Introduction* The Santa Barbara Corpus of
                Spoken American English is based on hundreds of
                recordings of natural speech from all over the
                United States, representing a wide variety of
                data in this corpus, please examine these samples
                of the recordings and transcripts: * Speech *
                Transcripts *Updates* There are no updates at this
                time.</dc:description>
        <dcterms:extent>Corpus size: 1677721 KB</dcterms:extent>
        <dcterms:medium>Distribution: Web Download</dcterms:medium>
        <dc:identifier>LDC2000S85</dc:identifier>
        <dc:identifier>https://catalog.ldc.upenn.edu/LDC2000S85</dc:identifier>
        <dc:identifier>ISBN: 1-58563-164-7</dc:identifier>
        <dc:identifier>ISLRN: 407-731-819-668-4</dc:identifier>
        <dcterms:bibliographicCitation>John W. Du Bois, et al.
                Santa Barbara Corpus of Spoken American English
                Part I LDC2000S85. Web Download. Philadelphia:
                Linguistic Data Consortium,
                2000</dcterms:bibliographicCitation>
        <dc:language olac:code="eng" xsi:type="olac:language"
                >English</dc:language>
        <dc:publisher>Linguistic Data Consortium</dc:publisher>
        <dc:publisher xsi:type="dcterms:URI"
                >https://www.ldc.upenn.edu</dc:publisher>
        <dc:relation xsi:type="dcterms:URI"
                >https://catalog.ldc.upenn.edu/docs/LDC2000S85</dc:relation>
        <dcterms:accessRights>Licensing Instructions for
                Subscription &amp;amp; Standard Members, and
                Non-Members:
                http://www.ldc.upenn.edu/language-resources/data/obtaining
        </dcterms:accessRights>
        <dcterms:license>LDC User Agreement for Non-Members:
                https://catalog.ldc.upenn.edu/license/ldc-non-members-agreement.pdf
        </dcterms:license>
        <dc:title>Santa Barbara Corpus of Spoken American English
                Part I</dc:title>
        <dc:type olac:code="primary_text"
                xsi:type="olac:linguistic-type"/>
        <dc:type xsi:type="dcterms:DCMIType">Sound</dc:type>
</olac:olac>
```

MARC RECORD

```
LEADER 03337nmm a22005173i 4500
001 8587995
006 m o u
007 cu |||||||u|||
008 190313s2000 pau u eng d
020    a| 1585631647
020    a| 9781585631643
024 8  a| LDC2000S85
024 8  a| 4077318196684 q| ISLRN
035    a| on1090038764
039    a| exclude
040    a| AEU b| eng e| rda c| AEU d| AEU
042    a| dc
043    a| n-us---
050  4 a| PE2808.8 b| .S26 2000
090    a| Internet Access b| AEU
245 0 0 a| Santa Barbara Corpus of Spoken American English Part I.
264  1 a| [Philadelphia, Pennsylvania] : b| Linguistic Data Consortium, c| [2000]
300    a| 1 online resource.
336    a| computer dataset b| cod 2| rdacontent
336    a| spoken word b| spw 2| rdacontent
337    a| computer b| c 2| rdamedia
338    a| unspecified b| zu 2| rdacarrier
500    a| LDC number: LDC2000S85.
500    a| Data samples are available on the LDC website.
506 1  a| Access restricted to authorized users and institutions.
520    a| The Santa Barbara Corpus of Spoken American English is based on hundreds of recordings of natural speech from all over the United States, representing a wide variety of people of different regional origins, ages, occupations, and ethnic and social backgrounds. It reflects
546    a| Content and documentation in English.
596    a| 44
650  0 a| English language x| Spoken English z| United States.
650  0 a| English language x| Variation z| United States.
650  0 a| English language x| Data processing v| Databases.
650  0 a| English language x| Spoken English x| Data processing v| Databases.
655  7 a| Sound recordings. 2| lcgft
700 1  a| Du Bois, John W.
700 1  a| Chafe, Wallace L.
700 1  a| Meyer, Charles.
700 1  a| Thompson, Sandra A.
856 4 0 3| University of Alberta Access u|
        https://docs.google.com/forms/d/e/1FAIpQLSd4VsEYOWoubQww-
        01W7IV2qDaAr4ctBJUhrJvfyN0GwoMuFQ/viewform z| Request Form
856 4 2 3| Dataset documentation u| https://catalog.ldc.upenn.edu/LDC2000S85
949    h| SUAIN z| LDC
926    a| Internet Access w| LC c| 1 i| 8587995-1001 l| INTERNET m| UAINTERNET r| Y s| Y t| E-RESOURCE u| 3/15/2019 z| LDC
```

CODE MAPPING

LDC metadata includes comprehensive coding of languages using ISO 639-3 codes. MARC language codes are aligned with ISO 639-2 codes, and MARC language names are aligned with Library of Congress Subject Headings (LCSH). These family relationships formed the basis of the code mapping.

In MARC records, language codes are used in the 041 field and the 008/35-37 fixed field. Language names are used in subject headings (650 field) and sometimes in notes fields (5XX fields).

ISO 639-3 to MARC (types of matches):

- Exact code match
- match with a **639-2 synonym**
- Match to a broader **Macrolanguage**
- Best fit based on language name, with reference to Ethnologue website

MARC to LCSH

- language names already aligned
- Classification Web used for verification
- standard heading constructed for each language
- additional heading constructed for spoken language datasets
- Backstage Library Works (library vendor) provided additional subject analysis

LCSH EXAMPLES
English language—Data processing—Databases.
English language—Spoken English—Data processing—Databases.

LANGUAGE CODES

While the **OLAC Language Extension** permits the use of codes from all three parts of **ISO 639**, OLAC normalizes codes to Part 3 in metadata it distributes through its OAI-PMH service. ISO 639-3 is maintained by SIL International.

The Library of Congress (LC) maintains both **ISO 639-2** and the **MARC Code List for Languages** on which it was based. The codes are the same, but the language names differ, because the MARC names have been aligned with terms from LC Subject Headings.

When LC developed **ISO 639-2**, it introduced **synonyms** for certain language codes, apparently to reconcile established library practice with the 2-letter codes found in ISO 639-1. "Bibliographic" (B) codes are the same as the MARC codes and based on language names in English. "Terminological" (T) codes are aligned with 639-1 codes. For example, 'per' (B) and 'fas' (T) are both valid codes for Persian.

ISO 639-3 incorporates all the 639-2 codes and expands the set considerably. In some cases – such as the family of Chinese languages – 639-3 introduced several new codes for languages that were represented by a single code in 639-2 and designated the 639-2 code a **Macrolanguage** in order to maintain backward compatibility. Macrolanguages increase the number of matches between 639-3 and MARC codes.

STYLESHEETS

- 101 ISO 639-3 language codes are mapped.
- Mapping notes are included in comments.
- When two or more 639-3 codes in a record map to a single 639-2 code, the 639-2 code is only added once.
- When the data and documentation are in English, the 041 is not added.

REFERENCES & LINKS

https://bit.ly/2UprJdy

THANK YOU

I would like to thank the U of A staff who contributed to this project: Anna Bombak (Data Team), Elizabeth Wallace (CSU), Shellon Miller (CSU), Céline Gareau-Brennan, Amanda Nagyl (Bib Services), Ian Bigelow (Bib Services), and anyone else I may have overlooked. Amanda and I worked together to refine the MARC output.

UNIVERSITY OF ALBERTA