9-21-2010

# Why Purchase When You Can Repurpose? Using Crosswalks to Enhance User Access

Teressa M. Keenan
*University of Montana - Missoula*, teressa.keenan@umontana.edu

# Why Purchase When You Can Repurpose? Using Crosswalks to Enhance User Access

## Abstract

The Mansfield Library subscribes to the Readex database *U.S. Congressional Serial Set, 1817-1994* (full-text historic reports of Congress and federal agencies). Given the option of purchasing MARC records for all 262,000 publications in the Serial Set or making use of free access to simple Dublin Core records provided by Readex, the library opted to repurpose the free metadata. The process that the Mansfield Library used to obtain the Dublin Core records is described, including the procedures for crosswalking the metadata to MARC21 and batch loading the bibliographic records complete with holdings information to the local catalog. This report shows that we successfully achieved our goals of dramatically increasing access to Serial Set material by exposing metadata in the local catalog and discusses the challenges we faced along the way. We hope that others tasked with the manipulation of metadata will be able to use what we learned from this project.

**Keywords:** metadata; crosswalks; repurposing data; MARC21; Dublin Core; MARCEdit.

## 1. Introduction

The University of Montan-Missoula (UM) is a doctoral institution serving a student population of approximately 12,000 undergraduates and 2,000 graduate students. Faculty and staff comprise a population of over 1,500. UM is part of the Montana University System with three affiliated campuses in Dillon, Helena, and Butte. The Maureen and Mike Mansfield Library (ML) holds the largest collection of books and media in Montana. ML's collections exceed one million volumes, 125,000 maps, 100,000 photographs, 50,000 media items, over 30,000 print and electronic journals, and access to over 300 databases. However, like many institutions its size, the library has a limited budget and large expenditures must be carefully considered. Thus when the library was faced with the option of purchasing vs. repurposing metadata to augment the catalog, they chose repurposing.

In January 2007 ML subscribed to the Readex database *U.S. Congressional Serial Set, 1817-1994,* a collection of full-text historic reports of Congress and federal agencies. Access to the database was provided through a database A-Z list and a subject-based Libguide. When Readex made bibliographic records available to subscribers, the library decided to add these to the local catalog in order to increase access to the database. Readex offered subscribers the option of purchasing a set of MARC format records for $25,000 or receiving a set of Dublin Core (DC) format records at no additional charge. The addition of a metadata librarian to ML's technical services unit in September 2008 provided the personnel resource needed to repurpose the free metadata.

## 2. The project

The Serial Set project was initiated in order to achieve a cost-effective way of increasing patron access to and use of the Serial Set database. Initial concerns centered on the quality of the bibliographic records and our ability to successfully crosswalk a less granular metadata schema such as Dublin Core into the more granular and structured rules of AACR2 and MARC. A test was run on a small set of records; the data was collected, crosswalked and then loaded into the library's test database. The results were then compared to the MARC records available for purchase (See Appendix A). After consultations between project team members, it was determined that the metadata in the DC records provided by Readex was compatible with the standards set for the library's bibliographic catalog. While it was acknowledged that the MARC records available for purchase more closely followed the conventions of AACR2 than the crosswalked records and that cataloging each title individually following AACR2 would produce more robust and complete bibliographic records, we felt that the information provided from the DC records would display appropriately in our catalog, would be sufficient for our goals and would enhance our users' ability to find, identify and access these resources. The project was broken down into three tasks:

1. collect the available Dublin Core records,
2. transform the XML into a compatible format, and
3. load the metadata into the library's Integrated Library System (ILS).

### 2.1. Collecting the data

The first available segment of Simple Dublin Core records provided by Readex included data for every publication contained in Serial Set Volumes 1 to 5377, approximately 168,800 records. The records were harvested over OAI-PMH with MarcEdit, a Windows-based freeware MARC editing tool already in use at the library (Reese, n.d.).

The interface was straightforward once the correct server address was obtained from Readex. The download process took approximately 10 hours. When complete, 6827 XML files, each containing data for approximately 40 individual titles, had successfully downloaded.

### 2.2. Transforming the data

The first step in transforming the metadata into a form that was ingestible by the library's Voyager catalog was to create a crosswalk from Dublin Core to MARC21. Existing DC to MARC crosswalks (LC, 2008; Dutta, 2003) were compared with the indexing capabilities of the local ILS and a working crosswalk was created specifically for this project. (See Appendix B)

We edited the existing XSLT in MarcEdit to match our project crosswalk specifications, but stumbled in our initial attempts to edit the script due to lack of experience with XML and XSLT. For example, we could not separate the year from the rest of the date field in the Dublin Core record until we learned about the substring-before function. Another issue we had with the bibliographic records after running the XSLT script was the inclusion of unwanted whitespace within the subfields in the 1XX and 7XX MARC fields. While the whitespace did not show in the public display of the catalog and did not appear to affect searching it was visible in the staff side display in the cataloging module of Voyager. We were unable to remove the whitespace with XSLT, but found success with find and replace in MarcEdit, so we added that step to its automated tasks. The following example contains extra whitespace in the date subfield:

```
710  10‡aU.S. Congress. House. Committee on the Judiciary ‡d          (1813- ).
```

Here is the same field, with the whitespace removed:

```
710  10‡aU.S. Congress. House. Committee on the Judiciary ‡d(1813- ).
```

Additional bibliographic data that was not included in the DC record was incorporated into the final MARC record through the XSLT script the batch editing features in the MarcEdit program. Additional data included:

- the government documents classification number stem (=086 0\ ‡a Y 1.1/2:),
- an authentication code to indicate that the record was converted from a simple resource description record in another syntax using the Dublin Core metadata

element set, and that the content of the record (descriptive elements and headings) may or may not follow any cataloging standard (=042 \\ ‡a dc)

- a reproduction note indicating that the material was provided by Readex and has restricted access (=533 \\‡a Electronic Reproduction ‡b Chester, VT. ‡c Newsbank, inc. ‡d 2005. ‡n Available via the World Wide Web ‡n Access restricted to Readex U.S. Congressional Serial Set subscribers)
- a local note indicating that the same material is also available in print (=590 \\‡a ML: This title is also available in print. See the "United States congressional serial set", Call number Y 1.1/2:)
- a public note in the electronic location and access field (=856 40 ‡z Connect to this title online.)
- holdings/item record information (=949 \\ ‡c uonetlib ‡d Y 1.1/2: ‡t useminet)

See Appendix C for the full XSLT stylesheet

### 2.3. Batch loading the data

In an effort to reduce the overall time and manual interaction with the transformation, editing and batch load process, we wrote a script to combine the numerous records into six batch files. Once the data were transformed, Voyager batch loading protocols were used to add the MARC bibliographic, holdings and item records to the catalog. A load profile was established on Voyager that defined the load parameters, including expected character set, fields to match on, how to handle matches (e.g., replace or merge), and location of holdings information. The batch files were loaded using parameters on the server, and system logs were reviewed for problems or errors upon completion.

Each of these batch files, roughly 64MB in size, required 18 hours to finish loading. Therefore, the timing of the load became an important consideration. The first file load was interrupted by our routine daily system back-up. Once this collision was discovered, further file loads were timed to not coincide with the back-up process. As the process of loading the files continued, we discovered that each subsequent file took longer to load. We decreased the processing time somewhat by loading the files without keyword indexing, then re-indexed after the project was completed.

In order to minimize load time in the future, records will not be processed with the above script. Instead, we'll use the batch edit feature of MarcEdit to combine the records into a number of smaller batches. This revised procedure will involve less manual interaction while avoiding the problems we encountered with very large files.

## 3. Project results (cost effectiveness & user access)

When completed, over 262,000 bibliographic, holding, and item records were added to the library's catalog. The process took approximately 46 staff hours and just over 260 computer hours to complete. Based on these statistics, repurposing of the metadata allowed the database to be enriched by the addition of about 856 titles per hour.

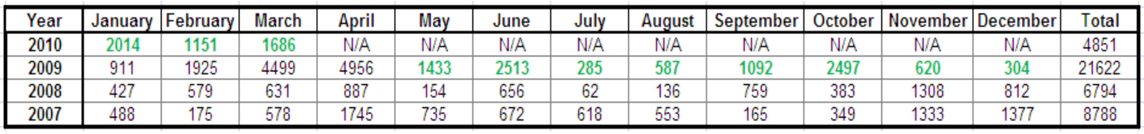
While the library did not have any initial purchase costs associated with the project (the metadata was freely available to database subscribers, and the software used was either open source or already in use for other applications), there were costs involved (Table 1). Most of the staff time was spent in planning, research, and the development of the metadata transformation scripts. Staff time estimates include salary and benefits. Approximations for computer time and other overhead costs such as electricity, etc. were not included in this analysis.

TABLE 1: Cost comparison chart

| | Repurpose | Purchase |
|---|---|---|
| **Data Acquisition** | $0.00 | $25,000.00 |
| **Staff Time (Systems Support)** | $614.80 | $412.46 |
| **Staff Time (Metadata Librarian)** | $514.25 | $257.25 |
| **Total Estimated Cost** | $1,129.05 | $25,669.71 |

System supplied statistics for use of the Serial Set database show an overall increase in use since the library subscribed in January 2007 (Table 2). Overall usage increased by 2,577 views in the nine months following the addition of individual title records to the catalog. The addition of bibliographic records to the library catalog is not the only factor affecting use of the database. For example, reference librarians include that database in their research classes. Also, in April 2009 the library celebrated its 100th year as a Federal Depository, an event which helped focus attention on this database. Initiating a better method of tracking use (including how users got to the database, rather than just the number of times a database was viewed), would provide better analysis opportunities for determining success of future projects intended to increase access to collections.

TABLE 2: Use statistics



| Year | January | February | March | April | May | June | July | August | September | October | November | December | Total |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 2010 | 2014 | 1151 | 1686 | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A | 4851 |
| 2009 | 911 | 1925 | 4499 | 4956 | 1433 | 2513 | 285 | 587 | 1092 | 2497 | 620 | 304 | 21622 |
| 2008 | 427 | 579 | 631 | 887 | 154 | 656 | 62 | 136 | 759 | 383 | 1308 | 812 | 6794 |
| 2007 | 488 | 175 | 578 | 1745 | 735 | 672 | 618 | 553 | 165 | 349 | 1333 | 1377 | 8788 |

## 4. Conclusions and future plans

The repurposing of the pre-existing metadata to increase access to digital material was an overall success. Repurposing the metadata proved to be a cost effective alternative to purchasing MARC records, saving the library over $24,500. Use statistics show a marked increase in database usage after the records were added to the catalog.

Because this project was the first of its kind at our library, the project team gained valuable knowledge that can be utilized in the future. Because of inexperience with XSLT and scripting in general, creating the crosswalk took a bit of trial and error. Additionally, the MarcEdit software was new to members of the project team. The reference section of this report includes materials that were helpful in learning the basics of XSLT and MarcEdit.

According to the Readex webpage, additional Dublin Core records will be available each month. (Readex, n.d.) The library plans to selectively harvest the new records, crosswalk them to MARC, and add the data to our catalog using the workflows established in this project. Future research plans include enhancing our understanding of XSLT and consideration of additional ways to repurpose existing metadata to enhance our users' experience.

## References

Dutta, B. (2003). Cataloguing Web Documents using Dublin Core, MARC 21. Presented at the Workshop on Digital Libraries: Theory and Practice, DRTC, Bangalore.

Library of Congress, Network Development and MARC Standards Office. (2008). Dublin Core to MARC Crosswalk. Retrieved March 30, 2010, from http://www.loc.gov/marc/dccross.html

Open Archives Initiative - Protocol for Metadata Harvesting - v.2.0. (n.d.). . Retrieved March 30, 2010, from http://www.openarchives.org/OAI/openarchivesprotocol.html

Otegem, M. (2002). Sams teach yourself XSLT in 21 days. Indianapolis  IN: Sams.

Readex. (n.d.). MARC Records. Retrieved March 31, 2010, from http://www.readex.com/readex/index.cfm?content=296

Reese, T. (2009a). YouTube - Add New Metadata Function. Retrieved March 31, 2010, from http://www.youtube.com/watch?v=3x5Ke81AoEU

Reese, T. (2009b). YouTube - Translating OAI metadata to MARC using MarcEdit. Retrieved March 31, 2010, from http://www.youtube.com/watch?v=gvBrMVH6j7U

Reese, T. (n.d.). MarcEdit Homepage: Your Complete Free MARC Software. MarcEdit. Retrieved March 30, 2010, from http://people.oregonstate.edu/~reeset/marcedit/html/index.php

## Appendix A : Metadata Comparison

The following example of the Simple Dublin Core and MARC records available through Readex (Readex, n.d.) followed by the Crosswalked and Edited MARC record created by ML demonstrates the similarities and differences in the granularity of the MARC records.

### Simple Dublin Core, Serial Set

```
- <record>
- <header>
<identifier>oai:docs.newsbank.com/sset.0FDBE2664851D648</identifier>
<datestamp>1817-12-08</datestamp>
<setSpec>sset</setSpec>
</header>
- <metadata>
-<oai_dc:dc
xmlns:oai_dc=”http://www.openarchives.org/OAI/2.0/oai_dc/”
xsi:schemaLocation=”http://www.openarchives.org/OAI/2.0/oai_dc/
http://www.openarchives.org/OAI/2.0/oai_dc.xsd”>
<dc:type> Motions and Resolutions</dc:type>
<dc:date>1817-12-08</dc:date>
<dc:coverage>Congressional Session Number 15th Congress, 1st Session; Session Volume Number
1</dc:coverage>
<dc:description>Senate Document</dc:description>
<dc:source>Serial Set Number 2 S.Doc. 3</dc:source>
<dc:identifier>http://docs.newsbank.com/select/serialset/0FDBE2664851D648.html
</dc:identifier>
<dc:language>English</dc:language>
<dc:title>In Senate of the United States, December 8, 1817. Mr. Sanford submitted the following
motion for consideration: Resolved, that the Committee of Finance inquire what alterations or
amendments may be requisite in the present system of collecting the duties charged...</dc:title>
<dc:creator>Sanford, Nathan, 1777-1838, Republican-Jeffersonian (NY)</dc:creator>
<dc:creator>U.S. Congress. Senate</dc:creator>
<dc:subject>Customs administration</dc:subject>
<dc:subject>Foreign trade</dc:subject>
<dc:subject>Imports</dc:subject>
<dc:subject>Tariffs and duties</dc:subject>
<dc:description>1 p.</dc:description>
</oai_dc:dc>
</metadata>
</record>
</GetRecord>
</OAI-PMH>
```

## Readex MARC Records

```
=LDR 01911cam 22003371i 4500
=001 NB00000000008
=003 Readex
=005 20090622155353.1
=006 m\\\\\\\\d
=007 cr\cn\||||||||
=008 070221s1817\\\\dcu\\\\\\s\\\f000\0\eng\d
=035 \\‡a(Readex)0FC9A3C9017886A0
=040 \\‡aReadex ‡c Readex
=110 1\‡aUnited States. ‡b Congress. ‡b Senate.
=245 10‡aIn Senate of the United States, December 8, 1817. Mr. Sanford submitted the following
motion for consideration: Resolved, that the Committee of Finance inquire what alterations or
amendments may be requisite in the present system of collecting the duties charged... ‡h [electronic
resource]
=260 \\‡aWashington, DC, ‡c 1817
=300 \\‡a1 p.
=440 \0‡aUnited States congressional serial set; ‡v serial set no. 2
=490 1\‡aSenate document / 15th Congress, 1st session. Senate ; ‡v no. 3
=500 \\‡aTitle taken from opening lines of text.
=533 \\‡aElectronic reproduction. ‡b Chester, Vt.: ‡c NewsBank, inc., ‡d 2005. ‡n Available via the
World Wide Web. ‡nAccess restricted to Readex U.S. Congressional Serial Set subscribers.
=540 \\‡aCopyright 2007 by NewsBank, Inc. All rights reserved.
=610 17‡aUnited States. ‡b Congress. ‡b Senate. ‡b Committee on Finance. ‡g (1816- ) ‡2 Readex
congressional thesaurus
=650 07‡aCustoms administration. ‡2 Readex congressional thesaurus
=650 07‡aForeign trade. ‡2 Readex congressional thesaurus
=650 07‡aImports. ‡2 Readex congressional thesaurus
=650 07‡aTariffs and duties. ‡2 Readex congressional thesaurus
=655 \7‡aMotions and Resolutions. ‡2 Readex congressional thesaurus
=700 1\‡aSanford, Nathan, ‡d 1777-1838. ‡u Republican-Jeffersonian (NY)
=830 \0‡aSenate document (United States. Congress. Senate) ; ‡v 15th Congress, 1st session, no. 3
=856 40‡uhttp://docs.newsbank.com/select/serialset/0FDBE2664851D648.html
```

## Crosswalked & Edited MARC Record

```
=000 01927cam a 00361Mi 00
=001 1552483
=005 20100704120842.0
=007 cr cn|||||||||
=008 090331s1817 dcu|||||s||||f||| 0|eng|d
=035 \\‡a(Readex)0FDBE2664851D648
=040 \\‡a Readex  ‡c Readex
=042 \\‡a dc
= 086 0\‡a Y 1.1/2:Serial Set Number 2
=110 1\‡a United States. Congress. Senate
=245 10‡a In Senate of the United States, December 8, 1817. Mr. Sanford submitted the following
motion for consideration: Resolved, that the Committee of Finance inquire what alterations or
amendments may be requisite in the present system of collecting the duties charged...‡h [electronic
resource]
=260 \\‡a [Washington, DC],  ‡c 1817.
=300 \\‡a 1 p.
=490 \0‡a Congressional Session Number 15th Congress, 1st Session; Session Volume Number 1
=500 \\‡a Title from opening lines of text.
=520 \\‡a Senate Document
=533 \\‡a Electronic Reproduction  ‡b Chester, VT.  ‡c Newsbank, inc.  ‡d 2005.  ‡n Available via
the World Wide Web  ‡n Access restricted to Readex U.S. Congressional Serial Set subscribers
=540 \\‡a Copyright 2005 by NewsBank, inc. All Rights Reserved.
=546 \\‡a English
=590 \\‡a ML: This title is also available in print. See the "United States congressional serial set", Call
number Y 1.1/2:
=650 07‡a Customs administration  ‡2 Readex congressional thesaurus
=650 07‡a Foreign trade  ‡2 Readex congressional thesaurus
=650 07‡a Imports  ‡2 Readex congressional thesaurus
=650 07‡a Tariffs and duties  ‡2 Readex congressional thesaurus
=655 7\ ‡a Motions and Resolutions  ‡2 Readex congressional thesaurus
=700 10‡a Sanford, Nathan,  ‡d 1777-1838,  ‡g Republican-Jeffersonian (NY)
=786 0\ ‡n Serial Set Number 2 S.Doc. 3
=856 41 ‡u
http://weblib.lib.umt.edu/redirect/proxyselect.php?url=http://docs.newsbank.com/select/serialset/0FD
BE2664851D648.html  ‡z Connect to this title online.
=949 \\ ‡c uonetlib  ‡d Y 1.1/2:  ‡t useminet
```

# Appendix B : DC to MARC Map

The following table contains the mapping/crosswalk between the Simple Dublin Core elements found in the Readex Serial Set records and MARC 21 bibliographic data elements

| Dublin Core | MARC 21 |
|---|---|
| Title | 245 10 ‡a (title statement/title proper) |
| Subject | 650 07 ‡a (subject added entry – topical term) ‡2 (source of thesaurus) |
| Description | 520 \\ ‡a (summary, etc. note)<br>300 \\ ‡a (physical description – extent) |
| Source | 786 0\ ‡n (data source note)<br>086 0\ ‡a (government document classification number) |
| Language | 546 \\ ‡a (language note) |
| Coverage | 490 \0 ‡a (uncontrolled series statement) |
| Creator | 110 1\ ‡a (main entry – corporate name)<br>100 1\ ‡a (main entry – personal name)<br>710 2\ ‡a (added entry – corporate name)<br>700 1\ ‡a (added entry – personal name) |
| Date | 260 \\ ‡c (date of publication, distribution, etc.) |
| Identifier | 856 40 ‡u (electronic location & access/URL) |
| Type | 655 \7 ‡a (index term – genre/form) |