Notes on Operations Repurposing MARC Metadata for an Institutional Repository

Working with Special Collections and University Press Monographs

Maureen P. Walsh

This paper describes the processes and workflows that transform Machine-Readable Cataloging (MARC) records found in The Ohio States University's library catalog into Dublin Core (DC) records for digital resources batch loaded into the Knowledge Bank, The Ohio State University's institutional repository. Two projects are described to illustrate the processes and workflows: the open-access monographs of The Ohio State University Press and the oral history collections of The Ohio State University Byrd Polar Research Center Archival Program.

An institutional repository, as both a software platform and a set of services, collects, organizes, preserves, and disseminates the digital output of an institution. The method of populating an institutional repository or adding digital content and associated metadata, varies by repository. Repositories may employ one or more deposit methods depending on their archiving and collection policies, the communities they serve, and repository staffing levels and resources. Deposit methods can include unmediated author (or his or her proxy) self-archiving, author (or his or her proxy) self-archiving mediated by repository staff, repository staff archiving on behalf of authors, and automated batch loading performed by repository staff. Institutional repositories also vary in their selection policies and, particularly in the United States, collections in institutional repositories represent a wide range of born digital and digitized material beyond peer-reviewed articles and electronic theses and dissertations.[1] As more scholarship is produced digitally and the cultural and intellectual resources of institutions are digitized, institutional repositories are seeing an ever-expanding source of content. In cases where this material is added to institutional repositories by repository or library staff and not by authors, the opportunity for an increasing flow of content is accompanied by the challenge of the metadata creation required to make that content accessible.

One way to address the challenge of creating metadata for institutional repositories is to repurpose (reuse) existing Machine-Readable Cataloging (MARC) metadata. Libraries have a wealth of descriptive metadata encoded in the MARC format in their library catalogs. Taking advantage of this robust legacy metadata and extending its semantic and descriptive value into new discovery environments and formats is an essential component of successful metadata management in today's heterogeneous metadata environment.[2] Repurposing library catalog metadata is

Maureen P. Walsh (walsh.260@osu.edu) is Metadata Librarian, The Ohio State University Libraries, Columbus, Ohio.

Submitted December 7, 2009; returned to author with request for revision and resubmission December 29, 2009; revision submitted February 27, 2010; tentatively accepted pending modest revision May 31, 2010; revision submitted June 19, 2010 and accepted for publication.

one way to reduce redundancy and increase efficiency in metadata creation for institutional repositories. This paper describes processes and workflows that use Extensible Stylesheet Language Transformations (XSLT) to transform metadata in MARC format in The Ohio State University's (OSU) library catalog to qualified Dublin Core (DC) format for the Knowledge Bank, OSU's institutional repository. The goal was to standardize procedures to facilitate metadata reuse and adaptation. Two projects are presented as case studies to illustrate an approach that might be applicable to other libraries looking to benefit from reusing existing metadata to automate metadata creation for an institutional repository. These case studies address creating metadata for a digitized collection of OSU Press monographs and, second, creating metadata for digital transcripts of oral history sound recordings held in a special collection. The processes described here also could be applied to creating metadata for theses, dissertations, and other cataloged materials that are digitized and added to an institutional repository, such as conference proceedings, exhibition catalogs, and technical reports.

Literature Review

While the literature is sparse regarding using MARC catalog records to create non-MARC metadata for institutional repositories, literature is available that describes repurposing non-MARC metadata to create MARC catalog records. Several institutions have implemented workflows to create MARC metadata for their library catalogs by repurposing non-MARC metadata. Most have focused on repurposing non-MARC metadata for electronic theses and dissertations (ETDs). For example, Surratt and Hill discussed a semiautomated workflow using Perl (a programming language) to create MARC catalog records from the metadata in an ETD database.[3] McCutcheon and colleagues described a semiautomatic process that also uses Perl to harvest OhioLINK ETD Center metadata, transform it to MARC, and add it to the library catalog.[4] Deng and Reese discussed the automatic harvesting and transformation of DC metadata for ETDs submitted to the DSpace repositories at Wichita State University and Oregon State University to create MARC catalog records using MarcEdit.[5]

Literature is also available that describes repurposing non-MARC metadata for institutional repositories. For example, Averkamp and Lee discussed workflows to repurpose non-MARC ProQuest metadata for batch loading ETDs into the University of Iowa's Digital Commons (bepress) repository.[6]

The examples the author found in the literature that discussed reusing MARC catalog records to create metadata for institutional repositories were conference presentations, posters, and unpublished reports in institutional repositories. Mundle and Thomas described how Simon Fraser University and the University of Adelaide Library, respectively, used Perl scripts to import metadata transformed from MARC records into DSpace.[7] Ng reported how the Hong Kong University Libraries used Perl to convert MARC records from Innovative Interface's Millennium catalog to DC for DSpace.[8] Robertson described how the University of Iowa used OCLC's Connexion to repurpose MARC records to add collections to CONTENTdm.[9] Branschofsky and colleagues discussed the batch importer developed by the Massachusetts Institute of Technology (MIT) to import MARC catalog records and transform them to DSpace-specific DC but did not describe the crosswalking method used.[10] Aside from the author's introduction to the use of XSLT to repurpose metadata, the use of XSLT workflows to repurpose MARC catalog metadata for institutional repositories has not been discussed in the literature.[11]

Background

The mission of the Knowledge Bank (https://kb.osu.edu/dspace) is to collect, preserve, and distribute the digital intellectual output of OSU's faculty, staff, and students.[12] The Knowledge Bank, a joint initiative of the OSU Libraries (OSUL) and the OSU Office of the Chief Information Officer, was first registered in the Registry of Open Access Repositories (http://roar.eprints.org) in September 2004. As of February 2010, the repository held 41,042 items in 1,207 collections. The Knowledge Bank uses DSpace (www.dspace.org), the open-source, Java-based repository software jointly developed by MIT Libraries and Hewlett-Packard. As a DSpace repository, the Knowledge Bank is organized by communities. The fifty-six communities currently in the Knowledge Bank include OSU administrative units, colleges, departments, journals, library special collections, research centers, and symposiums.

The structure of the Knowledge Bank follows the hierarchical arrangement of DSpace. Communities are at the highest level and can be divided into subcommunities. Each community or subcommunity contains one or more collections. All items, the basic archival element in DSpace, are contained within collections. Items consist of metadata and bundles of bitstreams (files). The selection of content for the Knowledge Bank at OSU is determined by the communities the repository serves. The collections in the Knowledge Bank contain a wide variety of items, including abstracts, articles, digital stories, journal issues, lectures, monographs, newsletters, oral history transcripts, photographs, presentations, proceedings, technical reports, and undergraduate theses. The formats of item content in the Knowledge Bank

include text, image, audio, and video.

The staff members working with the Knowledge Bank include people from OSUL's Technical Services, Information Technology (IT), and Preservation units, and the contracted hours of one systems developer from the OSU Office of Information Technology (OIT). The OSUL team members do not work full-time on the repository. The metadata management for the Knowledge Bank is the responsibility of the Scholarly Resources Integration Department (SRI) in Technical Services. The current SRI team working with the Knowledge Bank includes a librarian repository manager, two metadata librarians (including the author), two administrative and professional staff members, one graduate student assistant, and one student assistant.

The default metadata used by DSpace is a Qualified DC schema derived from the Dublin Core Library Application Profile.[13] An application profile is a schema that declares which metadata elements from one or more element sets, including locally defined sets, are used in an application or project and includes the policies and guidelines defined for the particular application or implementation.[14] The Knowledge Bank uses a locally defined extended version of the default DSpace Qualified DC schema that includes several additional element qualifiers. Metadata management for the Knowledge Bank is guided by a Knowledge Bank Metadata Application Profile (http://library.osu.edu/sites/techservices/KBAppProfile.php) and a core element set document for each collection within the repository derived from the application profile.

The purpose of the core element set for each collection is to provide guidance to contributors in the metadata creation process, record the variations of the application profile necessary to adequately represent the content of a particular collection, and improve both local and remote resource discovery. The core element set is adapted from the more general application profile specifically for each collection and describes the elements that compose the core set of metadata for a collection. The core element sets document metadata decisions on a project-by-project basis to ensure consistency and to provide a record for future reference. A core element set for a collection documents the minimum DC elements (and any qualifiers) used for a collection. For each element, the following is recorded: public display label, submission interface label, definition and usage guidelines, how supplied (system or depositor), obligation (which may be mandatory, required if available, optional), occurrence (repeatable or nonrepeatable), whether controlled or free text, recommended content schema, indexing, and usage example(s). The SRI librarians create the core element sets in consultation with Knowledge Bank community representatives. The core element sets serve as metadata guidelines for submitting items into the Knowledge Bank regardless of the method of ingest.

The two ways items are added to collections in the Knowledge Bank are direct, or intermediated, entry via the DSpace item submission user interface and via the DSpace batch item importer. Items are submitted directly to the Knowledge Bank via the item submission user interface by authors, community representatives, and SRI staff, students, and librarians. The submission process is customizable by collection and the metadata entry pages of the item submission user interface can include metadata fields unique to the collection, prepopulated default metadata, and dropdown menus with controlled vocabulary. The Knowledge Bank uses a general (default) input form for all collections not assigned to a specifically customized (alternate) input form. The SRI librarians create custom metadata entry pages by modifying the submission input forms XML file.

An input form is a customized set of pages through which submitters enter metadata. The input form determines the metadata fields available to the submitter, the order they are displayed, their labels, and their explanations. The input form also controls what fields are repeatable or required and the field input type, or how values are entered (i.e., name, date, dropdown menu, text box, etc.). The SRI librarians add prepopulated default-item metadata to the submission entry pages by modifying the collection item templates via the DSpace web user interface. The submission process also can include workflow steps for checking and editing metadata before items are made publically available. Individuals designated as collection administrators (most often these individuals are SRI librarians and staff, but they can also be community representatives) use the workflow option to check and edit the work of their community submitters. SRI staff and librarians use the workflow option to manage collections with an embargo option. For certain projects or collections, they use the workflow option to mediate deposit by authors, community representatives, SRI staff in training, or SRI student assistants.

The DSpace item importer is a command-line tool for batch loading items. The tool uses the simple archive format shown here.

```
archive_directory/
item_000/
dublin_core.xml—Qualified
    Dublin Core metadata
contents—text file containing
    one line per filename
file_1.pdf—files to be added as
    bitstreams to the item
file_2.pdf
item_001/
dublin_core.xml
contents
file_1.pdf
. . .
```

The archive is a directory of items each containing a subdirectory of item metadata, item files, and a contents file listing the bitstream file names. Each item's descriptive metadata are contained in a DC XML file. The format used by DSpace for the DC XML files is illustrated below.

```
<dublin_core>
<dcvalue element="title"
    qualifier="none">
    Understanding narrative
    </dcvalue>
<dcvalue element="creator"
    qualifier="none">Phelan,
    James, 1951-</dcvalue>
<dcvalue element="date"
    qualifier="issued">1994</
    dcvalue>
<dcvalue element="subject"
    qualifier="lcsh">Narration
    (Rhetoric)</dcvalue>
</dublin_core>
```

Staff members working with the Knowledge Bank have sought from the inception of the repository to be as efficient as possible in adding content. Using batch loading to populate the repository has been integral to that efficiency. During the last five years, 698 collections containing 32,188 items have been batch loaded, representing 78 percent of the items and 58 percent of the collections in the Knowledge Bank. These batch loaded collections vary from journal issues to photo albums, and the items include articles, images, abstracts, and transcripts. The creation of the metadata has varied by project. At times it has been created by SRI librarians and staff, but metadata also has been supplied by Knowledge Bank communities in consultation with an SRI metadata librarian or by a vendor contracted by OSUL. Metadata supplied in Excel spreadsheets or CSV files is mapped by an SRI metadata librarian to Qualified DC to prepare it for batch loading into the Knowledge Bank. An OIT or IT systems developer uses a custom programming script (Perl or Java) to migrate the data from the Excel spreadsheets or CSV files into the DSpace simple archive directory format, then batch loads the descriptive metadata and content files via the item importer. An SRI metadata librarian checks a random DC XML file generated by the script to verify expected results before running the batch process. After a successful run in the test mode of the importer, the batch is loaded into the staging instance of the Knowledge Bank used for testing and development. An SRI metadata librarian reviews the batch for expected results and identifies any loading errors that need to be corrected. After a successful load into the staging instance, the batch is loaded into the production instance (publically available version) of the Knowledge Bank.

The Knowledge Bank batch loading workflows using Excel spreadsheets or CSV files as a metadata source evolved through an iterative process of continual refinement and improvement. Batch loading works well for large collections where the metadata can be supplied by SRI, a Knowledge Bank community, or a vendor in an Excel or CSV file. The SRI librarians continually refine and revise the metadata workflows for new and ongoing collections in the Knowledge Bank that fall outside the parameters of the established batch loading processes as new ways to improve efficiency and reduce redundancy are identified. In cases where encoded metadata are available and can be automatically transformed into the DC XML for DSpace, adding that data to an Excel spreadsheet to batch load it into the Knowledge Bank would be redundant work. This paper describes the new batch loading workflows developed by the author for the OSU Press and the OSU Byrd Polar Research Center Archival Program where metadata was available in the library catalog.

Case Studies

The Ohio State University Press Open Access Monographs

The OSU Press Publications collection in the Knowledge Bank (http://hdl.handle.net/1811/131) contains open-access monographs no longer in print and new monographs for which the full-text is embargoed five years following date of publication. The OSU Press staff continue to add the new embargoed titles as they are published using a DSpace submission workflow mediated by SRI staff. The project to digitize and archive the open-access monographs began in August 2005. An SRI librarian created the core element set for the collection and designed the metadata workflow in consultation with the OSU Press community. The three hundred open-access titles contained in the collection were selected by OSU Press, digitized by a vendor funded by OSUL, and added to the collection by SRI staff and students. The open-access titles were digitized in three groups of one hundred and individually archived in the Knowledge Bank as the files were received from the vendor in 2006 and 2007. The monographs (metadata and PDF files) were added individually to the collection via the item submission user interface.

The metadata workflow for archiving the digital open-access monographs took advantage of the fact that the OSU Press print monographs were fully cataloged with MARC records in the library catalog. The workflow was designed to manually use the MARC catalog metadata to create DC item records in the Knowledge Bank. The core element set document listed the MARC data to be reused for DC fields including creator, date issued, description, description (table of contents), title, title (alternative), identifiers, publisher, relation (is part of series), and subject (Library of Congress Subject Headings (LCSH)). The SRI staff and students who submitted the items to

the Knowledge Bank used the title, author, date, and any other pertinent information found in the PDF file of the digital monograph to search for and identify the correct MARC record in the catalog for the matching print version of the monograph. The staff member or student copied the reusable data from the catalog record and pasted it into the appropriate fields in the input form as they entered metadata and uploaded the PDF for each title. Using the library catalog metadata this way significantly reduced the amount of keying needed to input the metadata.

While manually copying and pasting the catalog metadata for the repository items saved keying, it was still a labor-intensive process. To improve the workflow efficiency, the author developed a new procedure in September 2007 to automate the process of reusing the catalog metadata for the Knowledge Bank records following receipt of the final one hundred digitized monographs. The design of the new workflow began with the core element set document for the OSU Press collection. The author revised the document to serve as a descriptive metadata crosswalk for the transformation of the metadata from MARC to qualified DC. The one hundred MARC records for the transformation were extracted from OSU's Innovative Interfaces Millennium catalog. The author then used the MARC-to-MARCXML function in MarcEdit to transform the MARC output file to MARCXML.

The author used an existing MARCXML to DC XSLT stylesheet available from the Library of Congress as a template and modified it for the requirements of the project.[15] The first modifications to the template changed the DC schema used by the stylesheet. The Library of Congress stylesheet uses oai_dc, the XML schema for the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) and outputs the MARCXML metadata in the OAI format of DC. This OAI Simple DC schema was replaced with the qualified DC schema required for importing items in batch to DSpace. Figure 1 shows an example of each format. The author registered, or added, the modified stylesheet to the local application of the MarcEdit software for testing. An XML conversion function was added to the MarcEdit tool to use the modified stylesheet added to the software. The author then used the OSU Press MARCXML file as the test input file for the transformation. The new stylesheet successfully transformed the MARCXML file and output a file of Qualified DC in the DSpace format.

Following the conversion of the stylesheet to the DSpace Qualified DC schema, the author used the metadata crosswalk for the project as a guide to modify the MARC data fields in the stylesheet. The MARC fields not used in the OSU Press profile were removed from the stylesheet, including the 506 (restrictions on access), 521 (target audience), 530 (additional physical form available), 653 (index term—uncontrolled), 662 (subject added entry—hierarchical place name), and 856 (electronic location/access) fields. The various 76x, 77x, and 78x relation fields also were removed. The MARC fields needed for the transformation not present in the stylesheet were added, such as the MARC 440 and 830 series fields. The MARC table of contents field 505 was removed from the MARC 5xx fields for the DC description field and added as a separate DC table of contents description field. A DC contributor field was added and the MARC fields that were mapped to DC creator were split between creator and contributor. The MARC 6xx subject fields remaining in the modified stylesheet were restricted to LCSH by specifying a second indicator of "0" (zero) for each MARC 6xx data field. The MARC subfield codes also were modified on the basis of the requirements of the crosswalk. For example, the subdivisions "v" (form), "x" (general), "y" (chronological), and "z" (geographical) were added to the existing subfield codes for the MARC 6xx subject data fields. The author then added the revised stylesheet to the local MarcEdit software application. The MarcEdit XML conversion function for the project, modified to use the revised stylesheet, transformed the MARCXML file to Qualified DC. The result of the transformation was one XML file with all of the DC records for the OSU Press items. The next step in the new workflow involved preparing the DC metadata and the PDF files for batch loading into the Knowledge Bank.

```
<oai_dc:dc>
<dc:title>Carlyle and the search for
    authority</dc:title>
<dc:date>1991</dc:date>
<dc:subject>Authority in literature</
    dc:subject>
</oai_dc:dc>

<dublin_core>
<dcvalue element="title"
    qualifier="none">Carlyle and the
    search for authority</dcvalue>
<dcvalue element="date"
    qualifier="issued">1991</dcvalue>
<dcvalue element="subject"
    qualifier="lcsh">Authority in litera-
    ture</dcvalue>
</dublin_core>
```

Figure 1. Comparison of OAI Simple Dublin Core and DSpace Qualified Dublin Core

The OSU Press stylesheet still had limitations that needed to be addressed. The manual cleanup of the XML file included removing excess white space, unwanted punctuation, and the namespace declarations occurring within each DC record. Double-dash delimiters were also added between the subject subfields. After the author corrected the metadata, the OIT systems developer wrote a short Perl script that separated the DC records contained in the one XML file into individual dublin_core.xml files

and automatically creating the entire archive directory (PDF files and DC metadata). The author had manually pasted the PDF file names to the end of each record in the XML file during the metadata cleanup to facilitate this parsing. The OIT systems developer successfully batch loaded the completed archive directory (DC metadata and PDF content files) into the staging, or test, instance of the Knowledge Bank.

The successful batch could have been loaded into the production instance of the Knowledge Bank, but this new OSU Press workflow became a proof-of-concept exercise because the project's last group of one hundred monographs was manually entered into the Knowledge Bank for staff training purposes. Valuable experience was gained in the process of developing the automated workflow. The most important outcomes were the addition of a new automated workflow to repurpose library catalog metadata and a new batch loading process available for projects where the source metadata are contained in DC XML rather than in an Excel spreadsheet or a CSV file. As a result of the exercise, the author identified limitations of the new MARCXML-to-DSpace Qualified DC stylesheet that required further investigation to minimize the manual cleanup needed for the transformed metadata. The author refined the new stylesheet for use in a subsequent project for the OSU Byrd Polar Research Center Archival Program described below.

The Ohio State University Byrd Polar Research Center Archival Program Oral History Collections

The OSU Byrd Polar Research Center Archival Program is a collaborative effort of the Byrd Polar Research Center (BPRC) and the OSU Archives. The BPRC, recognized internationally as a leader in polar and alpine research, is named in honor of the polar explorer Admiral Richard E. Byrd. Part of OSUL's Special Collections, the BPRC Archival Program collects, preserves, and provides access to historical polar-region material. The BPRC community in the Knowledge Bank currently includes collections for reports, publications, conferences, photo albums, and two oral history collections: the Polar Oral History Program collection (http://hdl.handle.net/1811/6039) and the Antarctic Deep Freeze Oral History Project collection (http://hdl.handle.net/1811/35321). The Polar Oral History Program collection archives the program's transcripts of interviews documenting the early years of American polar exploration. The Antarctic Deep Freeze Oral History Project collection archives the transcripts of interviews with pioneers of South Pole exploration documenting United States involvement in Antarctica during the 1950s.

The Polar Oral History Program collection currently contains eighty-five transcript items. The first item was submitted to the Knowledge Bank in 2005. The metadata workflow for the collection was designed by a former metadata librarian in consultation with the community representative, the polar curator. The metadata librarian created the core element set document and a specialized input form for the collection. Following the workflow set up by the former SRI metadata librarian, the polar curator enters the items individually using the input form customized for the collection in groups of approximately five to fifteen interview transcripts. Similar to the availability of MARC records for the print versions of the digital OSU Press monographs, MARC records for the sound recordings used to transcribe the interviews are available in the library catalog. Approximately 80 percent of 104 interview audio tapes were represented in the catalog at the beginning of the Knowledge Bank project.

The core element set for the collection reused the available catalog metadata for the DC title, creator, contributor, date created, relation (is part of series), subject (LCSH), and subject (other) fields. The former metadata librarian added the values for the creator, contributor, subject (LCSH), and subject (other) fields to dropdown menus in the input form. The DC fields that have constant (default) metadata for the collection, including publisher, rights, relation (is part of series), sponsorship, and type were added to the collection item template used to prepopulate fields in the input form.

Although labor-intensive, the work done at the beginning of the project to create the customized input form significantly reduced the keystrokes required of the polar curator to submit each item. Because not all of the audio tapes were cataloged when the input form was created, the drawback to this workflow was the ongoing maintenance. During the last several years, new completed groups of transcripts included some that did not have metadata in the input form dropdown menus, either because the matching audio tapes were not yet cataloged or were cataloged after the original input form was created. The author added names and subjects to the dropdown menus as MARC records were added to the catalog by OSUL's Special Collections Cataloging (SCC).

The polar curator requested a new BPRC oral history collection for thirty-nine Antarctic Deep Freeze Oral History Project transcripts in October 2008. The author, in assessing the metadata requirements of the project, had several options. The existing metadata workflow for the Polar Oral History Program could have been extended to include the new collection. To use the same workflow for the new collection, the author would need to manually copy and paste all of the reusable catalog metadata to the XML input form (either the same

input form as used by the Polar Oral History program or a new input form modeled on it). Another option would have been for the Polar Curator to use an input form that had default data prepopulated via the item template, but did not have controlled vocabulary dropdown menus for the reusable, variable, catalog metadata. This second option would have eliminated the labor-intensive work on the part of a metadata librarian but would have shifted the rekeying to the polar curator who would copy and paste from the catalog records to input the items. Given the amount of redundant work and rekeying involved, the first and second options were not ideal. The author, in consultation with the polar curator, decided to try a third option. The new automated workflow tested with the OSU Press open-access monographs collection was modified and used for both oral history collections.

For the new workflow design process the author first addressed the existing Polar Oral History collection while SCC cataloged the Antarctic Deep Freeze audio tapes for the library catalog. The author used the XSLT stylesheet created for the OSU Press project as a template and modified it for the Polar Oral History collection. The fields for the new stylesheet were based on the Polar Oral History core element set. The constant data fields (default metadata) for the Polar Oral History collection, including publisher, rights, relation (is part of series), sponsorship, and type were added to the stylesheet as text elements, which serve to output the text enclosed within the element.

The author addressed many of the limitations of the OSU Press stylesheet in modifying the XSLT for the Polar Oral History Program. Functions to normalize space and remove unwanted punctuation were added. A template was added so that the XML namespace declaration would not occur in each DC record. Figure 2 illustrates the additions the author made to the "subfieldSelect" template to allow for alternate delimiters. An example of the modifications to the Polar Oral History stylesheet to create double-dash delimiters in subject fields is shown in the 650 data field in figure 3.

With no new Polar Oral History transcripts to batch load, the author tested the Polar Oral History stylesheet with transcripts that were already in the Knowledge Bank. Following the same process as with the OSU Press workflow, a list of catalog records was made and MARC records were output. The Polar Oral History stylesheet and the Library of Congress utility stylesheet modified for alternate delimiters were added to the local application of MarcEdit, a new XML conversion function for the project was created, and the MARC OUT file from the catalog was used for the

```
<xsl:template name="subfieldSelect">
  <xsl:param name="codes"/>
  <xsl:param name="delimeter">
    <xsl:text> </xsl:text>
  </xsl:param>
  <xsl:param name="altCodes"/>
  <xsl:param name="altDelimiter"/>
  <xsl:variable name="str">
    <xsl:for-each select="marc:subfield">
      <xsl:if test="contains($codes, @code)">
        <xsl:if test="contains($altCodes, @code)">
          <xsl:value-of select="$altDelimiter"/>
        </xsl:if>
        <xsl:value-of select="text()"/>
        <xsl:value-of select="$delimeter"/>
      </xsl:if>
    </xsl:for-each>
  </xsl:variable>
  <xsl:value-of select="substring($str,1,string-length($str)-string-
length($delimeter))"/>
</xsl:template>
```

Figure 2. The Library of Congress subfieldSelect Utility XML Stylesheet Template Modified for Alternate Delimiters

```
<xsl:for-each select="marc:datafield[@tag=650][@ind2=0]">
<dcvalue element="subject" qualifier="lcsh">
<xsl:call-template name="chopPunctuation">
<xsl:with-param name="chopString">
<xsl:call-template name="subfieldSelect">
<xsl:with-param name="codes">abcdvxyz</xsl:with-param>
<xsl:with-param name="altCodes">vxyz</xsl:with-param>
<xsl:with-param name="altDelimiter">-- </xsl:with-param>
</xsl:call-template>
</xsl:with-param>
</xsl:call-template>
</dcvalue>
</xsl:for-each>
```

Figure 3. The MARC 650 Data Field in the Polar Oral History Stylesheet

transformation. In contrast to the OSU Press workflow, the author did not generate a separate MARCXML file as an interim step. Instead, the MarcEdit option of beginning with a MARC file was chosen. With this option, the MarcEdit software does the prerequisite MARC to MARCXML conversion as part of the transformation to DC.

The amount of cleanup needed for the Polar Oral History DC metadata was significantly reduced from the author's previous attempt with the OSU Press stylesheet. The author, however, did need to manually remove the duplicate identifier fields from the DC XML file resulting from the duplicate identical MARC 099 local call number fields in the MARC records. Editing of the file was also required because the Polar Oral History core element set calls for variable metadata values that are contained in the MARC record but cannot be directly transformed from the MARC data. One example of this is the DC relation (is format of) field, which is used to indicate that the described resource is the same intellectual content of the referenced resource, but presented in another format. The core element set uses the relation (is format of) field for a statement about the audio tapes from which the transcripts are derived. The MARC 300 physical description field, which contains the number of sound recordings, was transformed into a DC description field. The relation (is format of) field was added to the stylesheet as a text element with a constant data statement and an underscore for the number of audio tapes. For each record in the DC XML file after the transformation, the author copied the number of audio tapes from the description field, added the number to the relation (is format of) field in place of the underscore, and deleted the description field. In the final editing step for the DC XML file, the author added the PDF file names for the transcripts to the end of each DC record in preparation for batch loading.

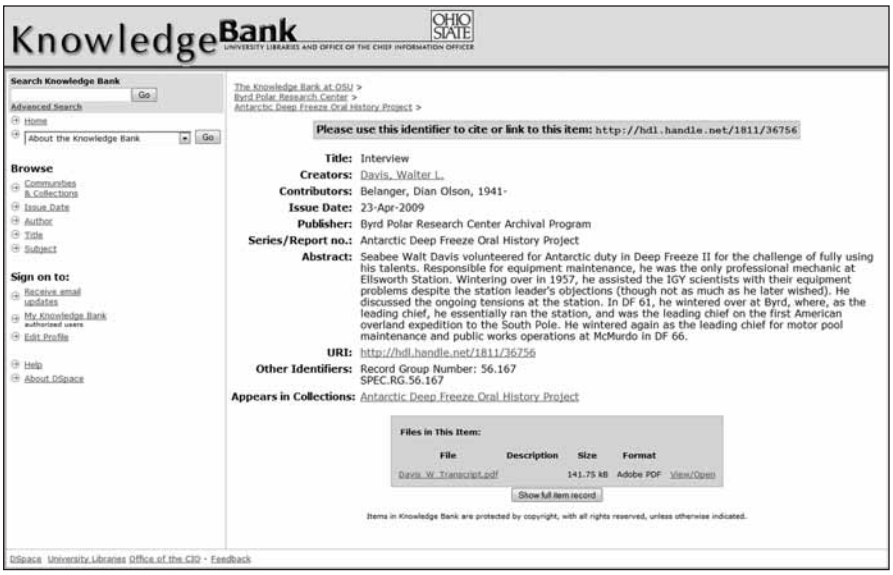

Figure 4. Example of a Knowledge Bank Antarctic Deep Freeze Transcript Item

The Polar Oral History DC and PDF files were prepared for batch loading using the same Perl script written for the OSU Press workflow. The OIT systems developer used the Perl script to create the archive directory, and the collection was successfully batch loaded into the staging instance of the Knowledge Bank. The author plans to use the new Polar Oral History Program workflow for the next set of available transcripts.

The author used the Polar Oral History XSLT stylesheet as a template for the Antarctic Deep Freeze Oral History Project workflow and created a core element set for the new collection in consultation with the polar curator. The core element set was used to determine the fields needed in the stylesheet. As with the Polar Oral History stylesheet, constant data fields for the Antarctic Deep Freeze collection were added as text elements to the stylesheet. The Antarctic Deep Freeze stylesheet also uses the same modified Library of Congress utility stylesheet used for the Polar Oral History workflow. The XSLT stylesheet used for the Antarctic Deep Freeze project is available in appendix A. When the cataloging for the Antarctic sound recordings was completed by SCC, the author gathered the MARC records in a review file and output them to a personal computer. The MARC OUT file was transformed to qualified DC utilizing the Antarctic Deep Freeze stylesheet added to the local application of the MarcEdit software. Appendix B includes an example comparison of Antarctic Deep Freeze metadata in the original MARC format and in DSpace qualified DC format for batch loading.

The DC XML file for the Antarctic Deep Freeze project required cleanup similar to that of the Polar Oral History records. For example, the author modified the relation (is format of) fields and removed the duplicate identifier fields resulting from duplicate fields in the MARC records. As with the previously described projects, the PDF file names were added to the end of each record, and the Perl script written for the OSU Press project was reused to create the DSpace archive directory. The new collection of thirty-nine transcripts was successfully batch loaded into the staging instance of the Knowledge Bank. The Polar Curator reviewed and approved the collection on the staging server, and

the collection was successfully batch loaded into the production instance of the Knowledge Bank on April 23, 2009. Figure 4 is an example of the public simple (short) item view for an Antarctic Deep Freeze transcript record in Knowledge Bank.

The automated workflow for the Antarctic Deep Freeze collection worked well. The XSLT transformation process went smoothly because of the previous work done to optimize it during the OSU Press and Polar Oral History tests. The author plans to further investigate XSLT options to reduce manual metadata cleanup for transformed records. However, the manual cleanup of the DC records could have been mitigated if different choices had been made regarding the desired Qualified DC metadata for the collection. Balancing metadata quality and completeness with efficiency is always an issue when creating metadata, and the same challenge is present when repurposing metadata. The author could have eliminated manual transformation procedures if the polar curator and author had decided to change the metadata profile. For example, the relation (is format of) field could have been a constant data field that stated audio tapes were available without providing the number.

The Perl script written by the OIT systems developer for the initial OSU Press test proved a successful way to create the batch loading archive directory for DC XML file source metadata. However, the author plans to further investigate an automated way to add the PDF file names to each record in the DC XML file needed for the current Perl script to create the batch loading archive directory.

Conclusion

As library budgets tighten and technical services staffing shrinks, the process of creating quality metadata for library resources must be made more efficient. Metadata creation, traditionally an expensive and resource-draining endeavor, can be streamlined in many ways. This paper detailed one: repurposing existing MARC catalog metadata for an institutional repository. The iteratively refined repurposing workflow described in this paper has opened new metadata management possibilities for OSUL. The author will continue to refine the automated XSLT workflow and investigate the application of the workflow to new projects. The author is currently working on an automated XSLT workflow to repurpose Knowledge Bank metadata for digital resources that OSUL has decided to represent in both the Knowledge Bank and the library catalog.

The repurposing workflows described here were designed with the goal of standardizing the procedures to facilitate metadata reuse and adaptation. The flexibility of the XSLT workflow promotes the reusability of the tools and processes while allowing for collection-specific refinements. The automated workflow can be modified for local requirements in other libraries that wish to use XSLT to transform MARC metadata for an institutional repository.

References

1. Elizabeth Yakel et al., "Institutional Repositories and the Institutional Repository: College and University Archives and Special Collections in an Era of Change," *American Archivist* 71, no. 2 (2008): 323–49.
2. Martin Kurth, David Ruddy, and Nathan Rupp, "Repurposing MARC Metadata: Using Digital Project Experience to Develop a Metadata Management Design," *Library Hi Tech* 22, no. 2 (2004): 153–65.
3. Brian E. Surratt and Dustin Hill, "ETD2MARC: A Semiautomated Workflow for Cataloging Electronic Theses and Dissertations," *Library Collections, Acquisitions, & Technical Services* 28, no. 2 (2004): 205–23.
4. Sevim McCutcheon et al., "Morphing Metadata: Maximizing Access to Electronic Theses and Dissertations," *Library Hi Tech* 26, no. 1 (2008): 41–57.
5. Sai Deng and Terry Reese, "Customized Mapping and Metadata Transfer from DSpace to OCLC to Improve ETD Work Flow," *New Library World* 110, no. 5/6 (2009): 249–64.
6. Shawn Averkamp and Joanna Lee, "Repurposing ProQuest Metadata for Batch Ingesting ETDs into an Institutional Repository," *Code4Lib Journal*, no. 7 (2009), http://journal.code4lib.org/articles/1647 (accessed Feb. 27, 2010).
7. Todd M. Mundle, "Digital Retrospective Conversion of Theses and Dissertations: An In House Project" (presentation, 8th International Symposium on Electronic Theses & Dissertations, Sydney, Australia, Sept. 28–30, 2005), http://adt.caul.edu.au/etd2005/papers/080Mundle.pdf (accessed Feb. 27, 2010); Steve Thomas, "Importing MARC Data Into DSpace" (technical report, Systems Department, The University of Adelaide Library, Aug. 6, 2006), http://hdl.handle.net/2440/14784 (accessed Feb. 27, 2010).
8. Alan Ng, "Converting Millennium ILS Bibliographic Records Into Dublin-Core XML Format for DSpace" (presentation, PNC 2009 Annual Conference and Joint Meetings, Taipei, Taiwan, Oct. 6–8, 2009), www.pnclink.org/pnc2009/english/PresentationMaterial/Oct06/06-Rm1-eResource2/06-eResource2-ppt-AlanNg.pdf (accessed Feb. 27, 2010).
9. Wendy C. Robertson, "Using MARC Records to Populate CONTENTdm," (presentation, Midwest CONTENTdm Users Group 3rd Annual Meeting, Indianapolis, Indiana, April 30, 2008), http://ir.uiowa.edu/lib_pubs/23 (accessed Feb. 27, 2010).
10. Margret Branschofsky et al., "Evolving Metadata Needs for an Institutional Repository: MIT's DSpace," *Proceedings of the 2003 International Conference on Dublin Core and Metadata Applications: Supporting Communities of Discourse and Practice—Metadata Research & Applications*, Seattle, Wash., 2003,

http://dcpapers.dublincore.org/ojs/pubs/article/view/753/749 (accessed Feb. 27, 2010).
11. Maureen P. Walsh, "Metadata Repurposing Using XSLT," in *More Technology for the Rest of Us: A Second Primer on Computing for the Non-IT Librarian*, ed. Nancy Courtney (Santa Barbara, Calif.: Libraries Unlimited, 2010): 125–39.
12. The Ohio State University Libraries, Knowledge Bank, Mission (2010), http://library.osu.edu/projects-initiatives/knowledge-bank/open-access-archiving/mission (accessed May 31, 2010).
13. Dublin Core Metadata Initiative Libraries Working Group, "DC-Library Application Profile (DC-Lib)," http://dublincore.org/documents/2004/09/10/library-application-profile (accessed Feb. 27, 2010).
14. Mary S. Woodley, Gail Clement, and Pete Winn, "DCMI Glossary," 2005, http://dublincore.org/documents/2005/11/07/usageguide/glossary.shtml (accessed Feb. 27, 2010).
15. Library of Congress, MARCXML to OAI Encoded Simple Dublin Core Stylesheet, www.loc.gov/standards/marcxml/xslt/MARC21slim2OAIDC.xsl (accessed Feb. 27, 2010).

Appendix A. MARCXML to DSpace Qualified Dublin Core Stylesheet

```
<?xml version="1.0" encoding="UTF-8"?>
<xsl:stylesheet version="1.0" xmlns:marc="http://www.loc.gov/MARC21/slim"
    xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
    xmlns:dc="http://purl.org/dc/elements/1.1/" xmlns:xsl="http://www.w3.org/1999/XSL/Transform" exclude-result-prefixes="marc">
    <xsl:import href="MARC21slimUtils4KB.xsl"/>
    <xsl:output method="xml" encoding="UTF-8" indent="yes"/> <xsl:param name="destfile"/>

<!--The Ohio State University Knowledge Bank: Antarctic Deep Freeze Oral History Program. MPW-->

<xsl:template match="/">
  <dublin_core>
    <xsl:apply-templates/>
  </dublin_core>
</xsl:template>

<xsl:template match="marc:record">
  <xsl:variable name="leader" select="marc:leader"/>
  <xsl:variable name="leader6" select="substring($leader,7,1)"/>
  <xsl:variable name="leader7" select="substring($leader,8,1)"/>
  <xsl:variable name="controlField008" select="marc:controlfield[@tag=008]"/>

<dublin_core>

<dcvalue element="identifier" qualifier="none">
  <xsl:text>Record Group Number: </xsl:text>
</dcvalue>

<xsl:for-each select="marc:datafield[@tag=099]">
  <dcvalue element="identifier" qualifier="none">
    <xsl:value-of select="marc:subfield[@code='a']"/>
  </dcvalue>
</xsl:for-each>

<xsl:for-each select="marc:datafield[@tag=245]">
  <dcvalue element="title" qualifier="none">
    <xsl:call-template name="subfieldSelect">
      <xsl:with-param name="codes">ab</xsl:with-param>
    </xsl:call-template>
  </dcvalue>
</xsl:for-each>

<xsl:for-each select="marc:datafield[@tag=100]">
  <dcvalue element="creator" qualifier="none">
    <xsl:variable name="c" select="."/>
    <xsl:value-of select="normalize-space($c)"/>
  </dcvalue>
</xsl:for-each>

<dcvalue element="contributor" qualifier="none">
  <xsl:text>Belanger, Dian Olson, 1941-</xsl:text>
</dcvalue>

<xsl:for-each select="marc:datafield[@tag=260]/marc:subfield[@code='c']">
  <dcvalue element="date" qualifier="created">
    <xsl:variable name="t" select="."/>
    <xsl:value-of select="translate($t,'.','')"/>
  </dcvalue>
</xsl:for-each>

<xsl:for-each select="marc:datafield[@tag=518]">
  <dcvalue element="description" qualifier="none">
    <xsl:value-of select="marc:subfield[@code='a']"/>
  </dcvalue>
</xsl:for-each>

<dcvalue element="publisher" qualifier="none">
```

Appendix A. MARCXML to DSpace Qualified Dublin Core Stylesheet (continued)

```
    <xsl:text>Byrd Polar Research Center Archival Program</
xsl:text>
   </dcvalue>

   <xsl:for-each select="marc:datafield[@tag=300]">
     <dcvalue element="description" qualifier="none">
       <xsl:value-of select="."/>
     </dcvalue>
   </xsl:for-each>

   <dcvalue element="relation" qualifier="isformatof">
     <xsl:text>_ audio tapes available in the OSU Archives</
xsl:text>
   </dcvalue>

   <dcvalue element="relation" qualifier="ispartofseries">
     <xsl:text>Antarctic Deep Freeze Oral History Project</
xsl:text>
   </dcvalue>

   <xsl:for-each select="marc:datafield[@tag=520]">
     <dcvalue element="description" qualifier="abstract">
       <xsl:value-of select="marc:subfield[@code='a']"/>
     </dcvalue>
   </xsl:for-each>

   <xsl:for-each select="marc:datafield[@tag=600][@ind2=0]">
     <dcvalue element="subject" qualifier="other">
       <xsl:call-template name="chopPunctuation">
         <xsl:with-param name="chopString">
           <xsl:call-template name="subfieldSelect">
             <xsl:with-param name="codes">abcdvxyz</
xsl:with-param>
             <xsl:with-param name="altCodes">vxyz</
xsl:with-param>
             <xsl:with-param name="altDelimiter">-- </
xsl:with-param>
           </xsl:call-template>
         </xsl:with-param>
       </xsl:call-template>
     </dcvalue>
   </xsl:for-each>

   <xsl:for-each select="marc:datafield[@tag=610][@
ind2=0]|marc:datafield[@tag=630][@ind2=0]|
   marc:datafield[@tag=650][@ind2=0]|marc:datafield[@
tag=651][@ind2=0]">
     <dcvalue element="subject" qualifier="lcsh">
       <xsl:call-template name="chopPunctuation">
         <xsl:with-param name="chopString">
           <xsl:call-template name="subfieldSelect">
             <xsl:with-param name="codes">abcdvxyz</
xsl:with-param>
             <xsl:with-param name="altCodes">vxyz</
xsl:with-param>
             <xsl:with-param name="altDelimiter">-- </
xsl:with-param>
           </xsl:call-template>
         </xsl:with-param>
       </xsl:call-template>
     </dcvalue>
   </xsl:for-each>

   <dcvalue element="description" qualifier="none">
     <xsl:text>The Antarctic Deep Freeze oral history project
was funded by a grant from the National Science Foundation
and supported by the Antarctic Deep Freeze Association. The
original paper copies and unaltered tapes have been deposited
in the library of the National Science Foundation.</xsl:text>
   </dcvalue>

   <dcvalue element="description" qualifier="sponsorship">
     <xsl:text>National Science Foundation</xsl:text>
   </dcvalue>

   <dcvalue element="description" qualifier="sponsorship">
     <xsl:text>Antarctic Deep Freeze Association</xsl:text>
   </dcvalue>

   <dcvalue element="language" qualifier="iso">
     <xsl:value-of select="substring($controlField008,36,2)"/>
   </dcvalue>

   <dcvalue element="type" qualifier="none">
     <xsl:text>Transcript</xsl:text>
   </dcvalue>

   <dcvalue element="rights" qualifier="none">
     <xsl:text>Restrictions: This item is not restricted</xsl:text>
   </dcvalue>

   </dublin_core>
   </xsl:template>
   </xsl:stylesheet>
```

Appendix B. Comparison of Original MARC Record Fields to DSpace Qualified Dublin Core XML

MARC Fields from Library Catalog Record	Dublin Core XML Record for Batch Loading
008 090220s1999 xxunnn t eng d	<dcvalue element="language" qualifier="iso">en</dcvalue>
099 SPEC.RG.56.167	<dcvalue element="identifier" qualifier="none">Record Group Number: 56.167</dcvalue>
099 SPEC.RG.56.167	<dcvalue element="identifier" qualifier="none">SPEC.RG.56.167</dcvalue>
100 1 Davis, Walter L	<dcvalue element="creator" qualifier="none">Davis, Walter L.</dcvalue>
245 10 Interview|h[sound recording] /|cWalter L. Davis [interview by Dian O. Belanger]	<dcvalue element="title" qualifier="none">Interview</dcvalue>
260 |c1999 518 Recorded on Sept. 24, 1999	<dcvalue element="date" qualifier="created">1999-09-24</dcvalue>
300 2 sound cassette (90 minutes) :|b1 7/8 ips	<dcvalue element="relation" qualifier="isformatof">2 audio tapes available in the OSU Archives</dcvalue>
520 Seabee Walt Davis volunteered… [truncated]	<dcvalue element="description" qualifier="abstract">Seabee Walt Davis volunteered …[truncated].</dcvalue>
700 1 Belanger, Dian Olson,|d1941-	<dcvalue element="contributor" qualifier="none">Belanger, Dian Olson, 1941-</dcvalue>
600 10 Davis, Walter L.|vInterviews	<dcvalue element="subject" qualifier="other">Davis, Walter L. -- Interviews</dcvalue>
610 10 United States.|bNavy|xMachinery|xMaintenance and repair	<dcvalue element="subject" qualifier="lcsh">United States. Navy -- Machinery -- Maintenance and repair</dcvalue>
650 0 Operation Deep Freeze	<dcvalue element="subject" qualifier="lcsh">Operation Deep Freeze</dcvalue>
651 0 Arctic regions|xDiscovery and exploration|vInterviews	<dcvalue element="subject" qualifier="lcsh">Arctic regions -- Discovery and exploration -- Interviews</dcvalue>
651 0 Polar regions|xDiscovery and exploration|vInterviews	<dcvalue element="subject" qualifier="lcsh">Polar regions -- Discovery and exploration -- Interviews</dcvalue>
651 0 Antarctica|xDiscovery and exploration|vInterviews	<dcvalue element="subject" qualifier="lcsh">Antarctica -- Discovery and exploration -- Interviews</dcvalue>
793 0 Antarctic Deep Freeze Oral History Project	<dcvalue element="relation" qualifier="ispartofseries">Antarctic Deep Freeze Oral History Project</dcvalue>
n/a	<dcvalue element="description" qualifier="none">The Antarctic Deep Freeze oral history project…[truncated].</dcvalue>
n/a	<dcvalue element="description" qualifier="sponsorship">National Science Foundation</dcvalue>
n/a	<dcvalue element="description" qualifier="sponsorship">Antarctic Deep Freeze Association</dcvalue>
n/a	<dcvalue element="publisher" qualifier="none">Byrd Polar Research Center Archival Program</dcvalue>
n/a	<dcvalue element="rights" qualifier="none">Restrictions: This item is not restricted</dcvalue>
n/a	<dcvalue element="type" qualifier="none">Transcript</dcvalue>