Notes on Operations

Automated Metadata Harvesting: Low-Barrier MARC Record Generation from OAI-PMH Repository Stores Using MarcEdit

By Terry Reese

For libraries, the burgeoning corpus of born-digital data is becoming both a blessing and a curse. For patrons, these online resources represent the potential for extended access to materials, but for a library's technical services department they represent an ongoing challenge, forcing staff to look for ways to capture and make use of available metadata. This challenge is exacerbated for libraries that provide access to their own digital collections. While digital repository software like DSpace, Fedora, and CONTENTdm expose bibliographic metadata through the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH), few organizations have a simplified method for harvesting and generating Machine-Readable Cataloging (MARC) records from these metadata stores. Fortunately, a number of tools have been developed that can facilitate the harvesting and generation of MARC data from these OAI-PMH metadata repositories. This paper will examine resources that enhance technical services staff's ability to use existing metadata, with specific focus on one of these current generation tools, MarcEdit, which was developed by the author and provides a one-click harvesting process for generating MARC metadata from a variety of metadata formats.

On December 11, 2007, Perry Willett, head of the Digital Library Production Service at the University of Michigan (UM) Library, posted a message to the XML4Lib electronic discussion list indicating that metadata for the public domain materials made available through the UM Library Google Books project were now available for Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) download.[1] The OAI-PMH protocol was primarily developed as a low-barrier method for interoperability between metadata repositories. Using the protocol, structured bibliographic metadata can be shared between repositories and other metadata harvesters. The announcement was significant in two ways. First, it represents the first such announcement by a member of the Google Books collaboration. Second, the announcement underscores a growing trend in digital library development—widespread harvestability of a project's digital items and its metadata. Announcements such as these represent a boon for libraries and their patrons. As more collections move into the digital space, library patrons cannot help but benefit. However, for library technical services offices, announcements such as this can present new challenges. This paper considers options for handling these challenges by focusing on one tool, MarcEdit (http://oregonstate.edu/~reeset/marcedit).

As more digital services like the UM Library Google Books project move their metadata into the public Web space, library technical services departments need to determine how they will make use of this new influx of available metadata. For sure, some libraries have become accustomed to the many issues dealing with non–MARC metadata within what is still largely a MARC–centric

Terry Reese (terry.reese@oregonstate.edu) is Cataloger for Networked Resources, Digital Production Unit Head, Oregon State University Libraries, Corvallis.

Submitted April 22, 2008; tentatively accepted, pending revision, and returned to author July 2, 2008; revised and resubmitted October 20, 2008, and accepted for publication.

universe. For libraries hosting digital collections or institutional repositories, challenges related to the representation of those digital objects within a library's many discovery tools like the OCLC Online Computer Library Center's (OCLC) WorldCat or local integrated library system (ILS) are commonplace. While most digital collection software (for example, DSpace, Fedora, and CONTENTdm) and many vendor product solutions (like NewsBank's Congressional Serials Set) provide the ability to harvest item metadata by using OAI-PMH, few libraries use these metadata streams to generate MARC records. The process of downloading, converting, and managing metadata records beyond the traditional MARC metadata workflow remains largely unexplored in many libraries. For those that do repurpose non–MARC metadata in some way, the process is often limited to a single service or metadata stream. For example, both Texas A&M University Libraries and the University of Virginia (UV) Library documented their efforts to develop site-specific metadata harvesters for converting bibliographic metadata for electronic theses and dissertation records submitted to their institutional repositories into MARC.[2]

Non–MARC metadata models for sharing digital metadata are not likely to disappear, and technical services departments will need to adjust to new forms of metadata acquisition. During the past twenty years, OCLC and the Library of Congress (LC) have provided libraries with a single, centralized metadata repository from which to download bibliographic metadata. While OCLC remains the largest database of available bibliographic content, the actual distribution of metadata today is becoming much more decentralized. Institutional repositories and digital collection software have played a role in moving the library from metadata consumers and creators to metadata distributors. For libraries looking to leverage content housed in digital collections, understanding and developing processes of harvesting and converting non–MARC metadata will be essential for moving forward.

Together, the Open Archive Initiative (OAI, www.openarchives.org) and library communities have worked in recent years to provide a number of tools to facilitate the harvesting and conversion of OAI-PMH–compliant metadata into other delivery formats, both non–MARC and MARC. Traditionally, these tools have been released as parts of "kits" or components that library developers could use in specialized conversion tools. However, while these tools and kits have provided library information technology (IT) departments greater access to bibliographic metadata, they have done little to help technical services departments deal with OAI-PMH data. More recently, OCLC released an updated version of its Connexion software that provides limited capabilities for metadata harvesting of up to one hundred records through OAI-PMH; the software supports various flavors of Dublin Core (DC). This is a step in the right direction, but it provides no flexibility for customizing the data conversion itself, thus making record creation a one-size-fits-all process. The flexible nature of non–MARC metadata formats coupled with the lack of a formal standard for inputting metadata within non–MARC formats has made metadata creation somewhat uneven and not easily managed using a generic conversion process. The issue is well known in cataloging circles, as noted in an article found in *Online Libraries and Microcomputers*.[3] Here the author notes the many challenges one encounters when attempting to crosswalk metadata from one format to another. The one-size-fits-all approach to metadata is problematic because of issues related to granularity and consistency. Crosswalking metadata from one level of granularity to another is always difficult. For example, when moving from a schema of high granularity like MARC to a less granular schema like DC, the loss of both bibliographic content as well as context is often unavoidable. For instance, MARC 21 has numerous fields to represent the "author" of an item with each field containing contextual information about that "author." In unqualified DC, this context and granularity is lost because all "authors" are placed into a single dc:author element. Likewise, metadata of lower granularity cannot easily be moved to schemas with higher granularity because context and content cannot be manufactured if it is not present within the original record. Second is the issue of consistency. Although all DSpace and CONTENTdm software platforms use DC as the method for primary markup, the best practices used when generating metadata vary widely, potentially varying between projects within a single institution. The lack of a national standard or shared best practices when creating non–MARC metadata has contributed to a high level of inconsistency in the metadata currently being produced. This inconsistency makes capturing subtle relationships expressed within the metadata difficult and can result in overly broad and only marginally useful MARC records generated using these generic translation processes.

Seeing a need for a process that both flexibly and reliably converts metadata from OAI-PMH metadata stores into bibliographic formats usable by its online catalog, Oregon State University (OSU) Libraries chose to use MarcEdit, a freely available client application (developed by the author) that offers default conversion support from OAI-PMH metadata to a number of different metadata formats. This paper will provide a brief discussion of MarcEdit's metadata harvesting functionality as well as provide a detailed description of two potential use cases. The first example details how OSU Libraries catalogers use MarcEdit to harvest unqualified DC metadata from electronic theses in its institutional repository and automatically generate MARC 21 records for inclusion into both the online catalog and WorldCat. Through this conversion process, OSU Libraries has been able to avoid expensive effort duplication and, more importantly, has developed a simple workflow that can be used by technical services staff to capture OAI-PMH metadata from any OAI-PMH provider and generate records for the OSU Libraries catalog or OCLC. The second example demonstrates how staff can generate MARC records from the UM Library Google Books metadata. The process will detail some of the problems that can be encountered while working with metadata from remote metadata repositories as well as ways of overcoming those challenges.

Literature Review

Given the pervasiveness of Extensible Markup Language (XML)–based metadata and the wide range of protocols that support and advertise the presence of available metadata, it is surprising that automated metadata harvesting, MARC record generation, and library staff–centric tools development is not more frequently addressed in the literature. Several articles detail the process of indexing and harvesting MARC data into other indexing systems like Solr (http://lucene.apache.org/solr). Likewise, tools like Villanova's VuFind (www.vufind.org) and UV Library's Project Blacklight (http://blacklight.betech.virginia.edu) have advanced discussions relating to MARC indexing outside of a non–MARC environment. Only a few articles discuss processes for reusing XML–based metadata formats in MARC environments, and fewer still have been written specifically for technical services staff. Most have concentrated on the potential for reusing existing metadata in one's institutional repository to generate MARC records for submitted electronic theses and dissertations.

Surratt and Hill's article on the development of a customized ETD2MARC processing documented how Texas A&M University Libraries was able to customize a process developed by UV Library to provide a semi-automated record generation tool. Integrated into their workflow, the tool provided a way for staff to automatically generate MARC records for items as they were submitted into their institutional repository.[4] The resulting files from the metadata translation were dirty, core-level MARC records, which were then reviewed and edited by a staff member and finally entered in the online catalog and sent to OCLC. Texas A&M University Libraries' conversion script allowed their catalogers to more efficiently process electronic theses and dissertations (ETDs) by making use of attached metadata. While the article provided a copy of the script used to perform the conversion process, little evidence suggests that other institutions were able to use the Texas A&M University Libraries' method to promote metadata repurposing at their own institutions. The reason lies in the implementation. The process documented by Surratt and Hill fulfills the needs of the organization but is so tightly coupled to the organization's workflow that it becomes unusable without significant revision when taken outside of that environment. In addition, the process of data conversion was moved outside of technical services, meaning that a firewall was placed between the catalogers and the developers that created the script.

An article by Kurth, Ruddy, and Rupp documents an ongoing metadata repurposing project at the Cornell University (CU) Library. Unlike the process documented by Surratt and Hill, the CU Library project looks at the development of a service to repurpose MARC metadata for use within one's digital library infrastructure.[5] Kurth, Ruddy, and Rupp note that metadata currently found within the online catalog could be used to enrich many of the digital services and projects at CU Library. However, to use this metadata, a system needed to be developed that broke down MARC metadata and reassembled it for use in the Text Encoding Initiative (TEI) and DC. What makes this system interesting is the cooperative relationship between CU Library's metadata services and its IT department. While the article notes that the IT department develops and maintains the MARC processing scripts and document type definitions (DTD) for validation and creates the Extensible Stylesheet Language Transformations (XSLTs) used to crosswalk MARCXML data to TEI or DC, the collection-specific MARC mappings were created in conjunction with stakeholders from within the library. Since metadata conversions feed metadata directly to specific digital projects, the conversion must be completely automated. In this case, that is possible because of the controlled nature of the metadata and the granularity of the destination metadata schema.

A 2005 article by this author in the *Journal of Map and Geography Libraries* described the process used by OSU Libraries to generate MARC records for Geographic Information Systems (GIS) datasets from the accompanying Federal Geographic Data Committee (FGDC) metadata records.[6] Using MarcEdit, OSU Libraries was able to create a generic XSLT stylesheet that could be used as a template for translating FGDC metadata to MARC 21 XML. Once in MARC 21 XML, MarcEdit is able to translate the metadata into MARC 21 as well as accommodate character set translations between the legacy MARC-8 and more current 8-bit Unicode Transformation Format (UTF-8). Because of the richness of data found within the FGDC data format, the MARC records generated from the

FGDC data sources often included much more detailed information than records generated without the FGDC metadata. Although this process is not fully automated because records are not harvested and translated automatically, the process is portable.

First, MarcEdit uses a framework that allows metadata, once converted to MARC 21 XML, to be translated to any metadata format registered with the application. For Oregon State University, that meant that once the FGDC crosswalk was developed, catalogers could produce records in MARC, MARC 21 XML, DC, DC Qualified, Metadata Object Description Schema (MODS), and Encoded Archival Description (EAD) records from a single FGDC source. Second, the user of the application has full control over the crosswalk itself, meaning that the cataloger is free to modify the conversion rules. This allows the cataloger to control the conversion between metadata formats with a greater level of granularity.

While the literature and toolsets for technical services focus on uses of existing XML–based metadata formats within existing MARC environments, a great deal of literature exists outside technical services on harvesting and repurposing metadata for the development of external services and metasearch repositories. Articles like Simons and Bird's "Building and Open Language Archives Community on OAI foundation" or Suleman and Fox's "Leveraging OAI Harvesting to Disseminate Theses" look at the OAI-PMH standard and the role that it can and has played in setting up large, ad hoc document communities.[7] Several data aggregations such as UM's OAIster (www.oaister.org) project, which provides a single point of query for more than 19 million records (as of December 2008), or Emory University's AmericanSouth.org project (now ceased), which focused on the aggregation of cultural and historical content, have been based on the concept of harvesting available metadata and repurposing it to draw connections and build virtual collections and local aggregations.[8] Out of these projects have come tools and frameworks that can be used to build additional metadata aggregations and services. The Metadata Migrator (www.metascholar.org/sw/mm), a self-contained application designed as a crosswalk for and generator of DC data files and that can be served as part of an OAI-PMH repository, is one such resource to come out of the MetaScholar initiatives (www.metascholar.org), a digital library project at Emory University. Many exemplary projects like Picture Australia (www.pictureaustralia.org) and the Networked Digital Library of Theses and Dissertations (www.ndltd.org) have been developed through the aggregation and harvest of OAI-PMH metadata, demonstrating the availability of metadata for many of the digital items currently being generated by researchers and universities around the world. Libraries and their technical services departments could take a cue from these projects as they look to collect and provide access to digital resources through their organization's primary discovery tools, which are often still online catalogs.

For libraries looking to use and expose their OAI-PMH–based metadata products, making the technical information about these resources available to the larger library community will continue to be a growing challenge. Metadata providers will need to consider how discovery takes place not just for items within their collections but also for the digital services that expose those collection. For this reason, projects such as OCLC's Digital Registry (www.oclc.org/registry), the Ockham Initiative (www.ockham.org), and the Joint Information Systems Committee Information Environment Registry (http://iesr.ac.uk) have worked to develop a flexible registry system for the sharing of technical metadata about digital collections. For technical services departments interested in reusing existing metadata for digital items, simply finding the information needed to access and capture that metadata may be a significant barrier that will continue to exist inthe immediate future. Fortunately, a number of open OAI metadata repositories are being developed to fill this need. In their article, "Current Developments and Future Trends for the OAI Protocol for Metadata Harvesting," Shreeves and colleagues made note of a number of OAI repositories being developed into a comprehensive knowledge base capable of providing the technical information users need to harvest metadata.[9] As they observed, metadata repository development represents the likely future for the OAI-PMH community as data harvesters look for reliable ways to retrieve technical information about a given metadata community and to discover other communities and projects that may be related.

Available Tool Sets

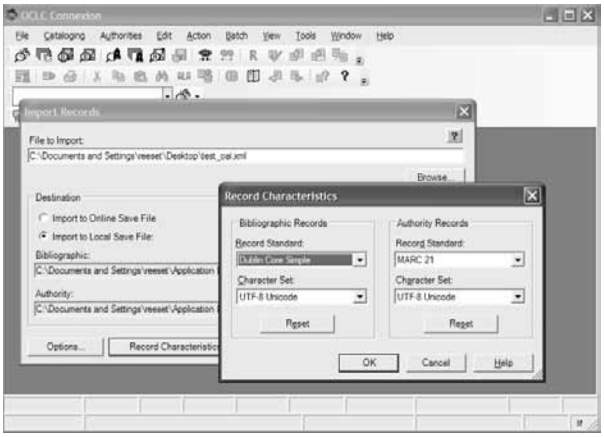
Presently, a number of effective development toolsets and software kits exist to provide OAI functionality to library tools. Developers interested in working with the OAI-PMH protocol are able to choose from components developed in a variety of languages, such as the Perl OAI modules, the Ruby OAI gem, or one of the many Java OAI harvesting kits; components have been readily available for some time for developers looking to build resources to aggregate metadata together. The OAI keeps track of a number of user-contributed tools and toolkits.[10]

The primary purpose of this paper is to look at resources that enhance technical services staff's ability to take advantage of existing metadata, not to examine resources developed for the developer community. While a rich ecosystem of developer-related tools exists for processing OAI-PMH metadata, these tools provide very little practical benefit to most technical services staffs and departments. To address this absence, this paper highlights two main classes of metadata harvesting tools currently available for technical services staff who wish to work with non–MARC metadata.

Innovative Interfaces' XML Harvester

Presently, many vendors have or are developing tools to facilitate the harvest of non–MARC metadata into the online catalog. ILS vendors like Innovative Interfaces have moved to create systems to streamline metadata harvesting directly into the online catalog. The Innovative Interfaces metadata solution known as XML Harvester, developed in cooperation with Michigan State University (MSU), is representative of most ILS vendor-supplied data harvesting tools because it provides one-way metadata conversion from a single data source into the online catalog. XML Harvester was used initially by MSU to generate MARC records in the online catalog from harvested EAD metadata, although today it can provide conversions from a number of different metadata formats.

XML Harvester's functionality is representative of most ILS vendor-supplied metadata harvesting applications. Since this class of applications tends to run at the server level, control over how metadata crosswalking is defined will vary in granularity and generally be available only to IT staff or those at the system level. Likewise, this class of tools tends to be designed to be single project solutions, meaning that a significant amount of time is generally required for set up and testing to harvest a single collection, overhead that must be reallocated each time a new collection is set to be harvested. Because translations are tailored to specific projects or collections, work done for one project cannot be shared or used when looking to harvest other collections. This places practical limits on the types of projects that these tools can support. While the tight coupling with the ILS generally simplifies the process of loading and updating harvested metadata, it does come at a price. XML Harvester, for example, can only be used to harvest metadata into the online catalog and Encore Platform rather than as an abstract harvesting tool for providing metadata conversion services. This does tend to put very specific limits as to how useful this class of tools can be in general, particularly when considering the wide range of databases and services library technical service departments are being asked to maintain. The ability to harvest metadata and convert it into many different formats will likely become more important with time, possibly shortening the shelf life for this class of applications.



File to Import: C:\Documents and Settings\veeset\Desktop\test_oai.xml

Destination

○ Import to Online Save File

● Import to Local Save File:

Bibliographic: C:\Documents and Settings\veeset\Application

Authority: C:\Documents and Settings\veeset\Application

Bibliographic Records

Record Standard: Dublin Core Simple

Character Set: UTF-8 Unicode

Authority Records

Record Standard: MARC 21

Character Set: UTF-8 Unicode

Figure 1. OCLC Connexion Metadata Extraction Tool

OCLC's Connexion

Some in the vendor community are beginning to provide better support for non–MARC metadata formats. For catalogers, the most interesting recent development for this is the inclusion of a metadata harvesting and crosswalking tool directly into OCLC's Connexion product (www.oclc.org/connexion). Given OCLC's influence and large number of member libraries, its software has the potential to simplify the metadata harvesting process for numerous libraries as well as lower the barriers to getting digital object metadata into the WorldCat database. As of version 2.10, the Connexion client provides OCLC members a set of basic metadata harvesting functionalities able to process records in a variety of DC flavors. The software is unique in the vendor community because it represents one of the first attempts by a vendor to shift responsibility for metadata harvesting and reuse from a library's IT staff to its technical services staff. Nevertheless, the current implementation offers little practical functionality.

Figure 1 illustrates the current functionality provided to the user. Presently, users wanting to automatically generate MARC records from OAI DC records must download the record set locally before initiating this process. For large datasets, like UM Library's Digital Books project, this workflow would be unfeasible because each OAI request returns only five hundred items. For example, using this method to generate the nearly one hundred thousand records made available through UM Library's Digital Books project would require harvesting the dataset two thousand times.

One of the most unique aspects of the OCLC's approach has been the decision, at least initially, to hide the metadata conversion process from the user. While this simplifies the overall metadata conversion process, it introduces a "fast food" approach to metadata conversion and is the process's greatest weakness. Given the number of ways that DC elements can be interpreted and implemented

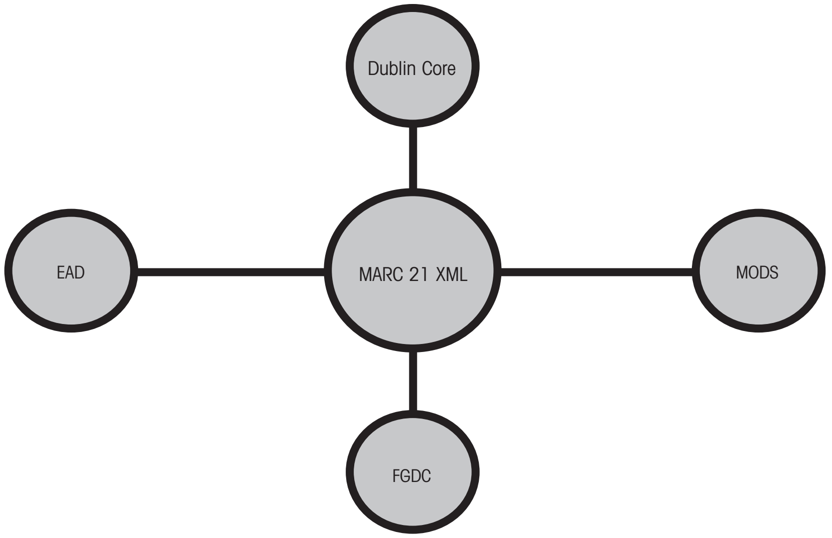


Figure 2. MarcEdit Spoke-and-Wheel Design

between collections within a single organization, such a one-size-fits-all approach to metadata extraction and generation is not likely to be useful for meaningful record generation. Despite its limitations, the OCLC Connexion metadata extraction tool is still a significant step toward mainstreaming metadata crosswalking for technical services staff.

MarcEdit

Overall, the vendor community has been making great strides toward simplifying the process of working with harvested metadata sets. However, at this point, their efforts still remain very project-based, making them marginally useful as general metadata conversion tools for the diverse datasets available to the library community. Each serves a need, but, as general metadata harvesting and conversion tools, their inability to allow catalogers to control metadata harvesting and customize the conversion rules is a serious impediment to adoption. For these reasons, OSU Libraries has used MarcEdit for its data harvesting and conversion needs.

MarcEdit is a freely available, metadata editing suite initially conceived in 1998 as a graphical user interface (GUI) replacement for the LC's DOS–based MARCBreakr and MARCMakr software. Originally designed primarily as a batch MARC editing tool, the program expanded the functionality found in MARCBreakr and MARCMakr by including the MarcEditor, a notepad designed specifically for the modification of batch MARC records. Metadata needs and formats have changed significantly since 1998, and MarcEdit has changed with them. Today, the name MarcEdit is almost a misnomer because the application no longer is simply a batch MARC editing tool. Instead, MarcEdit is an application suite of metadata editing tools, including character set conversion, XML crosswalking, and metadata harvesting.

In many respects, MarcEdit has a number of things in common with OCLC's Connexion application. They are both client-side applications, empowering users to work with data from many different sources. Likewise, the applications work with the OAI-PMH protocol and provide built-in data conversion rules for supported metadata formats. However, MarcEdit takes this one step further by providing users with the ability to customize the existing data conversion rules or create new data conversion rules. This allows users to harvest metadata from one of the supported metadata formats (DC, MODs, OAI MARC, or MARC 21 XML) as well as create conversion templates for additional metadata formats. It also allows users to customize existing conversion templates to reflect many variations in best practices used between projects. Users are given this customizability through XSLT. All of MarcEdit's metadata conversion rules are defined as XSLT templates.

Appendix A presents the entire XSLT stylesheet used for converting OAI MARC records to MARC 21 XML. This is a good example because it underlines how readily available this type of crosswalking information already has become. This particular stylesheet was derived from an XSLT stylesheet provided by the LC and is one of many such examples currently available to the library community.[11] Why a conversion to MARC 21 XML? MarcEdit uses a "wheel-and-spoke" method, with MARC 21 XML sitting at the center of that wheel. This architecture allows metadata conversions to be created without the need to know directly how the individual metadata elements relate to elements within different schemas. Once a new spoke has been added to the wheel, it becomes crosswalkable to any other spoke on that wheel.

Figure 2 provides an illustration of this approach. Using this model, an EAD record could be translated to any other metadata schema on the wheel without the need to know how the elements in the EAD record relate to elements in the destination format. A user simply needs to modify or create a new XSLT template to modify the formats and behaviors of MarcEdit's metadata conversion process. At one time, finding technical services staff with the ability to modify or create an XSLT document may have been an impediment, but the ubiquitous nature of XSLT has made this skill

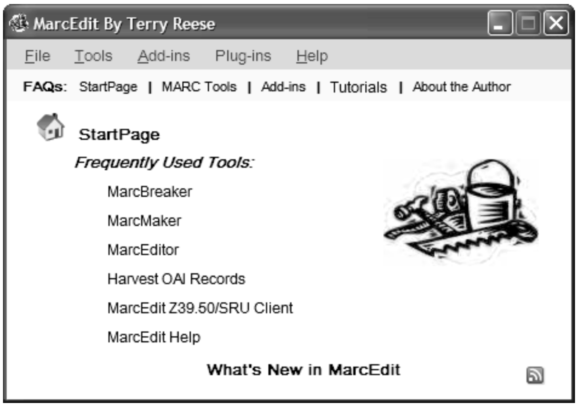


Figure 3. MarcEdit Welcome Screen

set much more common within the library community. At OSU Libraries, this functionality is what that makes MarcEdit's method for metadata harvesting so valuable. Given the dearth of metadata currently available in DC, the ability to customize metadata conversion rules is essential to accommodating the variety of best practices and input standards. Staff members also have the option of either accepting the template-generated metadata or continuing to be active participants in the metadata creation process.

Harvesting from OAI

Like Connexion, MarcEdit simplifies the process of harvesting OAI-PMH–based metadata. Upon startup, users are greeted with the MarcEdit welcome screen (see figure 3), which includes links to commonly used functionality. Here one will find a link to MarcEdit's OAI Data Harvester, which initiates the data harvesting service (see figure 4).

Once initialized, the user needs to provide the Metadata Harvester with only the host (URL) and set (collection name) for the set of records to be harvested. Users can optionally change the metadata type being requested from the server as well as define their own set of translation rules. Once set, the MarcEdit Metadata Harvester captures and translates the set's metadata records from the defined metadata type to MARC. No interaction is required by the user. Users who wish to do more granular data harvests can select the advanced settings link to use some of the optional parameters supported within the OAI-PMH specification. The advanced settings function reveals a cache of additional options that can be set to define what records are to be harvested by the Metadata Harvester (see figure 5).

Using the advanced settings, users have the ability to define a subset of records (using Start and End), individual records (using GetRecord) or resume harvesting a predefined record set (using the ResumptionToken). Additionally, the harvester can translate record data from Unicode to MARC-8 as well as simply harvest and save the raw XML metadata files to a local file system. The character conversion options should be of special value for libraries that still use systems that cannot load or recognize MARC records encoded in UTF-8. Functionality has been added for users wanting to harvest XML–based metadata and create records using the legacy MARC-8 character set. Again, users need not set any of these options to harvest OAI-PMH metadata, but they are available for more granular data capture.

Case Study: OSU Libraries Electronic Theses and Dissertation Record Generation

Getting Started

In January 2007, OSU joined a growing fraternity of universities whose students must submit electronic copies of their theses or dissertations in order to graduate. This policy shift by the graduate school was met with great excitement by OSU Libraries, which would take on the role of preserving and providing access for these materials through the library's institutional repository (IR) portal, ScholarsArchive@OSU. Within the IR, these materials could find a larger audience both inside and outside the university, potentially extending the reach of the research being done by the university.

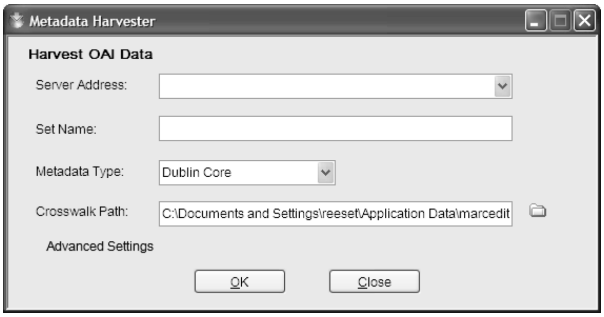
With these changes came a number of challenges for OSU Libraries' technical services department. Like most institutions, OSU Libraries had traditionally created original MARC records for OSU theses, adding the MARC records to the local ILS as well as the WorldCat database. Cataloging for these records was done as materials were submitted to OSU Libraries by the graduate school; technical services staff usually received all of a term's theses at one time. All record creation was performed using Connexion, meaning records were created once, dynamically becoming part of WorldCat, and then downloaded directly to the local library catalog.

The submissions of the theses and dissertations in electronic format, however, would be a much different process. First, unlike traditional print documents, electronic theses and dissertations would be submitted into the IR at any point during the term. Materials would first be vetted by the graduate school, then released to OSU Libraries, where they would be evaluated and then be made public. Technical services staff could no longer allocate fixed processing time for handling theses and dissertations because materials now would not be submitted on a fixed schedule. Secondly, metadata creation for these documents would shift from technical services staff to the document creators. When documents are submitted into the IR, submitters are required to provide metadata including abstracts and keywords. As a result, technical services staff began to use the metadata stored within the IR as the primary bibliographic data of record, meaning that metadata creation no longer took place in Connexion and was no longer being generated in MARC. This left technical services with two options: catalog materials twice (once in the IR and once in Connexion) or design a process that would allow metadata entered into the IR to be loaded directly into both WorldCat and the local catalog. Ultimately, the library choose the second option, not only to avoid the expense of rekeying records but also to support efforts to design processes that repurpose metadata rather than rekeying. As a side benefit, by using the metadata entered into the IR, the library was able to take advantage of an entirely new set of metadata elements: user contributed keywords and descriptions of the document. For materials like theses or dissertations, this information can be invaluable given the timeliness of topics, many of which are yet to be represented well within existing control vocabularies like the LC Subject Headings (LCSH).

Submission Process

The submission process for OSU Libraries' ETD program mirrors those used by a number of other institutions using DSpace as their IR platform. Materials are submitted directly into the IR by their authors; in this case, it is the graduate or PhD candidate. Each submitter answers a number of questions through the submission process, entering information about their paper and topic. This metadata forms the foundation for the MARC records creation later in the process. The submitter then chooses a distribution license and uploads the document to the IR.

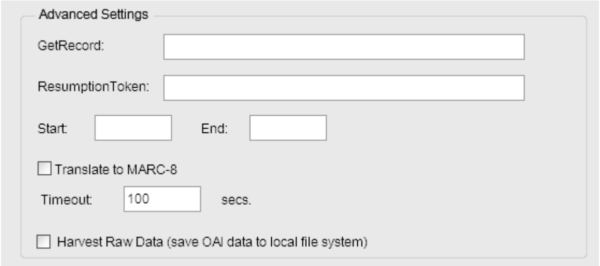
Once the item has been submitted to the IR, it is then vetted by the university's graduate school. This process ensures that only approved materials are actually archived in the IR. Once approved, the item is then forwarded into the IR's work queue, which is managed by technical services. At this point, a library staff member does original cataloging for the item in the IR. He or she then validates the user-submitted metadata and enters LCSH subject terms and any necessary descriptive notes for the document. When finished, the material is published to the IR and made available to the larger research community.



Metadata Type: Dublin Core

Crosswalk Path: C:\Documents and Settings\reeset\Application Data\marcedit

Figure 4. MarcEdit OAI-PMH Metadata Harvester



☐ Translate to MARC-8

Timeout: 100 secs.

☐ Harvest Raw Data (save OAI data to local file system)

Figure 5. Advanced Settings Function in MarcEdit Metadata Harvester

The question of whether MARC records are still needed was one with which OSU Libraries has struggled for some time. While having the records within both the local ILS as well as the WorldCat database was ideal, the reality was that staff simply did not have time to rekey data to create the MARC records. Moreover, given the increased accessibility of these documents through Web browsers like Google, questions arose regarding the need to continue producing MARC records. In the end, the library decided that having the data in WorldCat was important and set out to build a workflow that would keep staff from having to rekey the metadata from the IR into Connexion.

Automatic MARC Record Generation

MarcEdit offered a solution to the rekeying issue. As a DSpace repository, the library's IR could provide metadata for new and modified records entered into the IR through OAI-PMH. Unfortunately, the classified staff ultimately responsible for creating MARC records for IR items had no experience working with OAI-PMH or repurposing metadata from one schema to another. The ideal tool would automatically generate records from the metadata while essentially hiding this interaction from staff.

To meet this need, a special XSLT crosswalk derived from the default template was created to translate the DC metadata used for ETDs to their equivalent MARC fields.[12] This crosswalk varied from the vanilla DC-to-MARC 21 XML crosswalks provided by the LC because it used positioning within the metadata record to determine the context of some of the record's metadata, since all metadata being harvested through OAI-PMH was unqualified DC. Using information about the generated metadata, an XSLT stylesheet was created to restore context to the harvested metadata. Once the context for these metadata elements was reestablished, the ability to create good MARC 21 records became much easier.

After being created, this custom XSLT transformation was registered with the MarcEdit application, allowing staff to use a simple, wizard-based application to harvest OAI-PMH records and convert them directly into MARC. This process has allowed staff to develop a workflow that allows them to immediately process items when submitted to the IR while doing MARC record generation for those items at the end of each week. The records are reviewed for accuracy and then uploaded directly to WorldCat and the local ILS. Appendix B provides an example of records currently generated with the OSU Libraries record-generation process. The example demonstrates how Unicode characters outside of the MARC-8 character set are embedded into the document as well as how subjects and headings are analyzed to create records that will require little or minimal manual intervention during the later steps of the process. This process has become more and more automatic as time has progressed, and the XSLT stylesheet has been refined to allow for minimal review prior to uploading metadata to the local catalog and WorldCat.

Problems and Solutions

During the early testing phases of this process, several potential problems needed to be resolved. The first and most important issue was related to character encodings. All data loaded into the IR and harvested through OAI-PMH was encoded in UTF-8. While certainly desirable, OSU Libraries' local ILS was not set up to recognize Unicode data in MARC records. Likewise, presently, the OCLC's own importing tools will not allow for the import of Unicode data within MARC records. This meant that whatever harvesting tool OSU Libraries used to generate MARC records had to be able to facilitate some form of character set remapping between Unicode and MARC-8. Fortunately, MarcEdit's OAI Harvester includes the ability to remap metadata from Unicode to MARC-8 on the fly, quickly solving this issue for the library.

The harvesting of granular metadata using unqualified DC remains an issue today. While much of the context lost because of the generic nature of unqualified DC can be reclaimed through careful analysis of the metadata, one ambiguity that cannot be easily resolved is the differentiation between staff-submitted LCSH subject terms and user-submitted keywords. While the XSLT stylesheet can be coded to make a very good educated guess as to the nature of these elements, the ambiguity persists enough that review is required following record generation. Fortunately, a solution to this issue has been created by the DSpace community. The recently released DSpace 1.5 simplifies the inclusion of additional supported metadata formats for harvest through OAI-PMH. This should allow OSU Libraries to modify the XSLT transformation so that it uses a more granular XML schema for harvesting, such as qualified DC or MODs and so that it retains the context associated with each harvested element producing production-ready MARC records from the harvest.

Use Case: Automatic MARC 21 Record Generation for Remote Resources

Once established, the ability to harvest and repurpose metadata can fundamentally change how librarians collect materials. By lowering the barriers for creating MARC or, alternatively, other forms of records for addition to an institution's discovery application, technical service departments can empower collection development staff to evaluate a wider range of the electronically available materials being produced by research institutions. Likewise, technical services departments can represent documents within their local ILS that would have previously been considered out of reach. Examples of this at OSU Libraries are numerous, ranging from the capture of documents from a sister institution's IR to automated harvesting and record generation of tens of thousands of digital documents stored within CONTENTdm. Outside of OSU Libraries, libraries are starting to consider how they can leverage OAI-PMH metadata made available from vendors like NewsBank or how they can capture and represent tens of thousands of records from free metadata repositories like Project Gutenburg (www.gutenberg.org) or the LC's American Memory Project (http://memory.loc.gov/ammem).

One specific remote metadata set currently of interest to a growing number of institutions is UM Library's digital collections, or, more specifically, metadata from its Hathi Trust Digital Library (formerly MBooks) project (www.lib.umich.edu/mdp) for materials currently being scanned through Google's Book Scanning project. In December 2007, UM Library announced that it would be making available OAI-PMH harvestable metadata records for all the public domain materials captured through the project.[13] For the library community, this decision was significant because it was the first to come from any of the institutions partnering with Google. Not surprisingly, many libraries have started looking at how this content can be captured and loaded into their ILS systems.

While many libraries likely would have preferred that UM Library simply provide large downloadable metadata sets in MARC 21 format, they have essentially done this by making the metadata available for harvesting. While the OAI-PMH protocol requires that metadata be provided at least in unqualified DC, it does support the ability for metadata providers to make

```
<?xml version=”1.0” encoding=”UTF-8”?>
<OAI-PMH xmlns=”http://www.openarchives.org/OAI/2.0/” xmlns:xsi=”http://www.w3.org/2001/XMLSchema-instance”
   xsi:schemaLocation=”http://www.openarchives.org/OAI/2.0/ http://www.openarchives.org/OAI/2.0/OAI-PMH.xsd”>
 <responseDate>2008-12-23T13:10:31Z</responseDate>
 <request verb=”ListSets”>http://quod.lib.umich.edu/cgi/o/oai/oai</request>
 <ListSets>
  <set>
   <setSpec>hathitrust</setSpec>
   <setName>HathiTrust digital repository</setName>
   <setDescription>The “hathitrust” sets in this repository contain records from the HathiTrust digital repository,
  formerly MBooks. The HathiTrust digital repository is the access system to the digitized collections of some of the
  nation’s great research libraries. For more on the HathiTrust, view the website at http://www.hathitrust.org/.</setDe-
  scription>
  </set>
  <set>
   <setSpec>hathitrust:pd</setSpec>
   <setName>Public domain items worldwide</setName>
  </set>
  <set>
   <setSpec>hathitrust:pdus</setSpec>
   <setName>Public domain items according to copyright law in the United States</setName>
  </set>
  <set>
   <setSpec>dlps</setSpec>
   <setName>Digital Library Production Service (DLPS) digital objects</setName>
  </set>
  <set>
   <setSpec>dlpstext</setSpec>
   <setName>Digital Library Production Service (DLPS) text collections</setName>
  </set>
  <set>
   <setSpec>dlps:alajournals</setSpec>
   <setName>Abraham Lincoln Association Journals</setName>
  </set>
```

Figure 6. ListSets Response from the University of Michigan OAI-PMH Server

additional schemas available for harvest. UM Library has chosen to provide their metadata records both in DC and MARC 21 XML. Since MARC 21 XML records can be translated directly to MARC 21, one only needs to decide to harvest the metadata.

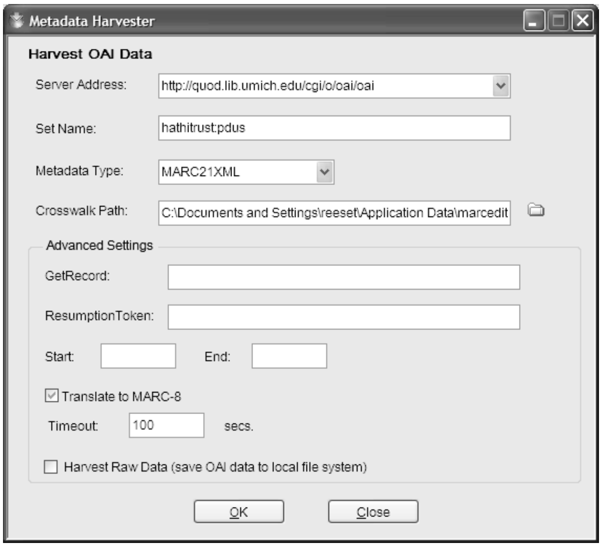
Using MarcEdit as an OAI harvester, the process is relatively simple. As noted above, OAI-PMH harvesting requires the definition of a base URL and the identification of the set to be harvested. UM Library is currently making metadata for numerous digital collections available through its OAI-PMH service, though it has separated its collection into three sets. These sets can be quickly identified by making a direct query to the OAI-PMH service, requesting the names and information needed to harvest the collection. In this case, the OAI-PMH command that would be used is the ListSets command. The ListSets command will return connection information about the collections being hosted on the server; figure 6 shows an sample response to the ListSets command.

The next step is to configure the MarcEdit to harvest the metadata. Figure 7 shows how to set the definitions in MarcEdit’s OAI-PMH Harvester. Using the configuration presented in figure 7, the MarcEdit OAI Harvester would capture all metadata records from the mbooks:pdus set, translate items from MARC 21 XML to MARC 21, and convert all UTF-8 data to MARC-8.

Several problems commonly occur during the harvesting process. The most frequent is server nonresponsiveness. During numerous test harvests of the UM Library collection, this author found that their OAI-PMH server would often drop harvesting requests after processing approximately one hundred thousand items. Most OAI-PMH harvesters provide some support for recovering from these types of failures, providing the information needed to resume the data harvest at the point where the connection was lost. MarcEdit uses a two-stage approach to recover from this error and resolve problems for users when possible. In the case of server nonresponsiveness, MarcEdit uses a tapered approach to data harvesting and issues multiple data requests at differing intervals to adjust for server nonresponsiveness. Additionally, if harvesting is stopped for any reasons, the user can resume harvesting metadata incrementally using the last resumption token processed by the software. This way, if the OAI-PMH server drops the harvesting connection, the request can be restarted where it stopped. In addition to dropped connections, other issues that may be present are errors within the metadata themselves (encoding errors) or the MARC–encoded data. When these records were first made available, a number of records within the harvested metadata set included invalid MARC data. Unfortunately, this represents a frequent problem with harvestable metadata. MarcEdit's OAI Harvester facilitates the correction and flagging of these types of issues by correcting metadata records with errors relating to the structural output of the records. For its part, UM Library quickly fixed these errors when reported, but invalid data elements are always an issue when dealing with metadata from remote sites.

Conclusion

As more institutions bring digital collections online, technical services staff will continue to face the growing issue of distributed metadata retrieval. Unlike their print cousins, today's institutional repositories and digital collections give rise to metadata of a distributed nature that require technical services departments to think creatively and produce workflows that encourage repurposing data. Tools like MarcEdit's OAI-PMH Harvester simplify that process for staff by allowing nontechnical users to harvest metadata in various formats without dealing with issues relating to XML validation or character encodings. For too long, technical services staff has viewed metadata harvesting and transformation as a job for library technology departments. As new tools and workflows continue to be developed, more technical services departments will likely turn to metadata harvesting and capture as a viable method of generating metadata for digital collections.



Server Address: http://quod.lib.umich.edu/cgi/o/oai/oai

Set Name: hathitrust:pdus

Metadata Type: MARC21XML

Crosswalk Path: C:\Documents and Settings\reeset\Application Data\marcedit

Advanced Settings

GetRecord:

ResumptionToken:

Start: End:

☑ Translate to MARC-8

Timeout: 100 secs.

□ Harvest Raw Data (save OAI data to local file system)

Figure 7. Setting Definitions in MarcEdit's OAI-PMH Metadata Harvester

References and Note

1. Perry Willett, "University of Michigan Announces OAI Harvesting of MBooks," online posting, Dec. 11, 2007, XML4Lib, http://lists.webjunction.org/wjlists/xml4lib/2007-December/005978.html (accessed Dec. 12, 2007).
2. Brian Surratt and Dustin Hill, "ETD2MARC: A Semi-Automated Workflow for Cataloging Electronic Theses and Dissertations," *Library Collections & Technical Services* 28, no. 2 (2004), also http://handle.tamu.edu/1969.1/588 (accessed Jan. 22, 2008); Cristina W. Sharretts and James C. French, "Electronic Theses and Dissertations at the University of Virginia Library," *International Conference on Digital Libraries: Proceedings of the Fourth ACM Conference on Digital Libraries*, 246–47 (New York: ACM, 1999), http://doi.acm.org/10.1145/313238.313429 (accessed Dec. 22, 2008).
3. "Challenges and Issues with Metadata Crosswalks," *Online Libraries & Microcomputer* 20, no. 4 (Apr. 2002): 1–4.
4. Surratt and Hill, "ETD2MARC: A Semi-Automated Workflow for Cataloging Electronic Theses and Dissertations."
5. 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): 154–65, http://lts.library.cornell.edu/lts/who/pre/upload/p153.pdf (accessed Dec. 22, 2008).
6. Terry Reese, "Bibliographic Freedom and the Future Direction of Map Cataloging," *Journal of Map & Geography Libraries* 2, no. 1 (2005): 67–97, http://ir.library.oregonstate.edu/dspace/handle/1957/16 (accessed Dec. 22, 2008).
7. Gary Simons and Steven Bird, "Building an Open Language Archives on the OAI Foundation," *Library Hi Tech* 21, no. 2 (2003): 210–18; Hussein Suleman and Edward A. Fox, "Leveraging OAI Harvesting to Disseminate Theses," *Library Hi-Tech* 21, no. 2 (2003): 219–27.
8. Martin Halbert, "The Metascholar Initiative: AmericanSouth.org and MetaArchive.Org," *Library Hi Tech* 21, no. 2 (2003): 182–98.
9. Sarah L. Shreeves et al., "Current Developments and Future Trends for the OAI Protocol for Metadata Harvesting," *Library Trends* 53, no. 4

(Spring 2005): 576–89.

10. Open Archives Initiative, PMH Tools, www.openarchives.org/pmh/tools/tools.php (accessed Dec. 22, 2008).

11. Library of Congress, MARC 21 XML schema stylesheet, www.loc.gov/standards/marcxml/xslt/OAIMARC2MARC21slim.xsl (accessed Dec. 22, 2008).

12. Details are available in the author's description of this process: See Terry Reese, "Oregon State University Electronic Theses DSpace (OAI-PMH) to MARC21XML Crosswalk" http://hdl.handle.net/1957/6300 (accessed Dec. 22, 2008).

13. Willett, "University of Michigan Announces OAI Harvesting of MBooks."

Appendix A. OAI MARC to MARC 21 XML Conversion

```
<?xml version=”1.0” encoding=”UTF-8”?>
<xsl:stylesheet xmlns=”http://www.loc.gov/MARC21/slim” xmlns:oai=”http://www.openarchives.org/OAI/1.1/oai_marc”
version=”1.0” xmlns:xsl=”http://www.w3.org/1999/XSL/Transform” exclude-result-prefixes=”oai”>
<xsl:template match=”/”>
   <collection xmlns:xsi=”http://www.w3.org/2001/XMLSchema-instance” xsi:schemaLocation=”http://www.loc.gov/
MARC21/slim http://www.loc.gov/standards/marcxml/schema/MARC21slim.xsd” >
      <xsl:apply-templates />
   </collection>
</xsl:template>

<xsl:template name=”OAI-PMH”>
       <xsl:for-each select = “ListRecords/record/metadata/oai:oai_marc”>
    <xsl:apply-templates />
    </xsl:for-each>
    <xsl:for-each select = “GetRecord/record/metadata/oai:oai_marc”>
    <xsl:apply-templates />
    </xsl:for-each>
</xsl:template>

 <xsl:template match=”text()” />
<xsl:template match=”oai:oai_marc”>
   <record xmlns:xsi=”http://www.w3.org/2001/XMLSchema-instance” xsi:schemaLocation=”http://www.loc.gov/MARC21/
slim
   http://www.loc.gov/standards/marcxml/schema/MARC21slim.xsd” >
      <leader>
         <xsl:text> </xsl:text>
         <xsl:value-of select=”@status”/>
         <xsl:value-of select=”@type”/>
         <xsl:value-of select=”@level”/>
         <xsl:text> 22 </xsl:text>
         <xsl:value-of select=”@encLvl”/>
         <xsl:value-of select=”@catForm”/>
         <xsl:text> 4500</xsl:text>
      </leader>
      <xsl:apply-templates select=”oai:fixfield|oai:varfield”/>
   </record>
</xsl:template>
<xsl:template match=”oai:fixfield”>
   <xsl:element name=”controlfield”>
      <xsl:call-template name=”id2tag”/>
      <xsl:value-of select=”substring(text(),2,string-length(text())-2)”/>
    </xsl:element>
</xsl:template>

<xsl:template match=”oai:varfield”>
    <xsl:element name=”datafield”>
        <xsl:call-template name=”id2tag”/>

        <xsl:attribute name=”ind1”>
            <xsl:call-template name=”idBlankSpace”>
                        <xsl:with-param name=”value” select=”@i1”/>
            </xsl:call-template>
        </xsl:attribute>

        <xsl:attribute name=”ind2”>
            <xsl:call-template name=”idBlankSpace”>
                        <xsl:with-param name=”value” select=”@i2”/>
            </xsl:call-template>
        </xsl:attribute>

        <xsl:apply-templates select=”oai:subfield”/>
    </xsl:element>
</xsl:template>

<xsl:template match=”oai:subfield”>
    <xsl:element name=”subfield”>
        <xsl:attribute name=”code”>
            <xsl:value-of select=”@label”/>
        </xsl:attribute>
        <xsl:value-of select=”text()”/>
    </xsl:element>
</xsl:template>

<xsl:template name=”id2tag”>
    <xsl:attribute name=”tag”>
        <xsl:variable name=”tag” select=”@id”/>
        <xsl:choose>
            <xsl:when test=”string-length($tag)=1”>
                        <xsl:text>00</xsl:text>
                        <xsl:value-of select=”$tag”/>
            </xsl:when>
            <xsl:when test=”string-length($tag)=2”>
                        <xsl:text>0</xsl:text>
                        <xsl:value-of select=”$tag”/>
            </xsl:when>
            <xsl:when test=”string-length($tag)=3”>
                        <xsl:value-of select=”$tag”/>
            </xsl:when>
        </xsl:choose>
    </xsl:attribute>
</xsl:template>
<xsl:template name=”idBlankSpace”>
    <xsl:param name=”value”/>
    <xsl:choose>
        <xsl:when test=”string-length($value)=0”>
```

```
            <xsl:text> </xsl:text>
        </xsl:when>
        <xsl:otherwise>
            <xsl:value-of select=”$value”/>
        </xsl:otherwise>
    </xsl:choose>
</xsl:template>
</xsl:stylesheet>
```

Appendix B. Sample Generated Record

```
=LDR 02358ntm 2200337Ia 45 0
=008 051012s2008\\\\xx\a\\\\\bm\\\000\0\eng\d
=040 \\$aORE$cORE
=049 \\$aOREV
=090 \\$aLD4330 2008$bMarcum, Wade R.
=100 1\$aMarcum, Wade R.
=245 10$aThermal hydraulic analysis of the Oregon State TRIGA Reactor using RELAP5-3D /$cby Wade R. Marcum.
=260 \\$cc2008.
=300 \\$axx leaves : $bill. ; $c29 cm.
=500 \\$aPrintout.
=502 \\$aThesis (M.S.)--Oregon State University, 2008.
=520 \\$aOregon State University has recently conducted a complete core conversion analysis as part of the Reduced Enrichment for Research and Test Reactors Program. The goals of the thermal hydraulic analyses were to calculate natural circulation flow rates, coolant temperatures and fuel temperatures as a function of core power for both the Highly Enriched Uranium (HEU) and Low Enriched Uranium (LEU) cores; for steady state and pulsed operation, calculate peak values of fuel temperature, cladding temperature, surface heat flux as well as critical heat flux ratio (CHFR) and temperature profiles in hot channel for both the HEU and LEU cores; finally, perform accident analyses for the accident scenarios identified in the Oregon State TRIGA{reg} Reactor (OSTR) Safety Analysis Report (SAR). RELAP5-3D Version 2.4.2 was used for all computational modeling during the thermal hydraulics analysis. This is a lumped parameter code forcing engineering assumptions to be made during the analysis. A single hot channel model&#x2019;s results are compared to that produced from more refined two and eight channel models in order to identify variations in thermal hydraulic characteristics as a function of spatial refinement.
=530 \\$aAlso available on the World Wide Web.
=583 \\$xIssue Date: 2008-04-03T23:16:26Z
=504 \\$aIncludes bibliographical references (leaves - ).
=650 \0$aTRIGA reactors.
=650 \0$aTRIGA reactors $xSafety measures.
=650 \0$aNuclear reactors $xFluid dynamics.
=650 \0$aHeat flux.
=650 \0$aRELAP5-3D.
=690 \\$aTheses, OSU$xNuclear Engineering.
=856 41$uhttp://hdl.handle.net/1957/8272
```