
Repurposing MARC metadata: using digital project experience to develop a metadata management design

*Martin Kurth
David Ruddy and
Nathan Rupp*

The authors

Martin Kurth is Head of Metadata Services, David Ruddy is Head of Systems Development and Production, Electronic Publishing, and Nathan Rupp is Metadata Librarian, all at Cornell University Library, Cornell University, Ithaca, New York, USA.

Keywords

Online cataloguing, Libraries, Design, United States of America

Abstract

Metadata and information technology staff in libraries that are building digital collections typically extract and manipulate MARC metadata sets to provide access to digital content via non-MARC schemes. Metadata processing in these libraries involves defining the relationships between metadata schemes, moving metadata between schemes, and coordinating the intellectual activity and physical resources required to create and manipulate metadata. Actively managing the non-MARC metadata resources used to build digital collections is something most of these libraries have only begun to do. This article proposes strategies for managing MARC metadata repurposing efforts as the first step in a coordinated approach to library metadata management. Guided by lessons learned from Cornell University library mapping and transformation activities, the authors apply the literature of data resource management to library metadata management and propose a model for managing MARC metadata repurposing processes through the implementation of a metadata management design.

Electronic access

The Emerald Research Register for this journal is available at www.emeraldinsight.com/researchregister

The current issue and full text archive of this journal is available at www.emeraldinsight.com/0737-8831.htm

Technical services staff in libraries have a long history of optimizing and documenting the processes they use to create and manage the metadata contained in MARC-based library management systems. With the expansion of library metadata processing into non-MARC schemes, metadata managers and practitioners are faced with the need to extend the tradition of careful MARC metadata management to all library metadata processes in an environment that is complicated by decentralization.

The decentralized situation in which libraries find themselves parallels that in all automated work environments following the emergence of relational databases, desktop workstations, and client-server architectures. These tools made it possible for internal data users to extract data from centralized organizational databases and create new, unique, standalone database applications on their desktops (Tannenbaum, 2002). In libraries that are building digital collections, information technology and metadata staff access the relational database files underlying library management systems to extract and manipulate MARC metadata sets to provide access to digital content via non-MARC schemes. The metadata processing environment in libraries that use MARC and non-MARC metadata schemes magnifies the decentralization and complexity common to automated workplaces.

In such libraries, metadata processing involves defining the relationships between metadata schemes (mapping), moving metadata between schemes (transformation), and coordinating the intellectual activity and physical resources required to create and manipulate metadata (metadata management). Mapping and transformation work is divided among metadata staff and information technology staff. The tools and electronic files used in metadata processing are often scattered throughout the library, on servers and on the workstations of individual staff members. Actively managing the non-MARC metadata resources used to build digital collections is something most of these libraries have only begun to do. Although the history of using non-MARC metadata for digital collections is a relatively short one, library metadata staff and information technology staff have already begun to notice the problems that result from the lack of metadata management. Access to digital collections inevitably suffers when staff try to recall metadata content decisions, recreate metadata repurposing workflows, or troubleshoot connection failures by locating metadata resource

Received 16 September 2003

Revised 1 November 2003

Accepted 7 November 2003



files among the innumerable folders of staff members who have handled them at various junctures in the metadata processing pipeline.

For libraries seeking to manage their metadata processing operations as a whole, an excellent starting point is the management of MARC repurposing efforts. We see four reasons for attending to the use of MARC metadata in non-MARC applications as a first step in the larger endeavor of library-wide metadata management. First, the MARC records in a library management system are typically the largest store of metadata in the library. The Cornell University Library (CUL) library management system, for example, contains 4.5 million bibliographic records and is arguably irreplaceable as a tool for retrieving and managing Cornell's unique array of physical and electronic resources. Second, repurposing MARC metadata necessarily involves mapping and transformation. Because the emerging library metadata environment requires mapping and transformation activities, MARC metadata repurposing is a representative subset of the library metadata environment as a whole. Third, the metadata mapping schematics and transformation processes used in MARC repurposing are sufficiently costly and complex to warrant optimization and documentation. Finally, MARC metadata repurposing inevitably results in data redundancy, with duplicative occurrences of source and derived metadata residing in various library systems. Carefully managing redundant metadata can avoid unnecessarily multiplying maintenance efforts and creating access points that should be identical but are instead divergent.

For the reasons just identified, this article will propose strategies for managing MARC metadata repurposing operations as the first step in a coordinated approach to library metadata management. In considering MARC metadata repurposing and metadata management, we will draw on our experiences as CUL staff members from three different library units who have participated in MARC metadata repurposing activities to support CUL digital collection development projects. Although CUL is a large library with decentralized metadata operations, we believe that our MARC metadata repurposing experiences are relevant to metadata and information technology staff in libraries of all sizes and configurations who are engaged in similar activities. As we have already observed, the complexities of MARC repurposing derive as much from data redundancy and automated work environments as they do from specific library configurations. Thus we believe we can prudently generalize from our digital project experience to recommend approaches to MARC metadata

repurposing and thereby library metadata management.

Our approach to MARC metadata repurposing will concentrate on three areas: mapping activities, transformation processes, and metadata management design. We will begin by relating significant CUL experiences in mapping MARC metadata to other metadata schemes and will use these experiences to recommend a model for managing mapping activities. Using a similar approach, we will then describe CUL experiences with metadata transformation processes and draw on those experiences to make general observations about metadata transformations. Next, guided by lessons learned from CUL mapping and transformation activities, we will apply the literature of data resource management to library metadata management and propose a model for managing MARC metadata repurposing processes through the implementation of a metadata management design. Having explored mapping, transformation, and metadata design, we will conclude with two short sections that look to the future. We will offer practical next steps for practitioners who wish to apply metadata management design to their operations and we will identify areas for further research into issues related to library metadata operations.

Mapping activities

Although authors and practitioners often use "mapping" and "crosswalking" interchangeably, in this article we use "mapping" to refer to the process of establishing relationships between semantically equivalent elements in different metadata schemes and use "crosswalk" or "map" to refer to a visual representation of mapping relationships (St Pierre and LaPlant, 1998; Woodley, 2000).

In this section we describe significant CUL digital project experiences in mapping MARC metadata to the TEI Lite and Dublin Core metadata schemes. We give particular attention to CUL's MARC-to-Dublin Core (DC) mapping effort because it represents the library's first coordinated approach to metadata mapping. We conclude the section by drawing on CUL mapping experiences to recommend a model for managing mapping activities.

MARC-to-TEI mapping experiences

CUL has derived metadata from MARC records for many of its digital library projects. In 1995-1996, for one of Cornell's earliest and largest digital conversion projects, the Making of America (MOA) (Cornell University Library,

1999), CUL supplied its scanning vendor with MARC records, which the vendor used to generate volume-level descriptive metadata records to accompany digitized page images. MOA project staff relied on the MARC-derived metadata primarily for file management, but they also applied it in an early MOA online delivery system. In 1998, the library implemented a digital collection delivery system developed at the University of Michigan to provide enhanced online access to MOA. This system required, as an ingest format, SGML-encoded TEI Lite documents containing bibliographic metadata, structural metadata, and full-text content generated from optical character recognition (OCR) output. To populate the TEI Header fields with bibliographic metadata, staff used the MOA vendor's records, those originally derived from CUL MARC records.

A number of CUL projects completed since MOA have required similar TEI Lite encoding of metadata and OCR. These projects include the Core Historical Literature of Agriculture (Cornell University Library, 2004a), the Samuel J. May Anti-Slavery Collection (Cornell University Library, n.d. a), Historical Math Monographs (Cornell University Library, 2004b), and the Home Economics Archive (Cornell University Library, 2004c). All of these projects have used metadata derived from MARC records and mapped into TEI Lite documents. Though the mapping schemes used in these projects were similar, they varied from project to project.

Metadata staff developing the Home Economics Archive (HEARTH), a "core electronic collection of books and journals in Home Economics and related disciplines . . . [p]ublished between 1850 and 1950" (Albert R. Mann Library, 2003), based the HEARTH MARC-to-TEI mapping scheme on the scheme used in the MOA project, though they changed some element mappings. For example, the HEARTH mapping differed from the MOA mapping in its handling of edition statements. Because many of the print versions of the books and journals included in HEARTH had been published in the late nineteenth and early twentieth centuries, the MARC records for them presented a mixture of pre-AACR and AACR cataloging rules. When HEARTH staff adapted these catalog records to describe the digital versions of those publications, they did not update the cataloging to reflect current rules; instead, they retained the original description and added fields representing the electronic aspects of the objects cataloged. In order to map edition statements from MARC records that followed different cataloging rules, HEARTH metadata staff wrote mapping protocols to recognize that in some cases

the edition statement would come from the 250 field, while in other cases it would come from the 500 field. In the latter cases, the mapping rule identified (via such abbreviations as "ed.") which 500 field represented the edition statement.

A similar example of how collection-specific mapping protocols differ involves the Samuel J. May Anti-Slavery Collection and the Historical Math Monographs collection. For the May Anti-Slavery Collection, project staff mapped the date of publication for the pamphlets in the collection from the 07 through ten character positions of the MARC 008 fixed field, because the MARC records used for the May Collection cataloged the original source documents, in this case the pamphlets. For Historical Math Monographs, however, CUL catalogers had created MARC records for the digitized versions of the original print books. The 008/07-10 character positions for these records contained the dates of the digital versions, whereas the publication dates of the original books were in the 008/11-14 character positions, so Historical Math Monographs staff mapped the 008/11-14 to TEI for that project.

The mapping schemes for MOA, HEARTH, May, and Historical Math Monographs reflect collection-specific variations that are common in MARC repurposing. CUL's experiences with these collections suggest that, given changes in cataloging practice over time and the requirements of individual digital collections, a single MARC mapping for all collections sharing a common metadata scheme is not possible.

Beginning coordinated mapping at CUL

As the number of digital collections at CUL continued to grow, and with them the number of MARC mapping schemes, CUL metadata librarians felt the need to develop a generalized, but CUL-specific, mapping scheme that they could use as the basis for collection-specific mappings in a number of different projects. Project staff would be able to consult the generalized mapping scheme in light of their project needs and revise it as they saw fit. To that end, CUL metadata librarians in early 2002 convened a group of staff who had worked on various digital library projects (including MOA, HEARTH, and May), along with other technical and public service stakeholders, to develop a CUL scheme for mapping MARC to DC. The group's organizers felt that bringing a representative group to consensus regarding MARC-to-DC mapping would generate a mapping scheme acceptable throughout the library system.

CUL metadata librarians chose DC for this effort because it was well on its way to becoming a metadata *lingua franca* (it had been approved by

NISO the previous fall and would be approved by ISO within a year), and because CUL had begun a project to implement Endeavor Information Systems' ENCompass product (Cornell University Library, 2003). ENCompass enables end users to search different digital library collections simultaneously by mapping their metadata schemes to a common element set. For that common element set, CUL ENCompass developers had selected DC.

CUL's DC Mapping Group began by consulting the MARC-to-DC Crosswalk created by the Library of Congress (n.d.). The DC Mapping Group wanted the CUL MARC-to-DC mapping to follow the LC Crosswalk whenever possible because the LC mapping had become a *de facto* standard for mapping MARC to DC. Although the CUL group based its work on the LC Crosswalk, it wanted to go beyond the LC map in three general ways. First, the LC Crosswalk did not address all MARC fields and the CUL group wanted to record its decisions regarding a greater number of MARC fields and subfields. Second, the DC Mapping Group wanted to apply recommendations found in the DC-Library Application Profile (Dublin Core Metadata Initiative, 2002) to the CUL MARC-to-DC map. Finally, the group wanted to expand the "Notes" section of the LC Crosswalk to reflect CUL practices with regard to such decisions as the treatment of initial articles in titles.

After the DC Mapping Group completed its work, members of the group presented their recommendations to CUL staff at a forum of the CUL Metadata Working Group. The group made some revisions to its MARC-to-DC map based on this additional input, and made its final recommendations available on the Web (Cornell University Library, n.d. b).

CUL's ENCompass Development Project quickly put the work of the DC Mapping Group to use. ENCompass Project staff extended the CUL MARC-to-DC Crosswalk to an even greater degree of granularity with regard to MARC subfields to support access to MARC-derived DC records in the ENCompass-based "Find Articles/Find Databases/Find e-Journals" system (Cornell University Library, n.d. c). Metadata librarians working on the ENCompass Project consulted the CUL MARC-to-DC Crosswalk regularly throughout the project, and then documented the revised MARC-to-DC mapping scheme they implemented in ENCompass. Documenting the ENCompass MARC-to-DC map enabled CUL ENCompass Project staff to offer their MARC mapping scheme to staff at other libraries who were implementing ENCompass. The efforts by the DC Mapping Group and the ENCompass

Project team were CUL's first attempts to record a generalized MARC mapping scheme and the application-specific mapping derived from it.

A MARC mapping model

Our experiences with MARC-to-TEI and MARC-to-DC mapping at CUL have led us to view metadata mapping in general, and MARC mapping in particular, as a series of refinements from an international or national standard to a library standard and ultimately to an application-specific map. Starting from a *de facto* international standard such as the Library of Congress MARC-to-DC Crosswalk has given CUL metadata librarians some confidence that local CUL mappings from MARC to DC will be compatible with other libraries' mappings based on the LC map. Similarly, developing local agreement around a more detailed metadata map should ensure an even greater consistency among the digital collections offered by a single library. And, as our experience with CUL MARC repurposing projects has shown, each digital project has its own peculiarities that call for collection-specific decisions regarding metadata mapping.

Throughout our mapping experiences, we have observed the clear need to document mapping decisions. We have found it worthwhile to share with our colleagues not only the library's standard mapping scheme, but also the specific mapping schemes developed for particular projects. Making all schemes available gives metadata librarians throughout the organization an opportunity to examine the collection-specific maps available and determine whether they can adapt any of them to their projects; if none of the maps is useful, metadata staff can choose the generalized library scheme and adapt it for their purposes.

The process of coordinating metadata mapping across a decentralized organization like CUL has proven to be no different than coordinating other decisions and processes that affect the organization as a whole. CUL's MARC-to-DC mapping experiences have confirmed the benefits of building a library-wide consensus and then disseminating results so staff who work with metadata can benefit from them. CUL metadata librarians' success with approaching metadata mapping from a library-wide perspective rather than an isolated project perspective led them to apply similar strategies to the metadata transformation processes that enact mapping rules.

Transformation processes

There is little agreement in the metadata literature regarding where metadata mapping ends and

metadata transformation, also called metadata conversion, begins. We view metadata mapping as the process of establishing semantic relationships between equivalent elements in different schemes and metadata transformation as the design and implementation of scripts and other tools that move mapped metadata between schemes. Representing a somewhat different view, St Pierre and LaPlant (1998) argue that a complete metadata map should define both semantic equivalents and transformation specifications. The boundary between detailed mapping specifications on one hand and transformation rules on the other is admittedly fuzzy. We include detailed mapping decisions, written as natural language instructions in lists or tables, within the boundaries of mapping; conversely, we include transformation rules written in languages such as Perl or XSLT within the boundaries of transformation.

In this section we turn our attention to metadata transformation processes. We describe CUL experiences with scripted transformation processes and we propose an XML/XSLT approach to transformations. We then compare the scripted and XML/XSLT approaches. We conclude the section by recommending strategies for MARC metadata transformations.

MARC metadata transformation experiences

As CUL's experiences with mapping MARC metadata have evolved, so have the processes the library has used to extract metadata from MARC records and transform it to another format – the transformation processes that implement the intellectual mapping work described in the previous section. The evolution of CUL transformation processes has followed a path similar to that of CUL mapping practices, from project-specific conversion processes to an increasing emphasis on reusable tools, processes, and workflows.

For early CUL digital collection projects such as MOA, scanning vendors derived descriptive metadata from MARC records before online delivery systems were in place to deliver the digital content. When CUL staff implemented the current delivery system for MOA, they used the vendor-supplied metadata records to populate TEI Header fields, rather than returning to the original, or “live”, MARC records in the library management system. The problems with this approach became more apparent with time. For one, the vendor had derived the MOA descriptive metadata from MARC without a clear sense of the delivery system's requirements; as a result, neither the vendor nor CUL staff had optimized the metadata records for any particular system. But more importantly, the MOA processing stream

had divorced the derived metadata from its source and there were no processes in place for regenerating it. The library had thereby created multiple versions of its metadata, with the resulting requirement that it be maintained in two places – the library management system (LMS) and the MOA delivery system. This experience of metadata replication, and its attendant problems of version control and duplicated maintenance effort, led CUL to work out a different approach to MARC metadata transformation for the May Anti-Slavery Collection.

As mentioned earlier, the metadata requirements for the May project were similar to those of MOA. The delivery system required TEI Lite encoded files, which included descriptive metadata in the TEI Header. This requirement called for a conversion process that would extract portions of the MARC record to populate a TEI Lite file according to the MARC-to-TEI mapping that project staff had devised. Rather than follow the MOA process, May Collection developers now wanted to avoid divorcing the derived bibliographic metadata from its source in the MARC record. What May project staff wanted to achieve instead was a way of regenerating the derived metadata easily and at will. This would allow CUL to maintain the bibliographic metadata for May Collection pamphlets in one place, the LMS. If CUL technical services staff corrected or altered May Collection records in the LMS, May project staff could rerun the transformation script, thereby updating the metadata used by the delivery system in a reasonably automatic way. This would obviate the need for redundant data entry and maintenance.

Implementing this transformation process involved CUL information technology staff. A programmer analyst wrote software to identify the relevant records in the LMS and extract fields from these records as specified in the MARC-to-TEI map. The extracted data was then encoded in XML and stored so that it could be picked up and combined with administrative metadata, structural metadata, and OCR output to produce the TEI files required by the delivery system. Once the programmer analyst began the transformation process, its operation was automatic.

The transformation process used in the May project had some clear advantages. Because the files required by the delivery system could easily be regenerated, it was relatively simple to rerun the entire transformation process if the bibliographic metadata was updated in the LMS. This approach allowed the LMS to remain a central site for maintaining bibliographic data, thus relying on well-established library procedures and workflows. The transformation process was

also quite flexible in terms of mapping and output. Working with a programmer analyst, librarians could request that any MARC field be extracted and mapped to any metadata output format. The flexibility of this “scripted” transformation process made it possible to reuse the same tools in subsequent projects that required similar conversions. As noted in the discussion of MARC mapping, while metadata librarians can fruitfully pass MARC mappings from project to project, collection-specific modifications are inevitable. The flexibility of the scripted approach to metadata transformations allows digital project staff to accommodate these modifications and create collection-specific conversion processes.

Experience has shown, however, that this manner of achieving needed flexibility is expensive. Because information technology staff must build the intellectual mapping work done by librarians into the transformation process, librarians must work closely with programmer analysts to install and test each new mapping. Every modification requires an analyst’s time, which is expensive in direct staff time requirements and can cause time delays in implementing project-specific MARC conversions. Further, scripted transformations are designed to run as batch processes initiated and directed by programming staff. In practice, however, single record transformations are often preferable and more efficient than regenerating thousands of records merely to accommodate a few changes. Though the scripted approach to MARC metadata conversion offers flexibility with regard to mapping and output, it is technologically inflexible with regard to scale. In other words, digital projects requiring metadata transformations also need flexibility that considers staffing and workflow efficiencies.

An XML/XSLT transformation process

The need for processes that accommodate different staffing requirements and workflow scenarios has prompted further evolution in CUL’s MARC metadata extraction and transformation strategies. For example, a CUL programmer has created a Web-based tool that automatically selects and extracts MARC records from the LMS based on criteria collected via an online form. Librarians can enter criteria, indicate various output parameters, and save and edit jobs. The extraction jobs run overnight and generate email notifications when they finish. Similar improvements in MARC metadata transformations are possible, perhaps by building on the features of the existing extraction tool. We envision an XML/XSLT transformation process that requires information technology

programming support to create the transformation tool but then allows staff in other units to carry out required work with few if any ongoing programming needs.

One strength of this approach to MARC metadata transformation is that it can use widely available XML and XSLT tools, taking advantage of the flexible mapping technologies they offer. XSLT processing essentially converts one XML document into another XML document by using instructions contained in an XSLT stylesheet. XSLT processors are widely available, and we have found that basic XSLT stylesheet creation and modification are skills that technical services staff can learn quickly.

A second strength of an XML/XSLT process for transforming MARC metadata is that its developers can build on existing work done with the MARCXML standard for XML encoding of MARC metadata. The Library of Congress (2003) has made a tool freely available precisely for this MARC-to-MARCXML conversion process. CUL has already implemented a Web-based interface to this tool for single record conversions of MARC records to MARCXML or DC. A fully functional implementation of this tool would require securing programming staff time to add the capability of batch processing MARC records. Once built, the modified tool could serve any project, as the conversion from MARC to MARCXML is uniform across MARC records.

The XML/XSLT approach to transformation accommodates collection-specific variations by converting MARCXML to other metadata schemes. It is XSLT’s straightforward handling of these variations that makes it so attractive and convenient. The subtle variations of MARC metadata mapping required by different digital collections can be scripted into an XSLT stylesheet by the librarians who are closest to a given collection and who have done the intellectual work of map preparation. Using the single record transformation tool just described, metadata librarians working on the CUL ENCompass Project have done just this, using XSLT to implement a MARC-to-DC map that met the project’s specific needs.

An XML/XSLT approach to MARC metadata transformations offers staffing, scheduling, and workflow advantages. Once library programmers build XML/XSLT transformation tools and put them in place, metadata staff can perform ongoing conversion work without information technology support. Metadata staff can also modify existing XSLT stylesheets or insert new ones without the need for programming skills. Using metadata staff rather than information technology staff for collection-specific transformation operations will

lower project staffing costs. With regard to scheduling, an XML/XSLT approach lets metadata practitioners decide when they will implement transformation modifications. XML/XSLT enables them to translate their own mapping decisions into transformation processes immediately by reusing or refashioning XSLT stylesheets as needed. Finally, using XML/XSLT offers exceptional flexibility with regard to transformation workflows. The same approach can be used for batch or one-off, record-by-record, processing. It might also be used for on-the-fly conversions, where library-developed tools import external MARC records (e.g. via Z39.50) and then convert them for user viewing or additional processing.

Considering MARC metadata transformations generally

Reflecting on our experience with scripted and XML/XSLT metadata transformations has prompted us to make these general observations regarding MARC metadata transformation processes. MARC metadata transformation work will inevitably encounter variation in digital project conversion needs. Transformation processes must accommodate this variation as a functional requirement in their design. To transform MARC metadata in a technically and economically efficient manner, transformation processes should be broadly available to library staff rather than centralized in information technology units. Libraries engaged in MARC repurposing need decentralized transformation processes that digital project staff can easily modify through routine and standardized methods such as the alteration or addition of XSLT stylesheets.

Decentralization of transformation processes, however, points to a critical need for a richer and more complete inventory and documentation of transformation components. The library staff who build and maintain transformation tools, including software and XSLT stylesheets, need to document them in a way that promotes their reuse. Without this intellectual control over the mechanisms by which librarians transform MARC metadata, it will be difficult to create the standardization in procedures necessary to facilitate the sharing, reuse, and adaptation of transformation tools.

Metadata management design

CUL experiences with MARC repurposing have highlighted the need to build library-wide consensus regarding mapping decisions between specific schemes; to document collection-specific

maps in order to share and reuse them; and to build, maintain, and document transformation tools in order to facilitate their sharing and reuse. In considering such topics as consensus, sharing, and reuse, we have moved from the specifics of mapping and transformation to metadata management. Metadata management comprises two interrelated aspects. First, it involves coordinating the intellectual work that drives metadata creation and manipulation activities and, second, it involves managing the tools and electronic files that result from metadata creation and manipulation processes. In this section we advocate meeting the needs of internal library users insofar as they are necessary to meet the needs of library end users. We explore the ways in which studying data resource management literature can inform our experiences as metadata practitioners to enable us to recommend potentially fruitful approaches to the ways in which libraries manage MARC metadata repurposing. And, last, we propose creating a metadata resource inventory as an initial step toward establishing a management design for MARC metadata repurposing.

Identifying the needs of internal users

In discussing metadata mapping activities and transformation processes in this article, we have identified the benefits to library operations that accrue when digital project staff share and reuse metadata resources. As members of library digital project teams, however, we have observed that sharing and reuse of metadata resources does not always occur. Instead, digital project staff often devise metadata maps, write transformation scripts, and create intermediate metadata files to meet digital collection project goals and deadlines without capitalizing on the latent value of those maps, scripts, and files to subsequent digital projects or to the ongoing maintenance of the collections for which they were created. Metadata operations would grow more integrated and effective if library metadata managers were to develop metadata processes that promoted sharing and reuse in order to meet the operational needs of the library staff who create and manage digital collections. Because the sustained vitality and integration of digital collections depends on digital project staff sharing metadata resources with each other over time, treating digital project staff as internal users with specific information needs is necessary to serving the long-term information needs of end users.

The value of developing metadata processes that meet the needs of both end and internal users is evident in some metadata management guidelines currently in effect. The UK Intra-Governmental

Group on Geographic Information (IGGI) has created two documents – “Principles of good data management” and “Principles of good metadata management” – that recommend best practices for information managers who deal with geographic data files and the metadata files that describe them (IGGI, 2000, 2002). The “Principles of good data management” contend that actively managing data by making it easier to access and thereby reuse improves operational activities. A significant aspect of the “Principles of good metadata management” is the principles’ emphasis on the benefit of metadata management for information managers as well as end users. The IGGI metadata principles also hold that actively managing metadata processes reduces the risk that metadata generation and transformation processes will be lost as an organization’s structure, staff, and activities change over time.

Consulting documents such as the IGGI data and metadata management principles prompted us to consult the literature of the data resource management discipline. The publications of the Data Management Association in particular document principles dedicated to managing the internal data needs of organizations in order to meet the data needs of the clientele that those organizations serve. According to the Data Management Association, data resource management “facilitates the stewardship of data” as a key organizational asset, making it more valuable to an organization through “planning, communication, control, coordination, and management” (DAMA Chicago, 2002). The discipline of data resource management seeks to manage the entirety of an organization’s data (including metadata) as an organic whole, regardless of the physical formats in which the data lie. Although the data resource management literature is largely targeted to private enterprise organizations, its principles are relevant to the metadata management needs of libraries building digital collections because they attend to internal data needs as intrinsic to external data needs.

Bringing data resource management to library metadata management

The benefits of applying data resource management principles to library metadata operations fall into two categories, cultural benefits and operational benefits. Cultural benefits relate to the adoption of values within the library that are useful to furthering its service mission, while operational benefits relate to the integration of useful practices into the library’s operations.

As we have discussed in this article, a library’s MARC metadata is connected to the metadata derived from it by mapping schematics and

transformation processes, all of which are represented by electronic files and likely paper documents stored throughout the library, often by many staff in several library units. Because such scattering of data and the staff members responsible for it is often the norm in libraries managing digital collections, it is important to note that applying data resource management to library metadata would not depend on centrally directing the creation and manipulation of all of the library’s metadata. Rather, it would involve, first, documenting existing metadata relationships and processes and, second, using that documentation as a primary resource when seeking to coordinate those relationships and processes. The two-part effort to describe and shape an organization’s metadata is what Tannenbaum (2002), who writes about data and metadata architectures in private enterprise settings, calls creating a meta-meta design for an organization.

Although we may be reluctant to use Tannenbaum’s terminology and introduce yet another “meta” into the library literature, we are nevertheless willing to endorse her advocacy for creating what we choose to call a library’s metadata management design in order to document and manage its metadata content relationships, processing applications, maintenance protocols, storage instances, and display occurrences. As a library begins to enjoy the operational benefits that creating a metadata design would lend to metadata activities, it would also begin to realize the cultural benefit of formally acknowledging the value of managing its metadata. Creating a metadata design would enable metadata and information technology staff to represent visually the location of an application-specific metadata scheme within the array of schemes used in the library. By calling on project staff to locate application-specific schemes within the library’s metadata design, the presence of a design would also require staff to establish the relationships between the project scheme and related schemes used elsewhere in the library. Treating project needs in a broader context in this way would maximize the chances for integration among projects and minimize the chances of duplicated effort and irrecoverable departures from existing access methods.

Underlying the effort to create a metadata management design for a library is the acknowledgement that a library’s metadata mapping schematics and transformation processes are valuable resources that serve the information needs of end users and as such warrant careful management. The principles of resource management call for an organization to optimize an

existing resource because the organization cannot afford an unlimited supply of it; to share and leverage a resource in as many ways as possible to maximize value and minimize cost; to anticipate a resource's requirements and fulfill them proactively; and to manage a resource carefully to make sure the organization uses it prudently, efficiently, effectively, and securely (DAMA Chicago, 2002). These values in managing resources may seem intuitively obvious, but we feel it is important to articulate them explicitly in the context of metadata design because the first two principles provide a justification for the cost of metadata design and the second two provide a rationale for undertaking it.

Related to the introduction of a metadata management design to a library's metadata operations is the application of the data resource management notion of "enterprise" to a library's metadata processes. Treating heretofore project-based metadata operations as an integrated whole through a metadata design is brought into clearer focus by seeing that whole as an enterprise. Here we draw on *Webster's Third International Dictionary* to define an enterprise as a planned, systematic, often complex venture undertaken to achieve a specific purpose. Enterprises frequently have the economic features of risk and cost. We have already noted the tendency in libraries to "look past" the needs of internal users in order to meet the needs of end users. When we look at a library's metadata generation activities in light of our definition of an enterprise, we can begin to see them not simply as by-products of an effort to serve end users, but as parts of a coordinated venture with a singleness of purpose whose careful management is essential to meeting the needs of end users. Insofar as data resource management principles focus on the enterprise, they seek to find integrated solutions that benefit an organization as a whole. Because of their collection-oriented history, digital library metadata creation processes can benefit from the holistic approach that an enterprise perspective brings to metadata repurposing.

Another potential benefit for libraries in drawing on the data resource management literature lies in incorporating the principle of data stewardship to manage the components of metadata mapping and transformation. The cultural value of stewardship is one of accountability. It holds that if information producers create or update data that other people in the organization can use, they will generate it, store it, and make it accessible to meet their colleagues' needs as well as their own (DAMA Chicago, 2002). At an organizational level, data stewardship requires buy-in from information managers and producers throughout

the organization. Applying data stewardship in library metadata management can achieve such operational benefits as enhancing communication and productivity among metadata and information technology practitioners and making it more likely that they will find and use tools generated in other parts of the library. Data stewardship also facilitates such metadata management outcomes as treating metadata as a shared resource, reducing metadata development and maintenance costs, minimizing the creation of redundant components, making it possible to develop new metadata applications faster by sharing and adapting existing maps and transformation tools, and making it more likely that new application-specific mappings and transformation tools will integrate with the existing metadata development environment.

Toward a design for MARC metadata repurposing

As we have already observed, creating a metadata management design for a library is a two-phase effort. First, library staff document existing metadata relationships and processes and, second, they use that documentation as a primary resource for coordinating those relationships and processes. Because this article focuses on MARC metadata repurposing as a representative subset of a library's metadata processes, our discussion of the documentation phase will focus on MARC repurposing.

To begin documenting MARC repurposing processes, we propose creating an inventory of the data files, mapping schematics, transformation processes, and systems that comprise the components of the library's current metadata repurposing efforts. The components of the metadata processing workflow for the May Collection project listed below are generally representative of those used in repurposing MARC metadata for digital projects at CUL. We have followed each entry in the inventory with observations about significant features of the component:

- (1) *The MARC bibliographic metadata, both content and content designations, as stored in the library management system.* For the May Collection, the MARC bibliographic metadata was used as a source for the description of the digitized pamphlets in the May Collection. The use of MARC bibliographic records for authority-controlled element values is complicated by the fact that the authoritative source for the value is not the bibliographic record, but rather its associated authority record. Some digital repository projects use MARC holdings metadata in combination with MARC bibliographic metadata as source metadata.

- (2) *The extract script or tool that selects and extracts the MARC bibliographic metadata for the project.* This script was written and is maintained by CUL information technology staff.
- (3) *The file that is the product of the extract in (2).* The extract file for the May project resides temporarily on a server in the library's server farm, under the control of the programmer who ran the extract process. For more recent digital projects, similar extract files reside on the desktop workstations of library staff members.
- (4) *The collection-specific MARC mapping used for the project.* The mapping used for the May project indicates which MARC fields from the records contained in the file (3) are to be captured and encoded in the descriptive metadata section of the XML metadata collection and storage scheme (5). As mentioned, we recommend deriving the project-specific mapping from a generalized mapping agreed on by stakeholders throughout the library.
- (5) *The XML metadata collection and storage scheme.* This XML scheme was designed by library staff to collect and store together in a single file several types of metadata about a digital library object in a transitional stage prior to the creation of TEI Lite files. The scheme includes descriptive, structural, and administrative metadata. Bibliographic metadata from MARC records are mapped into the descriptive section of the scheme. XML elements within this section are based on TEI Header elements.
- (6) *The transformation script that creates an XML file – meeting the specifications of the metadata storage scheme (5) – and populates the descriptive metadata of this file with MARC metadata elements from the extract (3), following the MARC mapping specified for this project (4).* Cornell information technology staff created and maintained the transformation script used for the May project.
- (7) *The XML file that is the product of the transformation in (6).* For the May project, these files reside on a library server. It should be noted that additional processes (with their own transformation scripts and data sources) also place metadata into these files. Examples of such metadata include the structural information contained in an image-to-page correspondence table as well as information about page image file names. All metadata about a particular document is eventually collected in this XML file.
- (8) *The transformation script that generates a TEI Lite file by taking metadata from the XML metadata storage files in (7) and integrating it with page-level optical character recognition data.* This transformation script was written by information technology staff, but not the same staff who maintain the transformation script in (6).
- (9) *The TEI Lite file that is the product of the script in (8).* This file is stored on a library server.
- (10) *The DTD used to validate the files in (9).* The DTD used is an abridged subset of TEI Lite. A validation check is done for quality control by the script in (8). The DTD is also used by the digital collection delivery system to validate files during ingest and to aid internal data management processes. This DTD resides on a library server.
- (11) *The project metadata as stored in the digital collection delivery system after the TEI Lite XML file is ingested.* A digital collection delivery system may or may not store metadata as XML. Regardless of whether the system stores the metadata as XML, it may not be able to output the metadata, including any changes made subsequent to ingest, as an XML file that validates against the DTD or schema used for the project.

Moving from an inventory like the one above to comprehensive documentation and coordination of MARC metadata repurposing will require further work. Establishing a metadata management design for MARC repurposing processes calls for a library to manage those processes as resources and to apply principles of stewardship to them. CUL, for example, has not yet brought library stakeholders together for a formal discussion of the costs and benefits of such efforts. Stakeholder discussions would provide an opportunity to raise the issue of identifying stewards for key metadata repurposing components. To bring stewardship to a library's metadata processes, we anticipate establishing these values for each component identified as critical to the process: its authoritative version; its location (server, filename, identifier); the staff or unit responsible for it; its supporting documentation (may be internal or external to the component); its backups; and the standards and policies it follows.

Proposals for library investigation

Our inventory of the metadata components involved in building the May Anti-Slavery Collection points to the work that still needs to be done to coordinate CUL's MARC metadata repurposing processes. As the authors of this article we have begun the documentation phase of creating a MARC metadata repurposing design, but as metadata practitioners at CUL we have yet to extend this work into the coordination phase or to introduce these ideas to our colleagues to generate the buy-in necessary for their

implementation. We expect that metadata practitioners at other libraries who are interested in improving their management of MARC repurposing operations may find themselves in similar positions. As our recommendations for immediate action, we offer the following list of practical next steps for practitioners wishing to apply the metadata management approaches we have discussed to their MARC repurposing operations:

- (1) Build library-wide consensus regarding metadata element decisions and generalized mappings. For example, CUL has recently drawn on its experience in reaching consensus about MARC-to-DC mapping to develop local consensus around preservation metadata elements (Cornell University Library, 2002).
- (2) Develop reusable transformation tools. CUL's experiences with its MARC-to-XML conversion tool and its MARC extraction tool are promising. CUL has many more transformation scripts currently in use that are not as well documented or as easily accessible by staff library-wide.
- (3) Organize meetings with stakeholders to discuss the costs and benefits of creating a MARC metadata repurposing design for the library. Metadata practitioners will want to use these discussions to demonstrate to their colleagues the value of stewardship in making tools and resource files more broadly accessible. Stakeholders' meetings will present opportunities to begin articulating the roles and responsibilities of metadata staff and information technology staff in metadata management.
- (4) Extend discussions of a MARC repurposing design to discussions of creating a library-wide metadata management design. Stakeholders in library-wide metadata management design will likely be more numerous than those interested in MARC repurposing design because library metadata activities typically extend beyond MARC repurposing.
- (5) Investigate the costs and benefits of taking the creation of a library-wide metadata management design yet further by investigating the creation of a metadata management repository of mapping schematics, transformation tools, data files, and other metadata resources. Creating a metadata repository would involve treating metadata components as persistent digital objects with persistent identifiers and descriptive metadata in order to facilitate their discovery and retrieval through a digital content delivery system. Building searchable

metadata repositories would make it easier for libraries to share their metadata mapping and transformation resources with each other.

The BellSouth Metadata Services Group has created a metadata repository similar to the one described here. The BellSouth repository includes information about the "databases, data transformations, interfaces, systems, metrics, components, Web content, XML artifacts, messaging structures, Web services and documents, all of which are accessible from a Web portal and cross-referenced by the use of the Dublin Core (DC) standard" (Stephens *et al.*, 2003). The repository contains 150,000 objects and provides such services as "data mapping, documentation, metrics, reusable components, naming standards, etc." (Stephens *et al.*, 2003).

Recommendations for further study

To provide a conceptual framework for the continuing work in metadata mapping, transformation, and management design we expect in the years ahead, we have identified three potentially fruitful areas for further research. These are:

- (1) Investigate the use of such architectures as the Dublin Core Abstract Model (Powell, 2003) and the METS external descriptive metadata (mdRef) element for linking MARC and MARC-derived metadata records, thus providing an infrastructure for refreshing MARC-derived metadata. Creating a library-wide MARC metadata repurposing design would identify the relationships among MARC and MARC-derived metadata occurrences. Diagramming these relationships would provide a foundation for establishing automated processes to refresh element occurrences with updated values. Linking among metadata value occurrences is an emerging activity that bears watching from a metadata management perspective.
- (2) Investigate the programmatic use of models such as the Simple Bucket Digital Object Model (Chandler and Westbrook, 2002) and the Master Metadata File model (Davis, 1998; Mandel, 1998) for collocating metadata elements and values intended to be shared among distributed metadata records. Some models for managing related metadata occurrences involve collocating metadata elements from related records in a single digital object. For example, metadata in a Master Metadata File object would comprise elements and values from multiple schemes

- including MARC. Proponents of collocative models argue that they facilitate management over time, assimilation of metadata content from multiple schemes, and delivery of metadata content into multiple schemes. Building a single file of metadata objects containing metadata from multiple applications would promote consistent data definitions, maintain consistent content for key elements, and formalize relationships among related digital objects (Davis, 1998).
- (3) Investigate the implications that building a MARC metadata repurposing design hold for creating Archival Information Packages (AIPs) in accord with the Open Archival Information Systems (OAIS) Reference Model. As we have observed, managing digital collections over time depends heavily on metadata processes and the files they produce. For AIPs that contain MARC-derived descriptive metadata, it may serve the ends of long-term preservation to establish links between the MARC-derived metadata contained in AIPs and the inventory of maps, tools, and files that generated that metadata. Establishing those links may also make it possible to refresh MARC-derived metadata in AIPs dynamically.

Conclusion

In light of the experience of metadata and information technology staff at Cornell University Library in repurposing MARC metadata for digital collection projects, we see the need to manage internal metadata processing more efficiently in order to serve end users reliably over time. Recent metadata mapping and transformation efforts indicate a trend in library metadata operations: Libraries will receive, create, and transform metadata in multiple schemes amid an increasingly complex and decentralized technological environment. The literature of data resource management indicates that libraries are not alone in facing this phenomenon; indeed, data resource management can provide useful strategies to library staff as they explore more systematic and comprehensive approaches to managing metadata operations.

This article advocates that libraries develop metadata management designs. Such designs will recognize the importance of mapping schematics and transformation processes to library operations and will propose management practices that optimize their use. In particular, we have argued for the need to build library-wide consensus on metadata mapping decisions and to document

mappings and transformation processes systematically in order to promote sharing and reuse. We see the value of approaching library metadata work as a coordinated enterprise (especially in a decentralized environment) and of reinforcing metadata coordination by establishing data stewardship responsibilities. Although library metadata and information technology staff have yet to discuss the many specific components of potential metadata management designs, the general features of a library metadata management design as outlined here offer metadata practitioners a useful framework within which to position their operational needs and potential solutions. Metadata management designs promise effective approaches for libraries seeking to optimize their significant investments in metadata work.

References

- Albert R. Mann Library (2003), "Home economics archive: research, tradition and history (HEARTH)", available at hearth.library.cornell.edu
- Chandler, A. and Westbrook, E.L. (2002), "Distributing non-MARC metadata: the CUGIR metadata sharing project", *Library Collections, Acquisitions, & Technical Services*, Vol. 26 No. 3, pp. 207-17.
- Cornell University Library (1999), "Making of America", available at: <http://cdl.library.cornell.edu/moa/>
- Cornell University Library (2002), "CUL working meeting on preservation metadata", available at: www.library.cornell.edu/iris/dpo/metadata.html
- Cornell University Library (2003), "CUL ENCompass Development Project", available at: <http://encompass.library.cornell.edu/>
- Cornell University Library (2004a), "Core historical literature of agriculture", available at: <http://chla.library.cornell.edu/>
- Cornell University Library (2004b), "Historical math monographs collection", available at: <http://historical.library.cornell.edu/math/>
- Cornell University Library (2004c), "Home economics archive", available at: <http://hearth.library.cornell.edu>
- Cornell University Library (n.d. a), *Samuel J. May Anti-Slavery Collection*, available at: www.library.cornell.edu/mayantislavery/
- Cornell University Library (n.d. b), "CUL MARC to Dublin Core Crosswalk", available at: http://metadata-wg.mannlib.cornell.edu/programs/docs/CUL_MARC_to_DC_Crosswalk.htm
- Cornell University Library (n.d. c), "Find articles/find databases/find e-Journals", available at: <http://find.library.cornell.edu>
- DAMA Chicago (2002), *Guidelines to Implementing Data Resource Management*, 4th ed., DAMA International, Bellevue, WA.
- Davis, S.P. (1998), "Managing and accessing the digital library", available at: www.columbia.edu/cu/libraries/inside/projects/metadata/presentation/nypl/nypl.ppt
- Dublin Core Metadata Initiative (2002), "DC-library application profile", available at: <http://dublincore.org/documents/2002/09/24/library-application-profile/>

- Intra-Governmental Group on Geographic Information (IGGI) (2000), "The principles of good data management", available at: www.iggi.gov.uk/achievements_deliverables/manage.pdf
- Intra-Governmental Group on Geographic Information (IGGI) (2002), "The principles of good metadata management", available at: www.iggi.gov.uk/achievements_deliverables/pdf/Guide.pdf
- Library of Congress (n.d.), "MARC to Dublin Core Crosswalk", available at: www.loc.gov/marc/marc2dc.html
- Library of Congress (2003), "MARCXML Site", available at: www.loc.gov/standards/marcxml
- Mandel, C. (1998), "Manifestations of cataloging in the era of metadata", available at: www.columbia.edu/cu/libraries/inside/projects/metadata/presentation/alctslita/alctslita.ppt
- Powell, A. (2003), "Dublin Core abstract model", available at dublincore.org/documents/2003/08/10/abstract-model
- St Pierre, M. and LaPlant, W.P. Jr (1998), "Issues in crosswalking content metadata standards", available at: www.niso.org/press/whitepapers/crswalk.html
- Stephens, R.T., Wilson, A. and Jenkins, B. (2003), "Best practices in metadata management", available at: www.wilshireconferences.com/award/Submissions/Bellsouth.pdf
- Tannenbaum, A. (2002), *Metadata Solutions: Using Metamodels, Repositories, XML, and Enterprise Portals to Generate Information on Demand*, Addison-Wesley, Boston, MA.
- Woodley, M. (2000), "Crosswalks: the path to universal access?", in Baca, M. (Ed.), *Introduction to Metadata: Pathways to Digital Information*, Getty Research Institute, Los Angeles, CA, available at: www.getty.edu/research/conducting_research/standards/intrometadata/2_articles/woodley/index.html