July 2011

Discovery Tools and Local Metadata Requirements in Academic Libraries

Mary S. Wood
San Jose State University, wood_mary@yahoo.com

Follow this and additional works at: https://scholarworks.sjsu.edu/ischoolsrj

Part of the Library and Information Science Commons

Recommended Citation

This article is brought to you by the open access Journals at SJSU ScholarWorks. It has been accepted for inclusion in School of Information Student Research Journal by an authorized administrator of SJSU ScholarWorks. For more information, please contact scholarworks@sjsu.edu.
Discovery Tools and Local Metadata Requirements in Academic Libraries

Librarians in the academy who work to optimize collection access face an onerous task given the complex nature of information and the information world. This complexity is reflected on academic library websites, which typically feature an online public access catalog (OPAC) as well as lists of database links, all of which function independently of one another. These distinct access choices require consolidation to facilitate efficient and effective searching for users who tend to engage in “Google-like” search behaviors. Discovery tools have been developed to address this problem, but their effectiveness depends on the quality of the metadata created to represent collection materials. This paper examines the current research on metadata creation practices employed in academic libraries to represent unique local collections—specifically institutional repository (IR) and special collections (SC) materials—in order to identify metadata quality issues that impede access to these resources in a discovery environment.

Problem Statement

To meet the scholarly research needs of students and faculty, academic librarians have traditionally been responsible not only for selecting resources but also for facilitating resource access. With the advent of electronic resources and the increasing demand for these resources among students and faculty, academic librarians must manage a widening and evolving array of electronic content. Not a week passes in an academic library without major electronic resource additions or revisions—new titles, changed titles, cancelled titles, platform upgrades and feature additions, or migrations to new platforms and vendors. Growth is not restricted simply to more online versions of print equivalents—more bibliographic databases, e-journals and e-books. “Born digital” genres are on the rise—numeric data resources, image galleries, multimedia reference works, and interactive tools. Over a remarkably short period of time, the range and complexity of commercial and open access electronic resources has expanded—from bibliographic indexes to full text, from electronic journals to electronic books, from text-based interfaces to GUI (graphical user interface), from plain text to digitized facsimile to “born digital” and multimedia. (Kichuk, 2010, p. 55) Given the complex nature of electronic resources, the academic librarian’s job of providing access can be quite challenging. Complicating this job further, publishers retain substantial control over electronic resource content and delivery (Bergstom, 2010; Evans & Saponaro, 2005; Torbert, 2008), and the subscription databases required for retrieving these materials are less controlled and consistent than OPACs (Collard & Whatley, 2011) and do not always function optimally (Beall, 2011; Gilbert, 2010). Consequently, an academic library’s electronic
collection can become quite extensive and complicated to search, yielding results that users may not sufficiently comprehend (Wolverton & Burke, 2009).

Federated search tools have shown promise for streamlining information retrieval by enabling users to query multiple databases simultaneously, albeit separately. Unfortunately, these tools have demonstrated inherent problems related to visibility and use, speed, and metadata inconsistencies (Collard & Whatley, 2011; Wolverton & Burke, 2009). Academic libraries are beginning to adopt new discovery tools to more effectively and efficiently streamline the search process in the hopes of facilitating better search results.

Discovery tools...allow the user—through a single search box—to search a base index of metadata as well as many of the library’s digital resources such as proprietary databases, the catalog, and institutional repositories. Mimicking the Google experience, results from both internal and external sources can be served up in a single relevancy-ranked batch. (Kenney, 2011, p. 25)

A notable advantage of discovery tools is that they can provide access not only to proprietary electronic content but to local collections as well. In their quests for information, users are more likely to use search engines than local metadata tools (Calhoun, as cited in Anderson, 2008), which are essential for finding unique materials held locally. Anderson (2008) suggests that providing access to these materials should become the focus of libraries, particularly academic libraries, if they are to remain viable in a changing information world. As Adamich (2010) aptly concludes,

the online catalog still appears to be an important tool for locating library materials and accessing information. However its scope and role may be modified to become more of a marriage with other similar tools used to manage library materials and information access. (p. 4)

Discovery tools provide an opportunity for such a marriage and promise to make unique local collections more accessible, but they also put these collections at greater risk for being overlooked if the corresponding metadata—which must be created in-house—is not sufficiently developed to translate into appropriate relevancy rankings for the materials it represents. To investigate this matter, the following literature review looks specifically at metadata creation practices that are associated with IR and SC materials and focuses on metadata quality issues that might hinder access to these local collections of unique resources in a discovery environment.

**Literature Review**

Before adopting a discovery tool, an academic library would be well advised to consider metadata quality issues that might impede access to the unique local materials it collects. This review of the recent academic library literature reveals
some of these issues. It focuses on IR and SC materials—the two primary types of unique materials that comprise local collections and thus may not be adequately represented by their associated metadata—and the processes through which metadata is created to represent these resources for retrieval.

**Institutional Repository (IR) Materials**

According to Chapman, Reynolds, and Shreeves (2009), an IR “collects, manages, and disseminates materials produced by an institution” (p. 310). In a comparison of IRs at three major universities in the United States—the University of Minnesota, Johns Hopkins University, and the University of Illinois at Urbana-Champaign—these researchers investigated metadata creation processes and examined variations in metadata creation strategies across institutions. The IRs under investigation used the same DSpace software package and represented mixed metadata environments in which multiple metadata creation workflows were performed by creators (faculty and other researchers), managers (repository staff), and catalogers (library staff). A number of metadata creation issues were considered: inconsistency in metadata within each repository, lack of authority control, and complex controlled vocabularies. These researchers found that the blanket DSpace metadata format (Dublin Core) did not provide the flexibility required for describing materials given the diversity of disciplines and content formats represented, and they noted that IRs should incorporate a variety of metadata schemes and controlled vocabularies to provide more granular description. They also pointed out the need for more comprehensive metadata capabilities to capture administrative, structural, and standardized author identifier elements. Differences in metadata creation strategies across the IRs under investigation reflected local efforts to work around these issues, and the authors concluded that “metadata tool investment would help to minimize the amount of customization each institution has to do in order to produce metadata that meets their requirements” (Chapman et al., 2009, p. 324). This suggests that a mixed metadata environment requires a variety of metadata creation tools to capture the mix and thereby enhance discovery.

In another study concerned with IR metadata quality that looked at multiple metadata creation tools, Birrell, Dunsire, and Menzies (2010) investigated the interoperability, duplication, and authority control of OPACs and IRs by surveying 85 academic libraries across the United Kingdom. They found little evidence of interoperability between the metadata systems at each institution despite considerable overlap in content scope, reflecting duplication resulting from differences in the types of metadata contained in each metadata system (i.e., the OPAC and one or multiple IRs). Like Chapman et al. (2009), Birrell et al. (2010) suggested improving institutional research metadata through the creation
and maintenance of many different metadata tools:

Duplication and scope overlap in fact serve a useful purpose in some instances, especially in an institution with multiple repositories. Different sites may be used by different types of user/community [sic], some feeling more at home (whether as depositors or users) when working within a departmental IR rather than a “centralized” one. Department IRs may be tailored around the needs of a specific user community (e.g., in terms of formats supported, subject granularity reflected, or tools built into the user interface). Some may prefer electronic journal catalogs where they can look at what other content appears in a given issue, rather than being limited to individual records; some researchers may prefer the interfaces of a “classic catalog,” and others those of a “Web 2.0” style Resource Discovery Platform. (p. 398)

They also made the critical point that in order for this type of multi-use system to work, institution-wide processes must be developed so that each metadata instance is based on an original source to which it is then linked, thereby establishing a network of links to all other related metadata instances (Birrell et al., 2010). Such collaborative processes should facilitate the development of the rich metadata content necessary for increasing relevance in a discovery tool search by providing a mechanism for identifying and filling metadata gaps.

Boock and Kunda (2009) provide an example of collaborative processing in their assessment of metadata workflows for print and electronic theses and dissertations at the Oregon State University Libraries. Workflow collaboration was driven by the need to streamline processes due to library staff reductions, resulting in a process in which IR metadata is generated first in Dublin Core using DSpace then mapped to Machine-readable cataloging (MARC) using a modified version of MarcEdit to generate OPAC metadata, a duplication of efforts that a discovery tool might render unnecessary unless the OPAC metadata is then enriched. However, they asserted that full subject analysis for these materials may no longer be warranted and could be discontinued to yield greater workflow efficiencies. They also claimed that the use of Library of Congress Subject Headings (LCSH) is impractical due to time and cost considerations and that the utility of LCSH to researchers is questionable, but they conceded that “a more thorough analysis is necessary before making a decision to discontinue the long-standing library practice” (Boock & Kunda, 2009, p. 303). However, as McCutcheon (2009) concluded in her comparison of keyword versus controlled vocabulary searching, “the one with the most tools wins” (p. 62), reinforcing the notion that metadata enrichment is essential to enhanced discovery.

Lubas (2009) looked at the issue of metadata enrichment in her research on thesis and dissertation metadata, which focused on electronic materials in an
attempt to identify best practices for metadata creation given a new electronic-only collection policy at the University of New Mexico. Her article addressed concerns about the duplication of metadata in DSpace and the OPAC, the enrichment of author submitted metadata, and the standardization of metadata creation. She pointed out that DSpace enables full-text searching and provides the ability to create descriptive metadata using Dublin Core; however, like Chapman et al. (2009), she lamented the lack of metadata elements in Dublin Core that are critical for thesis and dissertation discovery. Her literature review indicated that institutions have dealt with their IR systems’ limited thesis and dissertation metadata capabilities either by including supplemental metadata in linked MARC OPAC records or by creating descriptive metadata records in MARC that are housed in the OPAC but linked to full-text records housed in the IR system. These solutions required the development of standardized metadata creation processes and methods for correcting inadequacies and inconsistencies in the author-supplied metadata on which these processes have relied. According to Lubas (2009):

While during the early days the use of a simplified metadata element set such as Dublin Core may have seemed limiting, over the course of a decade of experience with electronic theses and dissertations metadata reveals [sic] that blending the use of qualified Dublin Core with harvesting and crosswalks, plus creating tools to encourage better results from author-generated metadata have proved useful. (p. 257)

However, the University of New Mexico had not yet progressed to this point at the time of her study:

Prior to this study [and the change to an electronic-only thesis and dissertation submission requirement] in spring 2009, the author-submitted Dublin Core metadata was not reviewed in detail or enhanced by a cataloger or metadata specialist. There was no connection between the metadata for the electronic version and paper version; no link for the electronic version was added to the MARC metadata for the paper version. (p. 254)

Based on her assessment of the thesis and dissertation data submitted under the electronic-only requirement, Lubas (2009) recommended a hybrid approach that commits to Dublin Core-based IR records, which can then be harvested directly, and then crosswalks those records to MARC for local discovery via the OPAC and remote discovery via the Online Computer Library Center (OCLC). While noting that discovery tools and OCLC can also harvest Dublin Core, she made the point that converting to MARC enables institutions to enhance records to fuller levels for more effective local and remote searching. She indicated that this should be accomplished through incremental metadata creation, starting first with the development of IR metadata for immediate full-text searching followed
by metadata enhancement through more time-intensive MARC record development to enhance discovery. McCutcheon, Kreyche, Maurer, and Nickerson (2008) also looked at an incremental approach to enhancing the discovery of electronic theses and dissertations in their study examining the development of Kent State University’s metadata creation process. These researchers described efforts to “promote and devise electronic thesis and dissertation (ETD) storage at OhioLINK’s ETD Center, to find efficient methods to represent these unique scholarly materials within the library’s catalog, and to foster the establishment of state-wide library catalog standards for ETDs” (McCutcheon et al., 2008, p. 41). The university uses a consortium’s IR rather than an institutional IR to provide its scholarly material with broad, immediate visibility using author-supplied metadata; the IR metadata is then extracted, modified, enhanced, and inserted into the OPAC through a semiautomatic process that was developed to create provisional records for immediate catalog access using a Perl program and the Open Archives Initiative for Metadata Harvesting Protocol (OAI-PMH); the resulting records are then enhanced using ETD-specific cataloging standards developed specifically for this purpose; and the completed MARC records are then shared with OCLC for further discovery enhancement (McCutcheon et al., 2008). This comprehensive, enriching metadata creation process seems ideal for optimizing ETD access via a discovery tool, but the authors noted that it requires great flexibility and collaborative cooperation on the part of catalogers and systems professionals.

Special Collections (SC) Materials

Like IR materials, SC materials are unique to their institutions, rely on locally created metadata for access, and as such are at risk for being overlooked in a discovery environment if metadata quality standards fall short. Han, Cho, Cole, and Jackson (2009) defined SC materials as “materials that need special care and arrangement, or collections of materials that have been assembled for specific themes” (p. 214). These researchers looked at CONTENTdm metadata created for the digital surrogates of different types of SC materials to support local access at the University of Illinois at Urbana-Champaign. Their objective was to investigate quality issues associated with mapping this metadata to Dublin Core using OAI-PMH to enable sharing with external metadata aggregators. They noted that although it is based on Dublin Core, CONTENTdm allows for more detailed description and is widely employed in academic libraries to create local metadata that can be tailored to capture unique attributes and contextual information pertaining to different SC materials; however, external aggregators require standardized metadata records, so sharing requires that CONTENTdm metadata be mapped to Dublin Core, which is the minimum standard for use with OAI-
PMH. Han et al., (2009) found that this process diminishes metadata quality due to inadequate mapping capabilities, and they identified six considerations for creating CONTENTdm metadata to facilitate interoperability with Dublin Core:

1. Balancing specificity and generality in defining unique local fields.
2. Deciding in advance which unique fields should be shared and which should remain local.
3. Being cognizant of how values will be created in the local environment and how they will translate to Dublin Core.
4. Maximizing the use of Dublin Core elements for labeling.
5. Using field names and definitions from other metadata standards that have crosswalks to Dublin Core.
6. Sharing the logic of mapping decisions with aggregators.

Since discovery tools aggregate local metadata as well as proprietary metadata, these findings bear consideration for ensuring that unique SC metadata is sufficiently developed for optimal harvesting and highlights the need for developing interoperable SC metadata at the local level.

In an article that illustrates how another university handled the metadata interoperability issues associated with SC materials, Hurford and Runyon (2011) described the management a born-digital collection of 30,000 orchid photographs that was donated to the Ball State University Digital Media Repository. According to the authors, managing a born-digital collection challenged traditional archival processing methods and description standards and required more than the usual amount of institutional collaboration to arrange, describe, edit, and make available to researchers; however, they were able to implement a metadata creation process that appears to address the issues identified by Han et al. (2009). Local metadata was created so that it would not only be accepted and displayed in a standardized way in the repository using CONTENTdm, but it could also mapped successfully to Dublin Core. This strategy addressed interoperability as part of the metadata creation process and would likely facilitate access via a local discovery tool. Unfortunately, this project has been labor-intensive and slow, resulting in a backlog of materials that cannot be accessed at all (Hurford & Runyon, 2011).

Nelson (2010) also described a labor-intensive metadata creation process in his paper focusing on the cello music collection at the University of North Carolina at Greensboro. This process has been so labor intensive that it has required the expertise of a specialist cataloger; the Cello Music Cataloger is a tenure-track librarian who performs both the technical services and the public services functions associated with managing this collection. The position reports to cataloging but liaises with the SC department and the music library, suggesting
considerable collaboration. Most of work to date has focused on high-quality, research-intensive, original cataloging to provide access to manuscript materials that are “uncommonly rich in the raw materials of original scholarship” (Nelson, 2010, p. 639). Collaboration has been reserved for the public services component of the position. This model has created organizational problems, not just because it is labor intensive, but also because it does not incorporate other media types, or leverage other metadata creation processes that have been developed and collaboratively implemented to support research more broadly throughout the university (Nelson, 2010). While traditional cataloging might provide the rich metadata required to increase the relevancy of selected collection items for local access using a discovery tool, other materials that are required to provide collection balance must be adequately represented as well, rather than overlooked altogether in the metadata creation process.

One of the biggest conflicts in the processing of SC materials has resulted from a “desire to provide superior physical care and descriptive strength for...collections and the resultant proliferation of...unprocessed backlog[s]” (Cox, 2010, p. 135). Greene and Meissner (2005) introduced the notion of more product, less process (MPLP) to resolve this conflict, arguing that the goal of processing should be to maximize user access to collections by eliminating tasks that are not productive and adopting minimal processing standards. To assess the impact of MPLP, Crowe and Spilman (2010) conducted a survey of American archivists and concluded that MPLP has been widely accepted and has improved processing backlogs, researcher access, and reference service outcomes. However, they noted that research on the effects of MPLP on descriptive practice is lacking. This is a critical issue for discovery given “Greene and Meissner’s recommendation that description of a collection should match the level of arrangement” (Crowe & Spilman, 2010, p. 122). Eisloeffel (2010) pointed out that while Greene and Meissner’s 2005 work “focused on larger twentieth century collections of records as their baseline, admittedly focusing on ‘the paper issues’” (p. 20), it also recognized that some SC materials are more retrieval intensive and therefore require item-level treatment. In a more recent work, Meissner and Greene (2010) pointed out that MPLP, in fact, is not a processing dictate but rather a guide intended to help practitioners balance resources, in order to accomplish their goals within their own institutional contexts, while achieving economies along the way. Toward the goal of optimizing discovery, making SC materials visible by eliminating backlogs must be balanced with adequate description.

This balance has been illustrated by Cox (2010) in his model of maximal processing, developed at the University of Massachusetts. The model’s first step is to provide comprehensive online access to holdings, whether they have been
processed or not. This is accomplished through a systematic review of each collection, the creation of standardized collection-level metadata, and the assignment of future processing priorities—the equivalent of MPLP. When implemented, this step resulted in “a relatively dramatic rise in use, with several smaller and previously obscure collections finding an audience along with the old, much used chestnuts” (Cox, 2010, p. 144) as well as a more expeditious workflow for processing new collections. The model’s second step is to queue up collections for full processing:

Arrangement and description are the keys to discovery, usability, and serviceability. Accessibility is not a binary, not a yes or no, but rather a continuum that extends from no description to full text availability, and as maximal processors, our goal is to push our collections as far along this spectrum as possible, privileging the benefits that accrue to our researchers over the limitations of our resources. (Cox, 2010, p. 145)

Finally, the model’s third step is post-description, which recognizes the dynamic nature of collections and allows for additions, corrections, updates, and reprocessing in order to meet changing collection needs over time.

In an article illustrating how the description of unique, museum-type objects can be enhanced using methods similar to those described by Cox (2010), Baca and O’Keefe (2009) reflected on their efforts to integrate collaboratively created metadata within a traditional MARC framework. These researchers provided an overview of the struggle librarians face in their attempts “to create the kind of immediate access and instant gratification that Google seems to offer, [when] the area of metadata standards is experiencing a period of profound evolution” (Baca & O’Keefe, 2009, p. 59). They noted that metadata realities require librarians to judiciously and carefully employ a combination of standards to ensure that metadata creation yields online access tools that are viable and effective:

Cataloging of unique, museum-type materials—whether in a library production system or elsewhere—require[s] different approaches, different standards, different skill sets and subject expertise. For most of these materials, the “item in hand” is not the source of core information, as it is for published materials. (Baca & O’Keefe, 2009, p. 60)

These researchers focused on two trends for facilitating metadata creation for these resources, so that “diverse data content standards and vocabulary tools can be integrated within the classic data structure/technical interchange format of MARC21 to better describe unique, museum-type objects, and to provide better end-user access and understanding” (Baca & O’Keefe, 2009, p. 59). The first trend is the use of schema-agnostic metadata like RDA, which is generally associated with MARC because of its origins in the cataloging world but can also be used effectively with other metadata schemes, like MODS and Dublin Core.

The second trend is incremental metadata creation, which can be performed...
collaboratively by trained staff from the variety of departments that participate in
digitization projects. Baca and O’Keefe (2009) claimed that this type of metadata
creation workflow can add to the intellectual value required for high-quality
description while saving time and costs, assuming that the technical infrastructure
required for activities like expert social tagging, as well as the organizational
support required for the assumption of new roles, can be put in place. Based on
these trends, they suggested employing a metadata creation process in which
static resources are provided with dynamic records that develop over time to
include a wide range of elements—capturing information about everything from
the original work to groups, collections, and items to surrogates to related
works—“in order to satisfy user expectations for one-stop information shopping”
(Baca and O’Keefe, 2009, p. 67).

Implications for Practice and Research

Baca and O’Keefe’s (2009) suggestion that dynamic records should be created
for static resources provides the necessary framework for developing the rich
metadata required for effectively accessing unique local materials using a
discovery tool. However, they noted that this approach requires academic
librarians to go beyond traditional cataloging practices that generate static records
for static resources (e.g., books) and dynamic records for dynamic resources (e.g.,
serials and websites). Academic librarians must also consider the research on
metadata creation for IR and SC materials, which highlights a number of quality
issues that should be assessed and resolved to ensure that a discovery tool, once
implemented, will yield optimal results. The research indicates that a variety of
metadata creation tools are needed to capture the range of elements associated
with different material types and that interoperability must be built into the overall
metadata creation model to ensure optimal access to local resources via a single
search interface. Furthermore, the research shows that standardization in the
creation of metadata is essential for maintaining metadata consistency and that
collaborative, incremental metadata creation processes are imperative for
balancing immediate local resource access with the enhancement of local resource
discovery.

Future research should address issues related to the quality of local metadata
created for a wider range of purposes, such as user and expert tagging for
improved keyword searching. As McCullough (2010) noted, discovery tools have
been developed to facilitate both types of tagging, so the effectiveness of these
and other broad local metadata creation processes should be evaluated. In
addition, future research should assess the impact of different metadata creation
processes—as well as the impact of changing these processes based on best
practices suggested by the literature—on search effectiveness using discovery
tools to locate local materials. To date, discovery tool outcome studies have been limited and have not addressed local retrieval issues, although this is likely to change once discovery tools become more widely used in academic libraries. Research examining and comparing different discovery tools and how they operate is more common. “The discovery approach is still in its infancy” (Notess, 2011, p. 47) and therefore warrants ongoing examination and development.

Conclusion

The tide has shifted in academic libraries with the advent of discovery tools and their promise of a more “Google-like” search experience, moving information retrieval away from complicated, librarian-centered methods and toward a more streamlined process that accommodates today’s user preferences. Discovery tools obviate the need to represent the vast array of proprietary electronic resources in the OPAC as well as in multiple stand-alone databases accessible via academic library websites, which is a boon to academic researchers who expect user-friendly search options, as well as to academic librarians who are responsible for ensuring that electronic resources are visible and accessible. However, these tools also create an imperative for enriching the metadata that represents unique local resources so these resources can withstand the relevancy competition that discovery tools impose during the information retrieval process. It is unfortunate that in the wake of more product, less process (Greene & Meissner, 2005), some practitioners have traded backlogs of local materials for inadequate processing while others have continued with exacting processing standards and have consequently missed opportunities to provide new materials in new formats to their users. “Ensuring users are always matched with the right resources for their need—and that they don’t get lost in the vastness of information available online—is a critical component of the [l]ibrary’s role” (McCullough, 2010, p. 10), one that requires institution-wide collaboration both to optimize and to assess user outcomes. The adoption of discovery tools should reduce the time required to manage proprietary electronic resources to such an extent that academic librarians will be able to focus on local collections needs, attendant metadata issues, and essential collaborations. These collaborations should extend to users, as well as to colleagues, and should focus on enriching metadata over time using multiple interoperable tools, developing research-based metadata creation standards, assessing metadata creation practices, and evaluating discovery tool outcomes. The implications of discovery tool adoption demand this shift in priorities to improve practice and advance knowledge in the field.
References
Birrell, D., Dunsire, G., & Menzies, K. (2010). Match point: Duplication and the scholarly record: The online catalog and repository interoperability study (OCRIS), and its findings on duplication and authority control in OPACs and IRs. *Cataloging & Classification Quarterly, 48*(5), 377-402. doi:10.1080/01639371003738723


doi:10.1080/03615260801970857

doi:10.1080/03615260902877019