Metadata Creation Practices in Digital Repositories and Collections: Schemata, Selection Criteria, and Interoperability

Jung-ran Park and Yuji Tosaka

This study explores the current state of metadata-creation practices across digital repositories and collections by using data collected from a nationwide survey of mostly cataloging and metadata professionals. Results show that MARC, AACR2, and LCSH are the most widely used metadata schema, content standard, and subject-controlled vocabulary, respectively. Dublin Core (DC) is the second most widely used metadata schema, followed by EAD, MODS, VRA, and TEI. Qualified DC's wider use vis-à-vis Unqualified DC (40.6 percent versus 25.4 percent) is noteworthy. The leading criteria in selecting metadata and controlled-vocabulary schemata are collection-specific considerations, such as the types of resources, nature of the collection, and needs of primary users and communities. Existing technological infrastructure and staff expertise also are significant factors contributing to the current use of metadata schemata and controlled vocabularies for subject access across distributed digital repositories and collections. Metadata interoperability remains a major challenge. There is a lack of exposure of locally created metadata and metadata guidelines beyond the local environments. Homegrown locally added metadata elements may also hinder metadata interoperability across digital repositories and collections when there is a lack of sharable mechanisms for locally defined extensions and variants.

Metadata is an essential building block in facilitating effective resource discovery, access, and sharing across ever-growing distributed digital collections. Quality metadata is becoming critical in a networked world in which metadata interoperability is among the top challenges faced by digital libraries. However, there is no common data model that cataloging and metadata professionals can readily reference as a mediation mechanism during the processes of descriptive metadata creation and controlled vocabulary schemata application for subject description.[1] The development of such a mediation mechanism calls for an empirical assessment of various issues surrounding metadata-creation practices.

The critical issues concerning metadata practices across distributed digital collections have been relatively unexplored. While examining learning objects and e-prints communities of practice, Barton, Currier, and Hey point out the lack of formal investigation of the metadata-creation process.[2] As will be discussed in the following section, some researchers have begun to assess the current state of descriptive practices, metadata schemata, and content standards. However, the literature has not yet developed to a point where it affords a comprehensive picture. Given the propagation of metadata projects, it is important to continue to track changes in metadata-creation practices while they are still in constant flux. Such efforts are essential for adding new perspectives to digital library research and practices in an environment where metadata best practices are being actively sought after to aid in the creation and management of high-quality digital collections.

This study examines the prevailing current state of metadata-creation practices in digital repositories, collections, and libraries, which may include both digitized and born-digital resources. Using nationwide survey data, mostly drawn from the community of cataloging and metadata professionals, we seek to investigate issues in creating descriptive metadata elements, using controlled vocabularies for subject access, and propagating metadata and metadata guidelines beyond local environments.

We will address the following research questions:

1. Which metadata schema(ta) and content standard(s) are employed in individual digital repositories and collections?
2. Which controlled vocabulary schema(ta) are used to facilitate subject access?
3. What criteria are applied in selecting metadata and controlled-vocabulary schema(ta)?
4. To what extent are mechanisms for exposing and sharing metadata integrated into current metadata-creation practices?

In this article, we first review recent studies relating to current metadata-creation practices across digital collections. Then we present the survey method employed to conduct this study, the general characteristics of survey participants, and the validity of the collected data, followed by the study results. We report on how metadata and controlled vocabulary schema(ta) are being used across institutions, and we present a data analysis of current metadata-creation practices. The final section summarizes the study and presents some suggestions for future studies.

Jung-ran Park (jung-ran.park@ischool.drexel.edu) is Assistant Professor, College of Information Science and Technology, Drexel University, Philadelphia, and **Yuji Tosaka** (tosaka@tcnj.edu) is Cataloging/Metadata Librarian, TCNJ Library, The College of New Jersey, Ewing, New Jersey.

Literature Review

As evinced by the principles and practices of bibliographic control through shared cataloging, successful resource access and sharing in the networked environment demands semantic interoperability based on accurate, complete, and consistent resource description. The recent survey by Ma finds that the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) and metadata crosswalks have been adopted by 83 percent and 73 percent of respondents, respectively. Even though the sample comes only from sixty-eight Association of Research Libraries (ARL) member libraries, and the figures thus may be skewed higher than those of the entire population of academic libraries, there is little doubt that interoperability is a critical issue given the rapid proliferation of metadata schemata throughout digital libraries.[3]

While there is a variety of metadata schemata currently in use for organizing digital collections, only a few of them are widely used in digital repositories. In her ARL survey, Ma reports that the MARC format is the most widely used metadata schema (91 percent), followed by Encoded Archival Description (EAD) (84 percent), Unqualified Dublin Core (DC) (78 percent), and Qualified DC (67 percent).[4] Similarly, a 2007 member survey by OCLC Research Libraries Group (RLG) programs gathered information from eighteen major research libraries and cultural heritage institutions and also found that MARC is the most widely used scheme (65 percent), followed by EAD (43 percent), Unqualified DC (30 percent), and Qualified DC (29 percent). The different levels of use reported by these studies are probably due to different sample sizes and compositions, but results nonetheless suggest that metadata use at research institutions tends to rely on a small number of major schemata.[5]

There may in fact be much greater diversity in metadata use patterns when the scope is expanded to include both research and nonresearch institutions. Palmer, Zavalina, and Mustafoff, for example, tracked trends from 2003 through 2006 in metadata selection and application practices at more than 160 digital collections developed through Institute of Museum and Library Services grants. They found that despite perceived limitations, use of DC is the most widespread, with more than half of the digital collections using it alone or in combination with other schemata. MARC ranks second, with nearly 30 percent using it alone or in combination. The authors found that the choice of metadata schema is largely influenced by practices at peer institutions and compatibility with a content management system. What is most striking, however, is the finding that locally developed schemata are used as often as MARC. There is a decline in the percentage of digital projects using multiple metadata schemata (from 53 percent to 38 percent). Yet the authors also saw a possible increase in the use of locally developed schemata as many projects added new types of nontextual digital objects that could not be adequately described by existing metadata schemata.[6]

There is a lack of research concerning the current use of content standards; however, it is reasonable to suspect that content-standards use exhibits patterns similar to that of metadata because of their often close association with particular metadata schemata. The OCLC RLG survey reveals that Anglo-American Cataloguing Rules, 2nd edition (AACR2)—the traditional cataloging rule that has most often been used in conjunction with MARC—is the most widely used content standard (81 percent). AACR2 is followed by Describing Archives: A Content Standard (DACS) with 42 percent; Descriptive Cataloging of Rare Materials with 33 percent; Archives, Personal Papers, Manuscripts (APPM) with 25 percent; and Cataloging Cultural Objects (CCO) with 21 percent.[7]

In the same way as metadata schemata, there appears to be a concentration of a few controlled vocabulary schemata at research institutions. Ma's ARL survey, for example, shows that the Library of Congress Subject Headings (LCSH) and Name Authority File (NAF) were used by most survey respondents (96 percent and 88 percent, respectively). These two predominantly adopted vocabularies are followed by several domain-specific vocabularies, such as Art and Architecture Thesaurus (AAT), Library of Congress Thesaurus for Graphical Materials (TGM) I and II, Getty Thesaurus of Geographic Names (TGN), and the Getty Union List of Artists Names (ULAN), which were used by between 30 percent to more than 60 percent of respondents.[8] The OCLC RLG survey reports similar results; however, nearly half of the OCLC RLG survey respondents ($N = 9$) indicated that they had also built and maintained one or more locally developed thesauri.[9]

While creating and sharing information about local metadata implementations is an important step toward increased interoperability, recent studies tend to paint a grim picture of current local documentation practices and open accessibility. In a nationwide study of institutional repositories in U.S. academic libraries, Markey et al. found that only 61.3 percent of the 446 survey participants with operational institutional repositories had implemented policies for metadata schemata and authorized metadata creators.[10] The OCLC RLG survey also highlights limited collaboration and sharing of the metadata guidelines both within and across the institutions. It finds that even when there are multiple units creating metadata within the same institution, metadata-creation guidelines often are unlikely to be shared (28 percent do not share; 53 percent sometimes share).[11]

A mixed result is reported on the exposure of metadata to outside service providers. In an ARL survey, the University of Houston Libraries Institutional Repository

Task Force found that exposing metadata to OAI-PMH service providers is an established practice used by nearly 90 percent of the respondents.[12] Ma's ARL survey also reports the wide adoption of OAI-PMH (83 percent). These results underscore the virtual consensus on the critical importance of exposing metadata to achieve interoperability and make locally created metadata useful across distributed digital repositories and collections.[13] By contrast, the OCLC RLG survey shows that only one-tenth of the respondents stated that all non-MARC metadata is exposed to OAI harvesters, while 30 percent indicated that only some of it was available. The prominent theme revealed by the OCLC RLG survey is an "inward focus" in current metadata practices, marked by the "use of local tools to reach a generally local audience."[14]

In summary, recent studies show that the current practice of metadata creation is problematic due to the lack of a mechanism for integrating various types of metadata schemata, content standards, and controlled vocabularies in ways that promote an optimal level of interoperability across digital collections and repositories. The problems are exacerbated in an environment where many institutions lack local documentation delineating the metadata-creation process.

At the same time, researchers have only recently begun studying these issues, and the body of literature is at an incipient stage. The research that was done often targeted different populations, and sample sizes were different (some very small). In some cases the literature exhibits contradictory findings about issues surrounding metadata practices, increasing the difficulty in understanding the current state of metadata creation. This points out the need for further research of current metadata-creation practice.

Method

The objective of the research reported in this paper is to examine the current state of metadata-creation practices in terms of the creation of descriptive metadata elements, the use of controlled vocabularies for subject access, and the exposure of metadata and metadata guidelines beyond local environments. We conducted a Web survey using WebSurveyor (now Vovici: http://www.vovici.com). The survey included both structured and open-ended questions. It was extensively reviewed by members of an advisory board—a group of three experts in the field—and it was pilot-tested prior to being officially launched. The survey included many multiple-response questions that called for respondents to check all applicable answers.

We recruited participants through survey invitation messages and subsequent reminders to the electronic mailing lists of communities of metadata and cataloging professionals. Table 1 shows the mailing lists employed for the study. We also sent out individual invitations and distributed flyers to selected metadata and cataloging sessions during the 2008 ALA Midwinter Meeting, held that year in Philadelphia.

The survey attracted a large number of initial participants (N = 1,371), but during the sixty-two days from August 6 to October 6, 2008, we only received 303 completed responses via the survey management system. We suspect that the high incompletion rate (77.9 percent) stems from the fact that the subject matter may have been outside the scope of many participants' job responsibilities. The length of the survey may also have been a factor in the incompletion rate.

The profiles of respondents' job titles (see table 2)

Table 1. Electronic mailing lists for the survey

Electronic Mailing Lists	E-mail Address
Autocat	autocat@listserv.syr.edu
Dublin Core Listserv	dc-libraries@jiscmail.ac.uk
Metadata Librarians Listserv	metadatalibrarians@lists.monarchos.com
Library and Information Technology Association Listserv	lita-l@ala.org
Online Audiovisual Catalogers Electronic Discussion List	olac-list@listserv.acsu.buffalo.edu
Subject Authority Cooperative Program Listserv	sacolist@listserv.loc.gov
Serialist	serialst@list.uvm.edu
Text Encoding Initiative Listserv	tei-l@listserv.brown.edu
Electronic Resources in Libraries Listserv	eril-l@listserv.binghamton.edu
Encoded Archival Description Listserv	ead@listserv.loc.gov

Table 2. Job titles of participants (multiple responses)

Job Titles	Number of Participants
Other	135 (44.6%)
Cataloger/cataloging librarian/ catalog librarian	99 (32.7%)
Metadata librarian	29 (9.6%)
Catalog & metadata librarian	26 (8.6%)
Head, cataloging	26 (8.6%)
Electronic resources cataloger	17 (5.6%)
Cataloging coordinator	15 (5.0%)
Head, cataloging & metadata services	15 (5.0%)

N = 227. Survey question: What is your working job title? (please check all that apply)

and job responsibilities (see table 3) clearly show that most of the individuals who completed the survey engage professionally in activities directly relevant to the research objectives, such as descriptive and subject cataloging, metadata creation and management, authority control, nonprint and special material cataloging, electronic resource and digital project management, and integrated library system (ILS) management.

Although the largest number of participants (135, or 44.6 percent) chose the "Other" category regarding their job title (see table 2), it is reasonable to assume that the vast majority can be categorized as cataloging and metadata professionals.[15] Most job titles given as "Other" are associated with one of the professional activities listed in table 4.

Thus it is reasonable to assume that the respondents are in an appropriate position to provide first-hand, accurate information about the current state of metadata creation in their institutions.

Concerning the institutional background of participants, of the 303 survey participants, fewer than half (121, or 39.9 percent) provided institutional information. We believe that this is mostly due to the fact that the question was optional, following a suggestion from the Institutional Review Board at Drexel University. Of those that provided their institutional background, the majority (75.2 percent) are from academic libraries, followed by participants from public libraries (17.4 percent) and from other institutions (7.4 percent).

Table 3. Participants' job responsibilities (multiple responses)

Job Responsibilities	Number of Participants
General cataloging (e.g., descriptive and subject cataloging)	171 (56.4%)
Metadata creation and management	153 (50.5%)
Authority control	147 (48.5%)
Nonprint cataloging (e.g., microform, music scores, photographs, video recordings)	133 (43.9%)
Special material cataloging (e.g., rare books, foreign language materials, government documents)	126 (41.6%)
Digital project management	101 (33.3%)
Electronic resource management	62 (20.5%)
ILS management	59 (19.5%)
Other	51 (16.8%)

Survey question: What are your primary job responsibilities? (please check all that apply)

Table 4. Professional activities specified in "Other" category in table 2

Professional Activities	Number of Participants
Cataloging & metadata creation	31 (10.2%)
Digital projects management	23 (7.6%)
Technical services	17 (5.6%)
Archiving	16 (5.3%)
Electronic resources and serials management	6 (2.0%)
Library system administration/ other	6 (2.0%)

N = 99. Survey question: If you selected other, please specify.

Results

In this section, we will present the findings of this study in the following three areas: (1) metadata and controlled vocabulary schemata and metadata tools used, (2) criteria for selecting metadata and controlled vocabulary schemata, and (3) exposing metadata and metadata guidelines beyond local environments.

Metadata and Controlled Vocabulary Schemata and Metadata Tools Used

A great variety of digital objects were handled by the survey participants, as figure 1 shows. The most frequently handled object was text, cited by 86.5 percent of the respondents. About three-fourths of the respondents described audiovisual materials (75.2 percent), while 60.1 percent described images and 51.8 percent described archival materials. More than 65 percent of the respondents handled electronic resources (68.3 percent) and digitized resources (66.7 percent), while approximately half handled born-digital resources (52.5 percent). The types of materials described in digital collections were diverse, encompassing both digitized and born-digital materials; however, digitization accounted for a slightly greater percentage of metadata creation.

To handle these diverse digital objects, the respondents' institutions employed a wide range of metadata schemata, as figure 2 shows. Yet there were a few schemata that were widely used by cataloging and metadata professionals. Specifically, 84.2 percent of the respondents' institutions used MARC; DC was also popular, with 25.4 percent using Unqualified DC and 40.6 percent using Qualified DC to create metadata. EAD also was frequently cited (31.7 percent). In addition to these major types of metadata schemata, the respondents' institutions also employed Metadata Object Description Schema (MODS) (17.8 percent), Visual Resource Association (VRA) Core (14.9 percent), and Text Encoding Initiative (TEI) (12.5 percent).

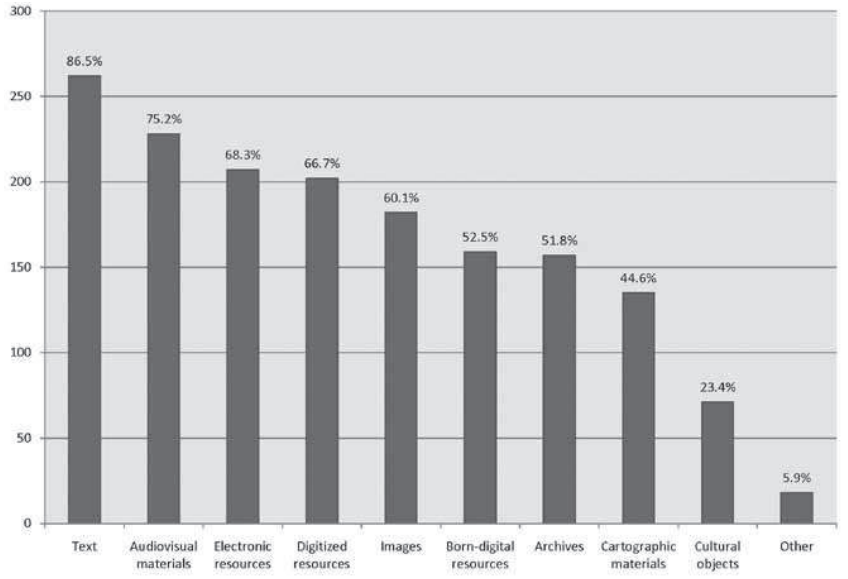

Figure 1. Materials/resources handled (multiple responses)

Survey question: What type of materials/resources do you and your fellow catalogers/metadata librarians handle? (please check all that apply)

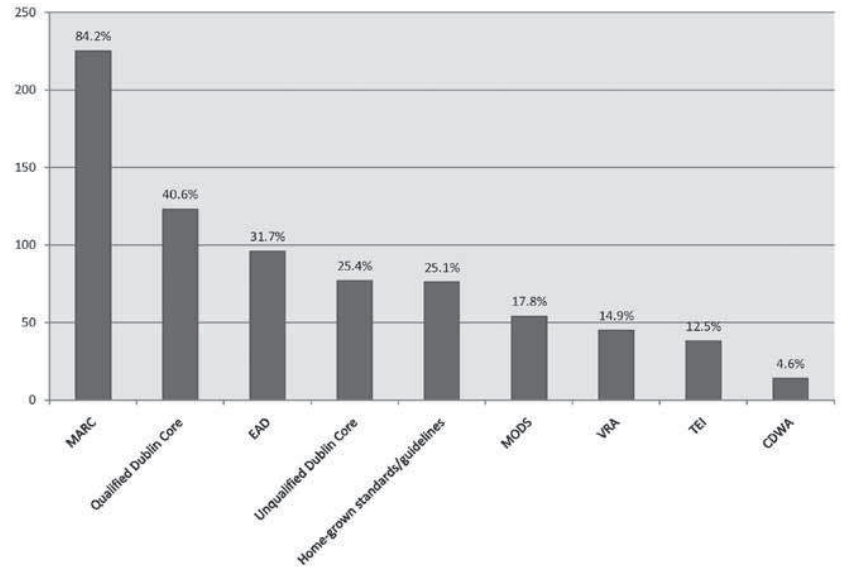

Figure 2. Metadata schemata used (multiple responses)

Survey question: Which metadata schema(s) do you and your fellow catalogers/metadata librarians use? (please check all that apply)

It is noteworthy that use of Qualified DC was higher than that of Unqualified DC. This result is different from the ARL survey and a member survey conducted

by OCLC RLG programs (as described in "Literature Review" on page 105).[16] In these surveys, Unqualified DC was more frequently cited than Qualified DC. One possible explanation of this less frequent use of Unqualified DC may lie in the limitations of Unqualified DC metadata semantics. Survey respondents also reported on problems using DC metadata, which were mostly caused by semantic ambiguities and semantic overlaps of certain DC metadata elements.[17] Limitations and issues of Unqualified DC metadata semantics are discussed in depth in Park's study.[18] In light of these results, examining trends of Qualified DC use in a future study would be interesting.

Despite the wide variety of schemata reported in use, there seemed to be an inclination to use only one or two metadata schemata for resource description. As shown in table 5, the majority of the respondents' institutions (53.6 percent) used only one schema for metadata creation, while approximately 37 percent used two or three schemata (26.2 percent and 10.3 percent, respectively). The institutions using more than three schemata during the metadata-creation processes comprised only 9.9 percent of the respondents.

Turning to content standards (see figure 3), we found that AACR2 was the most widely used standard, indicated by 84.5 percent of respondents. This high percentage clearly reflects the continuing preeminence of MARC as the metadata schema of choice for digital collections. DC application profiles also showed a large user base, indicated by more than one-third of respondents (37.0 percent). More than one quarter of the respondents (28.4 percent) used EAD application guidelines as developed by the Society of American Archivists and the Library of Congress, while 10.6 percent used RLG Best Practice Guidelines for Encoded Archival Description (2002). About one quarter (25.7 percent) indicated DACS as their content standard.

Homegrown standards and guidelines are local application profiles that clarify existing content standards and specify how values for metadata elements are selected and represented to meet the requirements of a particular context. As shown in the results on metadata schemata, it is noteworthy that homegrown content standards and guidelines constituted one of the major choices of participants, indicated by more than one-fifth of the institutions (22.1 percent). Almost two-fifths of the survey participants (38 percent) also reported that they add homegrown metadata elements to a given metadata schema. Slightly less than half of the participants (47.2 percent) indicated otherwise.

The local practice of creating homegrown content guidelines and metadata elements during the metadata-creation process deserves a separate study; this study only briefly touches on the basis for locally added custom metadata elements. The motivation to create custom metadata elements derives from the imperative to accommodate the perceived needs of local collections and users, as indicated by the two most common responses: (1) "to reflect the nature of local collections/resources" (76.9 percent) and (2) "to reflect the characteristics of target audience/community of local collections" (58.3 percent). Local conditions were also cited from institutional and technical standpoints. Many institutions (34.3 percent) follow existing local practices for cataloging and metadata creation while other institutions (18.5 percent) are making homegrown metadata additions because of constraints imposed by their local systems.

Table 5. Number of metadata schemata in use

Number of Metadata Schemata in Use	Number of Participants
1	141 (53.6%)
2	69 (26.2%)
3	27 (10.3%)
4 or more	26 (9.9%)

N=263. Survey question: Which metadata schema(s) do you and your fellow catalogers/metadata librarians use the most? (please check all that apply)

Table 6 summarizes the most frequently used controlled vocabulary schematas by resource type. By far the most widely used schema across all resource types was LCSH. The preeminence of LCSH evinces the critical role that it plays as the de facto form of controlled vocabulary for subject description. Library of Congress Classification (LCC) was the second choice for all resource types other than images, cultural objects, and archives. For digital collections of these resource types and digitized resources, AAT was the second most used controlled vocabulary, a fact that reflects its purpose as a domain-specific terminology used for describing works of art, architecture, visual resources, material culture, and archival materials.

While traditional metadata schemata, content standards, and controlled vocabularies such as MARC, AACR2, and LCSH clearly were preeminent in the majority of the respondents' institutions, current metadata creation in digital repositories and collections faces new challenges from the enormous volume of online and digital resources.[19] Approximately one-third of the respondents' institutions (33.8 percent) were meeting this challenge with tools for semiautomatic metadata generation. Yet a majority of respondents (52.5 percent) indicated that their institutions did not use any such tools for metadata creation and management. This result seems to contrast with Ma's finding that automatic metadata generation was used in some capacity in nearly

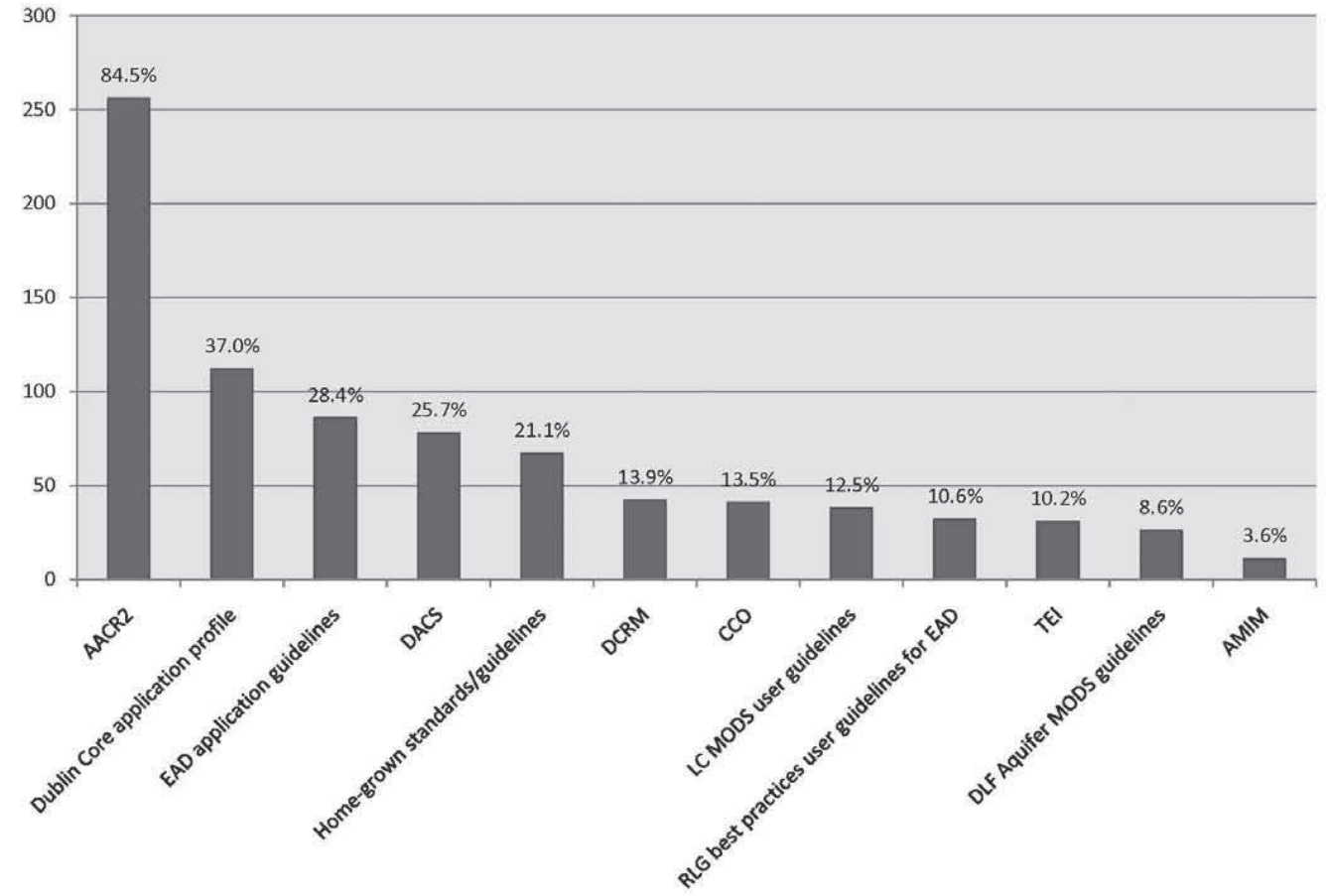

Figure 3. Content standards used (multiple responses)

Survey question: What content standard(s) and/or guidelines do you and your fellow catalogers/metadata librarians use? (please check all that apply)

two-thirds of ARL libraries.[20]

Because semiautomatic metadata application is reported in-depth in a separate study, we only briefly sketch the topic here.[21] The semiautomatic metadata application tools used in the respondents' digital repositories and collections can be classified into five categories of common characteristics: (1) metadata format conversion, (2) templates and editors for metadata creation, (3) automatic metadata creation, (4) library system for bibliographic and authority control, and (5) metadata harvesting and importing tools.

As table 7 illustrates, among those institutions that have introduced semiautomatic metadata generation tools, "metadata format conversion" (38.6 percent) and "templates and editors for metadata creation" (27 percent) are the two most frequently cited tools.

Criteria for Selecting Metadata and Controlled Vocabulary Schemata

What are the factors that have shaped the current state of metadata-creation practices reported thus far? In this section, we turn our attention to constraints that affect decision making at institutions in the selection of metadata and controlled vocabulary schemata for subject description.

Figure 4 presents the percentage of different metadata schemata selection criteria described by survey participants. First, collection-specific considerations clearly played a major role in the selection. The most frequently cited reason was "types of resources" (60.4 percent). This response reflects the fact that a large number of metadata schemata have been developed, often with wide variation in content and format, to better handle particular

Table 6. The most frequently used controlled vocabulary schema(s) by resource type (multiple responses)

	LCSH	LCC	DDC	AAT	TGM	ULAN	TGN	Other
Text	79.5% (241)	35.6% (108)	16.8% (51)	10.2% (31)	6.9% (21)	3.6% (11)	5.0% (15)	14.2% (43)
Audiovisual materials	67.3% (204)	25.1% (76)	12.9% (39)	9.2% (28)	8.6% (26)	4.0% (12)	5.0% (15)	14.5% (44)
Cartographic materials	44.9% (136)	17.5% (53)	7.3% (22)	5.0% (15)	4.3% (13)	1.3% (4)	4.3% (13)	6.3% (19)
Images	43.2% (131)	11.9% (36)	5.6% (17)	25.7% (78)	20.1% (61)	9.9% (30)	10.6% (32)	11.2% (34)
Cultural objects (e.g., museum objects)	20.1% (61)	7.3% (22)	4.3% (13)	13.2% (40)	6.3% (19)	4.6% (14)	3.0% (9)	7.9% (24)
Archives	44.2% (134)	11.6% (35)	6.3% (19)	11.9% (36)	6.6% (20)	3.0% (9)	2.6% (8)	12.2% (37)
Electronic resources	60.7% (184)	23.4% (71)	8.6% (26)	5.3% (16)	3.6% (11)	1.7% (5)	3.0% (9)	14.2% (43)
Digitized resources	51.8% (157)	15.5% (47)	5.0% (15)	15.5% (47)	10.2% (31)	6.6% (20)	7.6% (23)	15.2% (46)
Born-digital resources	43.9% (133)	13.5% (41)	5.6% (17)	8.3% (25)	7.3% (22)	4.3% (13)	4.6% (14)	13.9% (42)

Survey question: Which controlled vocabulary schema(s) do you and your fellow catalogers/metadata librarians use most? (Please check all that apply)

types of information resources. The primary factor in selecting metadata schemata is their suitability for describing the most common type of resources handled by the survey participants.

The second and third most common criteria, "target users/audience" (49.8 percent) and "subject matters of resources" (46.9 percent), also seem to reflect how domain-specific metadata schemata are applied. In making decisions on metadata schemata, respondents weighed materials in particular subject areas (e.g., art, education, and geography) and the needs of particular communities of practice as their primary users and audiences.

However, existing technological infrastructure and resource constraints also determine options. Given the prominence of general library cataloging as a primary job responsibility, "expertise of staff" (44.2 percent) and "integrated library system" (39.9 percent) appeared to highlight the key role that MARC continues to play in the metadata-creation process for digital collections (see figure 2). "Budget" also appeared to be an important factor in metadata selection (17.2 percent), showing that funding levels played a considerable role in metadata decisions.

Table 7. Types of semi-automatic metadata generation tools in use

Types	Response Rating
Metadata format conversion	38 (38.6%)
Templates and editors for metadata creation	26 (27.0%)
Automatic metadata creation	16 (16.7%)
Library system for bibliographic and authority control	15 (15.6%)
Metadata harvesting and importing tools	8 (8.3%)

N = 96. Survey question: Please describe the (semi)automatic metadata generation tools you use.

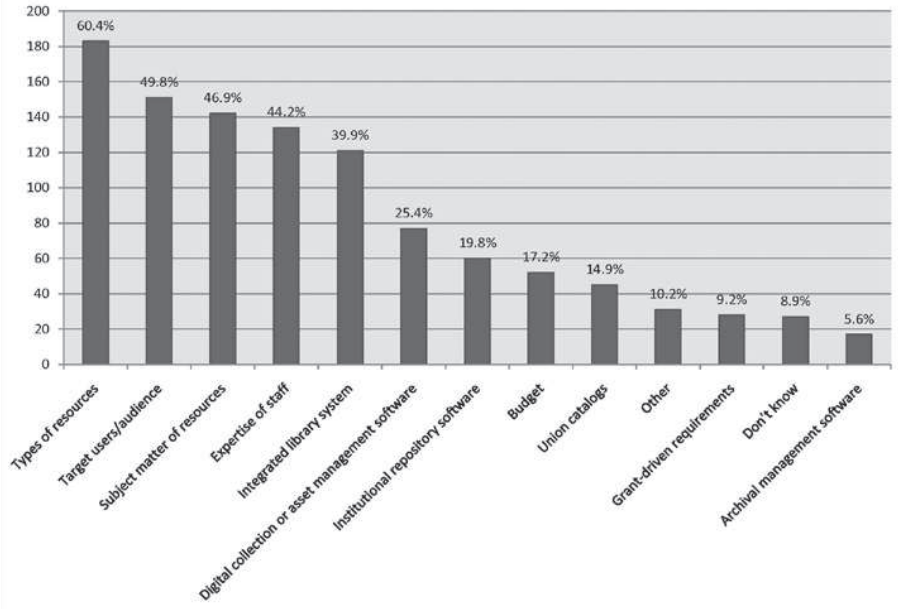

Figure 4. Criteria for selecting metadata schemata (multiple responses)
Question: Which criteria were applied in selecting metadata schemata? (please check all that apply)

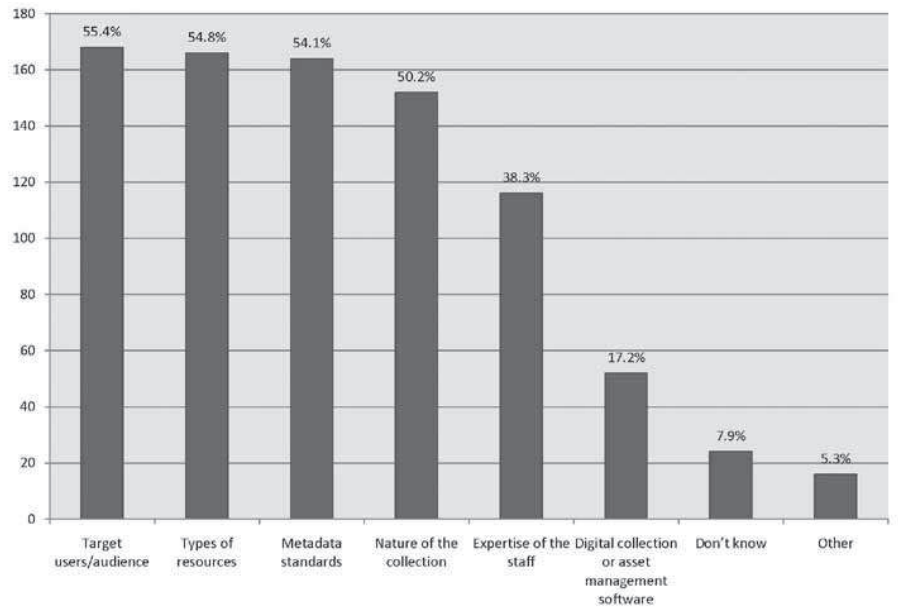

Figure 5. Criteria for selecting controlled vocabulary schemata (multiple responses)
Question: Which criteria are applied in selecting controlled vocabulary schemata? (Please check all that apply)

At the same time, it is noteworthy that while responses were not mutually exclusive, many respondents cited the software used by their institutions—i.e., "integrated library system" (39.9 percent), "digital collection or asset management software" (25.4 percent), "institutional repository software" (19.8 percent), "union catalogs" (14.9 percent), and "archival management software" (5.6 percent)—as a reason for their selection of metadata schemata. Metadata decisions thus seem to be driven by a variety of local technology choices for developing digital repositories and collections.

As shown in figure 5, similar patterns are observed with regard to selection criteria for controlled vocabulary schemata. Three of the four selection criteria receiving majority responses—"target users/audience" (55.4 percent), "type of resources" (54.8 percent), and "nature of the collection" (50.2 percent)—suggest that controlled vocabulary decisions are influenced primarily by the substantive purpose and scope of controlled vocabularies for local collections. A major consideration seems to be whether particular controlled vocabularies are suitable for representing standard data values to improve access and retrieval for target audiences.

"Metadata standards," another selection criteria frequently cited in the survey (54.1 percent), reflects how some domain-specific metadata schemata tend to dictate the use of particular controlled vocabularies. At the same time, the results also suggest that resources and technological infrastructure available to institutions were also important reasons for their selections. "Expertise of staff" (38.3 percent) seems to be a straightforward practical reason: the application of controlled vocabularies is highly dependent on the width and depth of staff expertise available. Likewise, when implementing controlled vocabularies in the digital environment, some institutions also took into account

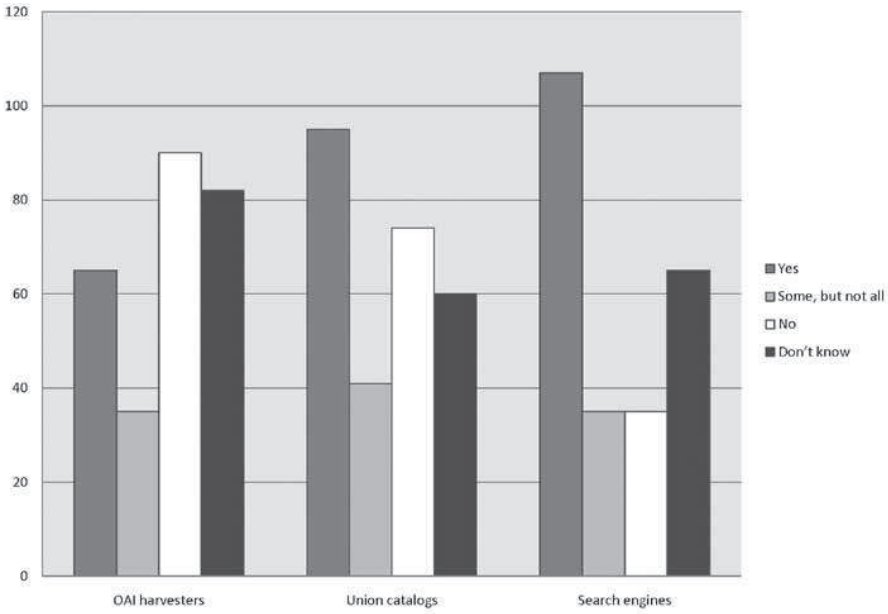

Figure 6. Mechanism to expose metadata (multiple responses)

Survey question: Do you/your organization expose your metadata to OAI (Open Archives Initiative) harvesters, union catalogs or search engines?

existing system features for authority control and controlled vocabulary searching, as exhibited by 17.2 percent of responses for "digital collection/or asset management software."

Exposing Metadata and Metadata Guidelines beyond Local Environments

Metadata interoperability across distributed digital repositories and collections is fast becoming a major issue.[22] The proliferation of open-source and commercial digital library platforms using a variety of metadata schemata has implications on the librarians' ability to create shareable and interoperable metadata beyond the local environment. To what extent are mechanisms for sharing metadata integrated into the current metadata-creation practices described by the respondents?

Figure 6 summarizes the responses concerning the uses of three major mechanisms for metadata exposure. Approximately half of respondents exposed at least some of their metadata to search engines (52.8 percent) and union catalogs such as OCLC WorldCat (50.6 percent). More than one-third of the respondents exposed all or some of their metadata through OAI harvesters (36.8 percent). About half or more of the respondents either did not expose their metadata or were not sure about the current operations at their institutions (e.g., 47.2 percent for search engines and 63.2 percent for OAI harvesters), a result that may be interpreted as a tendency to create metadata primarily for local audiences.

Why do many institutions fail to make their locally created metadata available to other institutions despite wide consensus on the importance of metadata sharing in a networked world? Responses from those institutions exposing none or not all of their metadata (see table 8) reveal that financial, personnel, and technical issues are major hindrances in promoting the exposure of metadata outside the immediate local environment. Some institutions are not confident that their current metadata practices are able to satisfy the technical requirements for producing standards-based interoperable metadata. Another reason frequently mentioned is copyright concerns about limited-access materials. Yet some respondents simply do not see any merit to exposing their item-level metadata, citing its relative uselessness for resource discovery outside their institutions.

As stated earlier, the practice of adding homegrown metadata elements seems common among many institutions. While locally created metadata elements accommodate local needs and requirements, they may also hinder metadata interoperability across digital repositories and collections if mechanisms for finding information about such locally defined extensions and variants are absent. Homegrown metadata guidelines document local data models and function as an essential mechanism for metadata creation and quality assurance within and across digital repositories and collections.[23] In this regard, it is essential to examine locally created metadata guidelines and best practices.[24] However, the results of the survey analysis evince that the vast majority of institutions (72.0 percent) provided no public access to local application profiles on their websites while only 19.6 percent of respondents' institutions made them available online to the public.

Conclusion

Metadata plays an essential role in managing, organizing, and searching for information resources. In the networked

Table 8. Sample reasons for not exposing metadata

Not all our metadata conforms to standards required
Not all our metadata is OAI compliant
Lack of expertise and time and money to develop it
IT restrictions
Security concerns on the part of our information technology department
Some collections/records are limited access and not open to the general public
We think that having WorldCat available for traditional library materials that many libraries have is a better service to people than having each library dump our catalog out on the web
Varies by tool and collection, but usually a restriction on the material, a technical barrier, or a feeling that for some collections the data is not yet sufficiently robust "still in a work in progress"

Survey question: If you selected "some, but not all" or "no" in question 13 [see figure 6], please tell why you do not expose your metadata.

environment, the enormous volume of online and digital resources creates an impending research need to evaluate the issues surrounding the metadata-creation process and the employment of controlled vocabulary schemata across ever-growing distributed digital repositories and collections. In this paper we explored the current status of metadata-creation practices through an examination of survey responses drawn mostly from cataloging and metadata professionals (see tables 2, 3, and 4). The results of the study indicate that current metadata practices still do not create conditions for interoperability.

Despite the proliferation of newer metadata schemata, the survey responses showed that MARC currently remains the most widely used schema for providing resource description and access in digital repositories, collections, and libraries. The continuing predominance of MARC goes hand-in-hand with the use of AACR2 as the primary content standard for selecting and representing data values for descriptive metadata elements. LCSH is used as the de facto controlled vocabulary schema for providing subject access in all types of digital repositories and collections, while domain-specific subject terminologies such as AAT are applied at significantly higher rates in digital repositories handling nonprint resources such as images, cultural objects, and archival materials.

The DC metadata schema is the second most widely employed according to this study, with Qualified DC used by 40.6 percent of responding institutions and Unqualified DC used by 25.4 percent. EAD is another frequently cited schema (31.7 percent), followed by MODS (17.8 percent), VRA (14.9 percent), and TEI (12.5 percent). A trend of Qualified DC being used (40.6 percent) more often than Unqualified DC (25.4 percent) is noteworthy. One possible explanation of this trend may be derived from the fact that semantic ambiguities and overlaps in some of the Unqualified DC elements interfere with use in resource description.[25] Given the earlier surveys reporting the higher use of Unqualified DC over Qualified DC, more in-depth examination of their use trends may be an important avenue for future studies.

Despite active research and promising results obtained from some experimental tools, practical applications of semiautomatic metadata generation have been incorporated into the metadata-creation processes by only one-third of survey participants.

The leading criteria in selecting metadata and controlled vocabulary schemata are derived from collection-specific considerations of the type of resources, the nature of the collections, and the needs of primary users and communities. Existing technological infrastructure, encompassing digital collection or asset management software, archival management software, institutional repository software, integrated library systems, and union catalogs also greatly influence the selection process. The skills and knowledge of metadata professionals and the expertise of staff also are significant factors in understanding current practices in the use of metadata schemata and controlled vocabularies for subject access across distributed digital repositories and collections.

The survey responses reveal that metadata interoperability remains a challenge in the current networked environment despite growing awareness of its importance. For half of the survey respondents, exposing metadata to the service providers, such as OAI harvesters, union catalogs, and search engines, does not seem to be a high priority because of local financial, personnel, and technical constraints. Locally created metadata elements are added in many digital repositories and collections in large part to meet local descriptive needs and serve the target user community. While locally created metadata elements accommodate local needs, they may also hinder metadata interoperability across digital repositories and collections when shareable mechanisms are not in place for such locally defined extensions and variants.

Locally created metadata guidelines and application profiles are essential for metadata creation and quality assurance; however, most custom content guidelines and best practices (72 percent) are not made publicly available. The lack of a mechanism to facilitate public access to local application profiles and metadata guidelines may

hinder cross-checking for quality metadata and creating shareable metadata that can be harvested for a high level of consistency and interoperability across distributed digital collections and repositories. Development of a searchable registry for publicly available metadata guidelines has potential to enhance metadata interoperability.

A constraining factor of this study derives from the participant population; thus we have not attempted to generalize the findings of the study. However, results indicate a pressing need for a common data model that is shareable and interoperable across ever-growing distributed digital repositories and collections. Development of such a common data model demands future research of a practical and interoperable mediation mechanism underlying local implementation of metadata elements, semantics, content standards, and controlled vocabularies in a world where metadata can be distributed and shared widely beyond the immediate local environment and user community. (Other issues such as semiautomatic metadata application, DC metadata semantics, custom metadata elements, and the professional development of cataloging and metadata professionals are explained in-depth in separate studies.)[26] For future studies, incorporation of other research methods (such as follow-up telephone surveys and face-to-face focus group interviews) could be used to better understand the current status of metadata-creation practices. Institutional variation also needs be taken into account in the design of future studies.

Acknowledgments

This study is supported through an early career development research award from the Institute of Museum and Library Services. We would like to express our appreciation to the reviewers for their invaluable comments.

References

1. Jung-ran Park, "Semantic Interoperability and Metadata Quality: An Analysis of Metadata Item Records of Digital Image Collections," *Knowledge Organization* 33 (2006): 20–34; Rachel Heery, "Metadata Futures: Steps toward Semantic Interoperability," in *Metadata in Practice*, ed. Diane I. Hillman and Elaine L. Westbrooks, 257–71 (Chicago: ALA, 2004); Jung-ran Park, "Semantic Interoperability across Digital Image Collections: A Pilot Study on Metadata Mapping" (paper presented at the Canadian Association for Information Science 2005 Annual Conference, London, Ontario, June 2–4, 2005), http://www.cais-acsi.ca/proceedings/2005/park_J_2005.pdf (accessed Mar. 24, 2009).

2. Jane Barton, Sarah Currier, and Jessie M. N. Hey, "Building Quality Assurance into Metadata Creation: An Analysis Based on the Learning Objects and E-Prints Communities of Practice" (paper presented at 2003 Dublin Core Conference: Supporting Communities of Discourse and Practice—Metadata Research & Applications, Seattle, Wash., Sept. 28–Oct. 2, 2003), http://dcpapers.dublincore.org/ojs/pubs/article/view/732/728 (accessed Mar. 24, 2009); Sarah Currier et al., "Quality Assurance for Digital Learning Object Repositories: Issues for the Metadata-creation process," *ALT-J* 12 (2004): 5–20.

3. Jin Ma, *Metadata*, SPEC Kit 298 (Washington, D.C.: Association of Research Libraries, 2007): 13, 28.

4. Ibid., 12, 21–22.

5. Karen Smith-Yoshimura, *RLG Programs Descriptive Metadata Practices Survey Results* (Dublin, Ohio: OCLC, 2007): 6–7, http://www.oclc.org/programs/publications/reports/2007-03.pdf (accessed Mar. 24, 2009); Karen Smith-Yoshimura and Diane Cellentani, *RLG Programs Descriptive Metadata Practices Survey Results: Data Supplement* (Dublin, Ohio: OCLC, 2007): 16, http://www.oclc.org/programs/publications/reports/2007-04.pdf (accessed Mar. 24, 2009).

6. Carole Palmer, Oksana Zavalina, and Megan Mustafoff, "Trends in Metadata Practices: A Longitudinal Study of Collection Federation" (paper presented at the Seventh ACM/IEES-CS Joint Conference on Digital Libraries, Vancouver, British Columbia, Canada, June 18–23, 2007), http://hdl.handle.net/2142/8984 (accessed Mar. 24, 2009).

7. Smith-Yoshimura, *RLG Programs Descriptive Metadata Practices Survey Results*, 7; Smith-Yoshimura and Cellentani, *RLG Programs Descriptive Metadata Practices Survey Results*, 17.

8. Ma, *Metadata*, 12, 22–23.

9. Smith-Yoshimura, *RLG Programs Descriptive Metadata Practices Survey Results*, 7; Smith-Yoshimura and Cellentani, *RLG Programs Descriptive Metadata Practices Survey Results*, 18–21.

10. Karen Markey et al., *Census of Institutional Repositories in the United States: MIRACLE Project Research Findings* (Washington, D.C.: Council on Library & Information Resources, 2007): 3, 46–50, http://www.clir.org/pubs/reports/pub140/pub140.pdf (accessed Mar. 24, 2009).

11. Yoshimura and Cellentani, *RLG Programs Descriptive Metadata Practices Survey Results*, 24.

12. University of Houston Libraries Institutional Repository Task Force, *Institutional Repositories*, SPEC Kit 292 (Washington, D.C.: Association of Research Libraries, 2006): 18, 78.

13. Ma, *Metadata*, 13, 28.

14. Smith-Yoshimura, *RLG Programs Descriptive Metadata Practices Survey Results*, 9, 11; Smith-Yoshimura and Cellentani, *RLG Programs Descriptive Metadata Practices Survey Results*, 27–29.

15. For the metrics of job responsibilities used to analyze job descriptions and competencies of cataloging and metadata professionals, see Jung-ran Park, Caimei Lu, and Linda Marion, "Cataloging Professionals in the Digital Environment: A Content Analysis of Job Descriptions," *Journal of the American Society for Information Science & Technology* 60 (2009): 844–57; Jung-ran Park and Caimei Lu, "Metadata Professionals: Roles and Competencies as Reflected in Job Announcements, 2003–2006," *Cataloging & Classification Quarterly* 47 (2009): 145–60.

16. Ma, *Metadata*; Smith-Yoshimura, *RLG Programs Descriptive Metadata Practices Survey Result*.

17. Jung-ran Park and Eric Childress, "Dublin Core Metadata Semantics: An Analysis of the Perspectives of Information Professionals," *Joural of Information Science* 35, no. 6 (2009): 727–39.

18. Park, "Semantic Interoperability."

19. Jung-ran Park, "Metadata Quality in Digital Repositories:

A Survey of the Current State of the Art," in "Metadata and Open Access Repositories," ed. Michael S. Babinec and Holly Mercer, special issue, *Cataloging & Classification Quarterly* 47, no. 3/4 (2009): 213–38.

20. Ma, *Metadata*, 12, 24. The OCLC RLG survey found that about 40 percent of the respondents were able to generate some metadata automatically. See Smith-Yoshimura, *RLG Programs Descriptive Metadata Practices Survey Results*, 6; Yoshimura and Cellentani, *RLG Programs Descriptive Metadata Practices Survey Results*, 35.

21. Jung-ran Park and Caimei Lu, "Application of Semi-Automatic Metadata Generation in Libraries: Types, Tools, and Techniques," *Library & Information Science Research* 31, no. 4 (2009): 225–31.

22. Park, "Semantic Interoperability"; Sarah L. Shreeves et al., "Is 'Quality' Metadata 'Shareable' Metadata? The Implications of Local Metadata Practices for Federated Collections" (paper presented at the 12th National Conference of the Association of College and Research Libraries, Apr. 7–10, 2005, Minneapolis, Minnesota), https://www.ideals.uiuc.edu/handle/2142/145 (accessed Mar. 24, 2009); Amy S. Jackson et al., "Dublin Core Metadata Harvested through OAI-PMH," *Journal of Library Metadata* 8, no. 1 (2008): 5–21; Lois Mai Chan and Marcia Lei Zeng, "Metadata Interoperability and Standardization—A Study of Methodology Part I: Achieving Interoperability at the Schema Level," *D-Lib Magazine* 12, no. 6 (2006), http://www.dlib.org/dlib/june06/chan/06chan.html (accessed Mar. 24, 2009); Marcia Lei Zeng and Lois Mai Chan, "Metadata Interoperability and Standardization—A Study of Methodology Part II: Achieving Interoperability at the Record and Repository Levels," *D-Lib Magazine* 12, no. 6 (2006), http://www.dlib.org/dlib/june06/zeng/06zeng.html (accessed Mar. 24, 2009).

23. Thomas R. Bruce and Diane I. Hillmann, "The Continuum of Metadata Quality: Defining, Expressing, Exploiting," in *Metadata in Practice*, ed. Hillman and Westbrooks, 238–56; Heery, "Metadata Futures"; Park, "Metadata Quality in Digital Repositories."

24. Jung-ran Park, ed., "Metadata Best Practices: Current Issues and Future Trends," special issue, *Journal of Library Metadata* 9, no. 3/4 (2009).

25. See Park, "Semantic Interoperability"; Park and Childress, "Dublin Core Metadata Semantics."

26. Park and Childress, "Dublin Core Metadata Semantics"; Park and Lu, "Application of Semi-Automatic Metadata Generation in Libraries."