Alexandra M. Chassanoff. Metadata Quality Evaluation in Institutional Repositories: A Survey of Current Practices. A Master's Paper for the M.S. in I.S degree. December, 2009. 48 pages. Advisor: Jane Greenberg

Metadata plays an important role in the discovery, access, and use of materials in institutional repositories (IRs). Thus far, little empirical research been conducted to assess and evaluate metadata quality practices in place. This study begins to address that gap in knowledge by gathering data on current practices and procedures relating to metadata quality and evaluation in institutional repositories. A survey was distributed to individuals at ARL-member institutional repositories with knowledge of their institution's metadata procedures. The survey specifically gathered data on what metadata practices were in place and whether quality control procedures were being used. Forty respondents provided results that offer a state of the art view into the current metadata quality practices in place at IRs. Survey results indicate that metadata activities may not yet be streamlined into institutional workflow. For most institutions, metadata quality checking is a manual process, with only a small percentage (4%) employing the use of automated tools. Additionally, institutions rely on users as much as repository to staff to discover quality problems. Other results indicate that the majority of institutions surveyed are maintaining documentation relating to metadata policies. For example, 75% of respondents reported that their institution had developed either minimum metadata requirements or metadata submission guidelines for contributors. Overall, these results reflect the challenges and growing pains facing institutions as they adapt to managing materials in the digital world.

Headings:

Institutional Repositories--United States

Surveys--Metadata

Metadata--Quality Control

Metadata--Management

METADATA QUALITY EVALUATION IN INSTITUTIONAL REPOSITORIES: A SURVEY OF CURRENT PRACTICES

by
Alexandra M. Chassanoff

A Master's paper submitted to the faculty
of the School of Information and Library Science
of the University of North Carolina at Chapel Hill
in partial fulfillment of the requirements
for the degree of Master of Science in
Information Science.

Chapel Hill, North Carolina

December 2009

Approved by

_____

Jane Greenberg

Table of Contents

Introduction

The growth of digital resources in the last thirty years has presented both unique opportunities and problems to the higher education community. On the one hand, the emergence of a networked environment has enabled scholars to have wider access to materials. The changing landscape has created new norms and expectations for service; researchers no longer have to visit the library to retrieve journal articles, unlike even ten years ago. On the other hand, the subsequent effective management of these digital resources requires the presence of robust infrastructures.

One response to the growing body of digital scholarship has been the development of academic institutional repositories in the last decade. Lynch (2003) defines institutional repositories as consisting of "a set of services that a university offers to the members of its community for the management and dissemination of digital materials created by the institution and its community members" (Defining institutional repositories section, para. 1). Their diverse holdings can consist of many different types of resources (scientific data sets, journal articles, unpublished research) and a host of different materials and formats (posters, digital images, video, text files, presentations, among other types). In her survey of American academic institutional repositories, McDowell (2007) found the majority of content to be student-produced, including over 93,000 electronic theses and dissertations (Composite results section).

The management and stewardship of digital scholarship for use by the wider community is largely enabled through the creation and maintenance of metadata. Greenberg (2001) defines metadata as "structured data about data that supports the discovery, use, authentication, and administration of information objects" (p. 918). Metadata acts as a physical surrogate for digital objects by describing its corresponding properties. In the context of institutional repositories, it follows that successful metadata creation will support and enable the discovery, access, and use of digital resources. Presumably, unsuccessful (or poor quality) metadata creation will impact the ability of users to carry out these same functionalities. Consequently, institutional repositories should include practices to evaluate metadata quality as part of their infrastructure.

Many research studies have been undertaken exploring the relationship between metadata quality and digital repository usage (see Anderson, 2006; Bruce & Hillman, 2004; Lagoze et al., 2006, Shreeves et al., 2005). Yet relatively little empirical research has been conducted on metadata quality and evaluation in the specific domain of institutional repositories. There are currently no standardized methods or procedures in place for evaluating metadata; instead, only conceptual frameworks for evaluation and suggested quality criteria indicators exist. Furthermore, the number of institutional repositories employing metadata quality control practices is yet unknown.

This research seeks to address this gap in knowledge by asking the following questions:

1. What metadata practices are currently in use at institutional repositories?
2. Are institutional repositories employing metadata quality control procedures?

The results of this research provide a state-of-the-art perspective on current metadata quality evaluation practices and procedures in institutional repositories. Additionally, the following results are reported on for US academic institutional repositories:

- Overall impressions of metadata quality
- The percentage with quality control procedures in place
- The percentage using quality-enforcing structures such as controlled vocabularies
- The percentage providing metadata guidelines for depositors
- The percentage with existing documentation on metadata quality
- Descriptions of how repositories discover quality issues

Overall, these results provide baseline data for determining quality measures and validation points that can better guide metadata generation during the submission process.

Literature Review

2.1 The domain of institutional repositories

The emergence of institutional repositories in academic settings is a relatively recent phenomenon. Lynch & Lippincott (2005) conducted one of the first investigations of the state of academic institutional repository deployment in the United States through their work with the Coalition for Networked Information (CNI). One limitation of the study was its small sampling frame (124), roughly half of all U.S. doctoral-granting research universities, with 97 survey participants. However, the study did offer an early glimpse into deployment statistics; 41 repositories were reported to be fully operational. One complicating factor in measuring deployment has been the reluctance of researchers to define institutional repositories for survey participants. For example, the CNI study requested that respondents complete their survey with their own view of what constitutes an institutional repository (Survey of US higher education institutions section, para. 6).

Similarly, the absence of a widely-agreed upon definition of institutional repositories has been problematic. A variety of organizational models based on differing visions have been proposed. In a SPARC position paper, Crow (2002) writes that institutional repositories should serve as alternative publishing models for scholarly communication. He explains, “Institutional repositories can provide an immediate and valuable complement to the existing scholarly publishing model, while stimulating

innovation in a new disaggregated publishing structure that will evolve and improve over time" (p.5). On the other hand, Lynch, who provides the key definition in use for my research study, advocates that institutional repositories supplement, rather than replace, traditional publishing models. A key distinction is Lynch's emphasis on the lifecycle management of digital scholarship. Such a distinction has wide-ranging implications for the establishment of a model which must support access, use, and preservation services.

The development of the Open Archive Initiative (OAI) in 2000 proved to be highly influential on the organizational model of institutional repositories. The OAI was born out of the EPrint community, and was initially concerned with providing interoperable standards and guidelines for disseminating open-access digital content (Lagoze, 2005). An important manifestation of these efforts was the creation of the OAI-PMH, a standard protocol establishing for harvesting metadata from different OAI-compliant data providers. The OAI-PMH standards enable the large-scale aggregation of compliant resources, both within individual repositories and throughout an established network or domain. As Hitchcock (2007) explains, "For the first time institutions such as universities have the ability to capture, store *and disseminate* copies of the published work of their own researchers. The importance of this cannot be understated" (Evolution of institutional repositories, para. 2). The protocol has since been widely adopted for use by institutional repositories; according to the online University of Illinois OAI-PMH Data Provider Registry, there are currently 2,156 OAI-compliant repositories actively providing data world-wide.

2.2 Metadata quality in institutional repositories

The unique qualities that characterize academic institutional repositories also greatly influence the quality of metadata. For instance, harvesting through the OAI-PMH aggregates together metadata from a variety of disparate resources. Yet successful harvesting requires syntactically correct metadata element usage. In 2004, the Canadian Association of Research Libraries concluded an analysis of metadata records with the following: "Metadata inconsistency and incompleteness are presenting a significant challenge to the effective harvesting and searching of institutional repository records" (Jordan & Shearer p.2). Moreover, harvesting tools like OAIster have exposed the vast quality problems found in metadata. Spelling mistakes, inconsistent data entry on author and title fields, malformed subject descriptions, and non-standardized date formats are just a few of the problems exposed through metadata harvesting (see Barton, Currier & Hey 2003; Ward 2003).

The lack of a standard metadata application profile or namespace for institutional repositories complicates matters; metadata elements can be drawn from a variety of metadata schemas with different conformance standards and the potential for semantic context loss. For example, Repository A may use the Dublin Core element "format" to describe the dimensions of an item. On the other hand, Repository B may be using the Library of Congress' Metadata Object Description Schema (MODS) metadata schema for their descriptive metadata. With a higher number of descriptive elements, the MODS schema typically allows for a more granular description items. Repository B may describe the resources of an item through the use of multiple qualifier elements like "extent", "internet media type", "form" and "digital origin".

Another complicating factor is the discrepancies in required metadata element usage across repositories. Repository A may require that all deposited journal articles have a corresponding abstract submitted by the author(s). However, Repository B may instead require that journal authors select keywords or subject headings to help describe their works. These discrepancies complicate the potential for automated quality evaluation procedures, contributing to the persistence of quality problems by failing to establish conformance standards. As Duval, Hodgins, Sutton, & Weibel (2002) have argued, "Communities of practice should be encouraged to further specify standards of practice for a given metadata standard that will encourage uniformity of descriptions within a given domain" (Mandatory Versus Optional Elements section, para. 2).

The few empirical investigations of metadata quality in institutional repositories have focused on evaluating the aggregated output of harvesters within the Open Archive community. In her examination of OAI-compliant data providers, Ward (2003) found that, on average, only 8 of 15 required Dublin Core elements (p.316). Cole and Shreeves (2004) looked at a pilot implementation of a collection and item-level metadata registry. They found an immense diversity of controlled vocabulary usage, thereby impacting the effectiveness and utility of using metadata for interoperability. They suggest that it would be useful for the OAI-PMH to help define a "quality, shareable metadata" that enumerates what attributes or metrics of quality produce truly interoperable metadata. The authors conclude that while aggregation serves a useful function, more guidance is needed on how best to optimize metadata. Efron (2007) examined the degree to which OAI-compliant institutional repositories made use of the standard 15 Dublin Core elements. He sampled harvested data from 23 repositories and found an average number

of 18.28 Dublin Core elements occurring per record. “Date” and “identifier” were the two most frequently occurring elements, followed by “title,” “language,” and “format.” The terms “contributor,” “rights,” and “coverage” were used the least and “source” was not utilized at all.

While this research highlights the importance of interoperability for repositories, it obscures the need to evaluate the context and required functionalities of metadata. It also does little to establish measures for analyzing good metadata quality. As Rothenberg (1996) writes, “The appropriateness of using a database for some purpose cannot even be defined--let alone evaluated--until that purpose is specified. For these reasons, it is important to focus on the evaluation and assessment of data quality, in addition to its improvement” (Evaluating data quality in order to improve it, para. 2)

2.3 Conceptual frameworks for metadata quality

Approaches to metadata quality evaluation frequently draw upon the vast body of data quality literature that exists. This connection is legitimized by Orr, whose sixth rule of data quality states that “laws of data quality apply equally to data and metadata” (p. 68). Given this relationship, and the relative newness of metadata creation processes, it is appropriate to briefly review Rothenberg’s discussion of data quality in order to conceptualize frameworks for metadata quality evaluation.

Rothenberg (1996) proposes that the concept of data should be understood as an abstracted model of reality. The representation choices of how to model that data efficiently inevitably impacts its evaluation. Data quality can be defined as “a measure of the suitability of data for its intended purpose (or range of purposes)” (section 1,

overview). Data should be evaluated in the *context* of its intended use; the results of these evaluations should then be utilized to make improvements to quality. Data quality assurance has tended to focus on validation of output ("the best data possible") rather than on performing explicit evaluations of purpose and intent ("how good is it?"). Thus, a first step in evaluating data quality should be to specify the functionality it needs to provide and the tasks it needs to support.

Rothenberg lists categories of what he terms "metadata" at three distinct levels: the database level, the data element (or data-dictionary) level, and the data value level. Each level contains a number of requirements that can be used to measure quality. At the database level, for example, metadata should capture "description" and "meaning" of the database, as well as its "intended use/range of purposes" and "constraints." Cumulatively, they can be seen to comprise an essential foundation for evaluating metadata quality.

Rothenberg's metadata categories should be understood as conceptual requirements, rather than the more functionally-oriented elements found in standardized metadata schemas. For instance, Rothenberg states that at the data element level, the attributes of "resolution, precision, and intended/expected accuracy" should be expressed. Still, the ideas that persist underneath these term structures can be mapped to specific schema elements. Moreover, Rothenberg's conceptual framework can be effectively transposed onto the standard metadata application profile, which is made up of the metadata schema, metadata element, and element value.

While there has been no official consensus on conceptual and operational definitions of metadata quality, a variety of frameworks have been explored and tested. Stvilia and Gasser (2008) propose a value-based assessment of metadata quality, a baseline model that evaluates the value of metadata in the context of particular activities. Established models can be used to contextualize and identify metadata requirements for successful activity completion. For example, the authors mention the FRBR bibliographic model, which outlines specific discovery tasks of "find, identify, select, and obtain" (p.12). They stress two different approaches, analytical and empirical, which can be used to estimate levels of quality for designated community. The former conceptualizes metadata requirements for particular activities, whereas the latter "helps to infer the actual or active model for quality of a particular data provider or end-user" (p.16).

To test their model, the authors used aggregated, unqualified Dublin Core metadata records harvested by the IMLS Digital Collections and Content project. Their sample consisted of approximately 150,000 objects collected from more than 20 different data providers. Their findings indicate a difference in Dublin Core element usage, based on provider, provider type, and object type variables. The total number of distinct elements used was much higher for academic libraries (21) than for public libraries (14) and museums (17). While Dublin Core best practice guidelines suggest a minimum of eight distinct elements, the authors suggest that perhaps that number be inflated to eleven. They conclude that different types of data providers will use different baseline quality requirements. (p. 72).

Barton, Currier & Hey (2003) advocate separating out the concepts of structure and content in quality evaluations, with an emphasis on the latter. They describe a number of common data quality issues which occur in metadata records, including spelling mistakes, inconsistent data entry on author and title fields, malformed subject descriptions, and non-standardized date formats. Though the authors primarily conduct a review of current literature, their stated intention is to "stimulate debate in the area of quality assurance for metadata creation across a range of communities of practice" (p. 8). Such debates can prove essential in prompting further research.

2.4 Metadata quality indicators

Moen, Stewart & McClure (1998) analyzed the metadata record content at GILS, a US Government-produced network information service. They give two levels of quality assessment measures: compliance with document requirements and utility/appropriateness of elements in supporting the intent(s) of the user(s). Based on a review of the literature, researchers identified a set of 23 criteria for assessing metadata quality, including accuracy, comprehensiveness, content, consistency, timeliness, and usability, among others. Using both quantitative and qualitative content analysis techniques, researchers evaluated approximately 3,500 metadata records. They narrow down to 3 general criteria effective for measuring quality: "accuracy," "completeness," and "delineation of information resource type" (pp. 249-254).

Bruce and Hillmann (2004) cite the risk in assessing quality by enumerating "defects." They argue that research focusing on syntactical errors sacrifices an "organized view of the forest to an overly-specific appreciation of the trees" (p. 2). They have

identified seven characteristics of good quality metadata, based in part on the Quality Assurance Framework for statistical data, developed by Statistics Canada and adapted for metadata quality analysis by Paul Johnais (2002). Metadata should have *completeness*; elements in the target schema should be utilized fully for description. It should be described with *accuracy*, both syntactically and figuratively. Its *provenance* should be disclosed, including creation information as well as any transformations undergone. It should *conform to expectations*; chosen element sets and application profiles should both support and reflect community needs and user requirements. Metadata should have *logical consistency and coherence*, enabling perceptions to conform to established standards or definitions. It should have *timeliness*, with attention paid to both the currency of the description and any associated lagging. Finally, metadata must have *accessibility,* in both the physical and intellectual sense. If metadata is disassociated with the object it is describing, it lacks physical accessibility. Similarly, an object described in alien terms to its user community can be said to lack intellectual accessibility.

The authors go on to construct a three-tiered approach to determining metadata quality through automated means. They reason that automatic metadata validation techniques were chosen because they were the most cost-effective. The first tier quality indicators consist of automatic validation of XML schemas and declared namespaces. The second tier indicators are the presence of controlled vocabularies and the population of both discovery-oriented and community-tailored elements. Finally, the third-tier indicators can include an application profile that conforms to a metadata standard or the full provenance information.

Shreves, Riley, and Milewicz (2006) characterize quality metadata as shareable metadata qualified for exchange with other distributed systems. Shareable metadata should possess what the authors refer to as the "six C's." First, metadata content should be optimized for sharing, describing the resource sufficiently enough for intended usage. Second, metadata records should be consistent in both their presence and absence. For example, if a field is missing consistently across all records, an aggregator is able to effectively ignore that field in display and search. Third, shareable metadata records are coherent. That is, users should be able to interpret them at first-glance. Values should appear in appropriate fields and elements should not be repeated. Fourth, shareable metadata should have context. In other words, metadata should be able to be understood regardless of the domain or local context it was created for. The authors recommend the inclusion of collection-level information when possible to augment meaning. Fifth, shareable metadata records rely on the establishment of communication between service providers and data providers. For instance, data providers can disclose what content standards or controlled vocabularies were used in record creation. Finally, shareable metadata records must conform to recognized standards. Without conformance, records are at risk of not being aggregated by data harvesters.

2.5 Metadata quality evaluation techniques

Empirical studies on metadata quality evaluation techniques have been scarce. As the literature shows, research has primarily focused on identifying the criteria for evaluating metadata quality. However, there are a few exceptions to this rule. Nichols et al. (2008) compare and contrast two metadata analysis tools in use at New Zealand libraries. They interview repository managers about their experiences with both tools and

make recommendations for the development of future metadata quality tools. Hughes (2004) reported on the construction of an infrastructure to support metadata quality assessment within a specific domain, the Open Language Archives Community (OLAC), a consortium of linguistic data archives. The author recommends examining both individual metadata records and collection-level metadata records, against "a baseline of broader community practice, as well as for compliance to external standards" (p.320).

In their survey of current metadata practices in digital repositories, Park and Lu (2008) discovered many recurring problems that underscore the challenges of good quality metadata creation. The authors examined 659 metadata item records from digitized image collections from three digital repositories using Dublin Core metadata schemas. Overall, the authors found two main issues affecting metadata quality that fall in line with other empirical research. First, there was inconsistent or inaccurate usage of elements. For example, physical descriptions of items were mapped to both the Dublin Core "description" and "format" elements. Moreover, there was confusion surrounding the presumed correct use of the Dublin Core "type" and "format" elements. Further clarification was also needed on the appropriate usage of the three Dublin Core elements: "creator", "contributor" and "publisher". Second, they noted the frequency of null values in provenance-oriented metadata elements, such as contact or acquisition information.

Dushay & Hillmann (2003) discussed techniques used for large-scale metadata evaluation of bulk data the National Science Digital Library. Both the metadata registry and OAI-harvested data gathered by the authors exposed a significant amount of metadata quality problems which the authors bucketed into broad categories. Their first

problem category was an absence of data from critical Dublin Core elements such as "format" and "type"; for repositories seeking to establish search capabilities based on resource type or format, these null values prove to be a central concern. Secondly, they found a number of examples where incorrect data was inputted. For example, they found that creator names were often repeated in the language element. The third category of data quality problems were confusing data, likely the result of merged databases or inaccurately placed HTML tags within values. The final problem category the authors delineate is insufficient data, where the OAI-PMH minimal requirements of simple Dublin Core for harvesting subsequently removed the necessary context for interpretation of metadata (p.3).

The authors go on to discuss a number of metadata evaluation processes developed and tested to address these problems. One approach was to use an XML schema interface tool like XMLSpy for both random sampling and easy display of possible errors. However, the authors found that "reviewing more than a handful of metadata records using this method was tedious at best and ultimately unsatisfactory, primarily because it provided no pattern of error, nor any convenient way of determining the extent of a discovered problem within a file" (p. 3). Next, they used Microsoft Excel for visual review of the data. They sorted by element name and then sub-sorted by values within the elements. While this approach allowed for easier detection of data errors, it was ultimately not scalable for large amounts of data. Their third evaluation technique was the use of a visual graphical analysis tool called Spotfire, which displays up to six data dimensions simultaneously. Evaluators were able to detect problem patterns in both individual elements and across collections; the tool also enables users to make changes

directly to problem data. Thus, the authors conclude that the use of data visualization software can "significantly improve efficiency and thoroughness of metadata evaluation, both before and after transformation" (p. 9).

Few digital repositories appear to make their evaluation techniques for measuring quality metadata available to the public. One exception is the California Digital Library, which has produced formal specifications for a metadata processing tool to improve quality. They disclose broad guidelines for assessing metadata quality, as well as functional requirements for evaluating metadata for these guidelines. To address the latter, the specifications require that: (1) All elements present should be listed along with a percentage of non-empty elements of each type; (2) An output list should be produced of all non-empty occurrences; (3) The number of duplicate instances of the same element should be totaled; (4) List all non-duplicative content for specific elements and the number of times the content occurs; (5) Identification of patterns across records.

Methodology

3.1 General Description

In order to explore current metadata quality practices at academic institutional repositories, a thirteen-question survey was designed and administered online using the Qualtrics software, which is available through the Odum Institute at the University of North Carolina at Chapel Hill (see Appendix A). Survey questions were adapted from previous survey research on issues of importance to institutional repository communities of practice, including overall metadata practices (Ma 2007), institutional leadership (Bailey 2006), preservation efforts (PREMIS Implementation Group 2003), and general usage statistics (McDowell 2007). Self-administered, online questionnaires were chosen as a method to gather survey data as they presented the most convenient option for respondents. As Babbie (2007) advises, "Anything you can do to make the job of completing and returning the questionnaire easier will improve your study" (p. 260).

3.2 Participants

The sample population for this research study was comprised of Association of Research Libraries (ARL) members with operational institutional repositories. ARL is a not-for-profit organization comprised of 123 member libraries from research institutions in North America. This subject population was chosen for two main reasons. First, they could be easily identified through institutional lists and/or membership. Secondly,

previous research studies on metadata practices in IRs had looked specifically at ARL-member institutions (Ha 2007). It was anticipated that this research could offer an updated, state-of-the-art assessment of current metadata practices in IRs which could be compared and contrasted with previous studies.

The subject population recruited for this research was staff members who were most familiar with the current metadata practices at their respective IRs. Specifically, this study sought respondents who could provide information about their institution's creation and management of metadata, current metadata schema(s), guidelines and quality control procedures, and overall impression of metadata quality. It was anticipated that the results from this sample population could be quantified to reflect the "trends, attitudes or opinions of a population" (Creswell 2003, p.12). It was estimated that there would be between 30 and 45 subjects participating.

3.3 Procedures

Survey participants were recruited using purposive sampling techniques with two primary approaches. The first approach consisted of a recruitment email sent to the Scholarly Publishing and Academic Resources Coalition (SPARC) Institutional Repositories Discussion List. SPARC is an alliance of research and academic libraries, with a coalition of over 200 North American members under the branch of the Association of Research Libraries.

A second, concurrent approach consisted of contacting individual repository contacts collected from three aggregated repository directories. The first was the OpenDOAR website, which lists academic open-access repositories. The second was the

University of Illinois OAI Registry site, which compiles all OAI-compliant repositories. The third was the Registry of Open Access Repositories (ROAR), which lists registered users of institutional ePrint software. All three sites were manually scanned and narrowed down purposefully to limit to the appropriate sampling frame of ARL-member institutions.

After receiving Institutional Review Board (IRB) approval, the Principle Investigator send a recruitment email to the SPARC-IR list serve (see Appendix B) and to identified individual contacts at repositories (see Appendix C), asking for their participation in a survey on current metadata practices and procedures at their institutional repository. The recruitment email stated that the questionnaire would take no more than 20 minutes and that participation would be completely voluntary and anonymous. A final, reminder email was sent one week later. The survey closed one month after the initial recruitment email was sent. No compensation was provided for participation in the survey.

3.4 Data Analysis

Descriptive statistical analysis was used to quantify responses of current practices in use at institutional repositories. Quantitative data was converted into tables and figures using Qualtrics to report results in a graphical format. Univariate data was analyzed mainly for percentages and frequency distributions. Qualitative data was exported to Excel and reviewed for themes and patterns. Based on analysis, major themes were identified and then coded by frequency of terms.

Results

Fifty-five participants began this survey, with a total of forty respondents providing responses useful for data analysis. All questions were optional and the data presented here are based upon the response rates for each question.

4.1 Metadata practices within IR settings

Survey respondents were first asked to identify their job titles in a free-text box. Twenty-three respondents answered this question, with the majority of responses claiming an array of different job titles. Eleven respondents reported a position with the term "librarian" in the title, with four respondents self-identifying as metadata librarians.

Table 1: Respondents' job titles

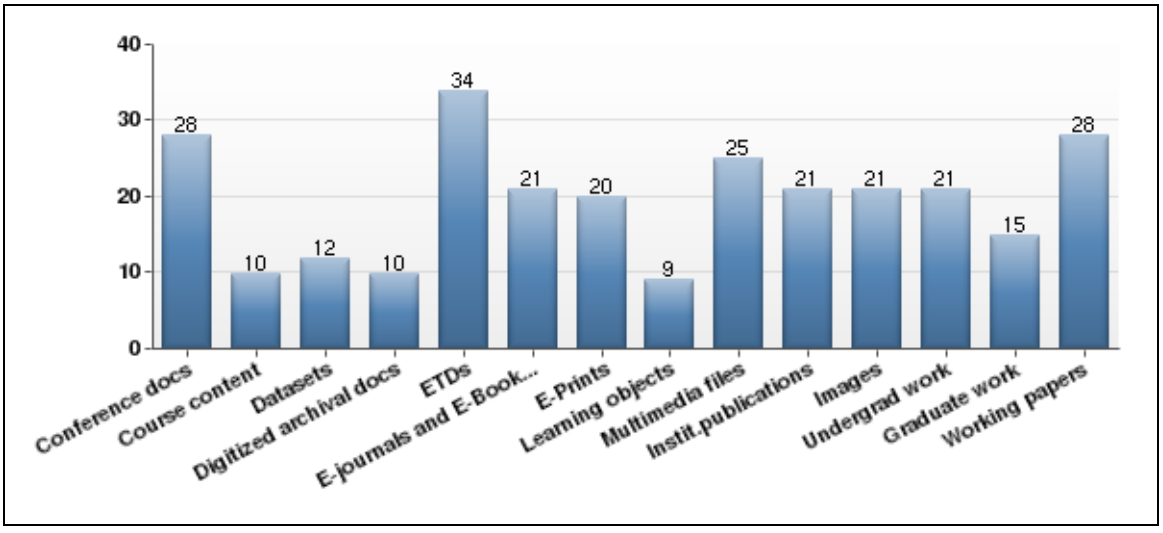
Text Response	# Respondents
Assessment & Scholarly Communications Services Coordinator	1
Collection Development/E-Resources Librarian	1
Digital Collections Librarian Head, Metadata Services	1
Digital Initiatives Librarian	1
Digital Library Specialist	1
Digital Repository Coordinator	3
Digital Repository Librarian	1
Director, Digital Library and Archives	1
Electronic Acquisitions	1
Head, Technical Services Dept.	1
Head of Cataloging	1
Head of Digital Services and Scholarly Communication	1

Text Response	#Respondents
Institutional Repository Manager	1
Metadata Librarian	4
Scholarly Communication Librarian	1
Senior Associate University Librarian	1
Serials Team Leader & Scholars' Bank Coordinator	1
Technology and Metadata Librarian	1

4.2 Institutional holdings

In an effort to assess how repository content is managed, survey participants were asked to disclose their institutional holdings. Respondents could select multiple types of content. Thirty-eight institutions responded to this question, with the majority (34, or 89%) reporting that their institution held electronic dissertations or thesis. Working papers and technical reports tied with conference proceedings and presentations (held by 28 institutions, or 83%) as the second most common digital object(s).

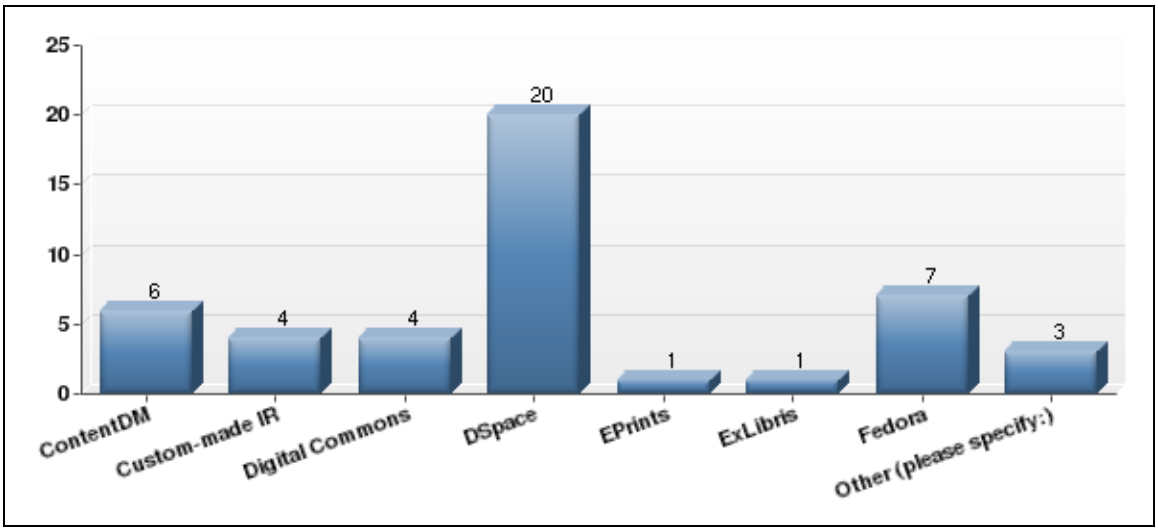
Figure 1: Institutional holdings



4.3 Software platforms

A total of thirty-eight institutions responded to a survey question about their software platform(s) in use, checking all that apply. Twenty respondents (or 54%) reported using DSpace, while ContentDM and Fedora were mentioned by 6 respondents, respectively (16%). Most institutions were only using one software platform rather than a combination of applications.

Figure 2: Software platforms

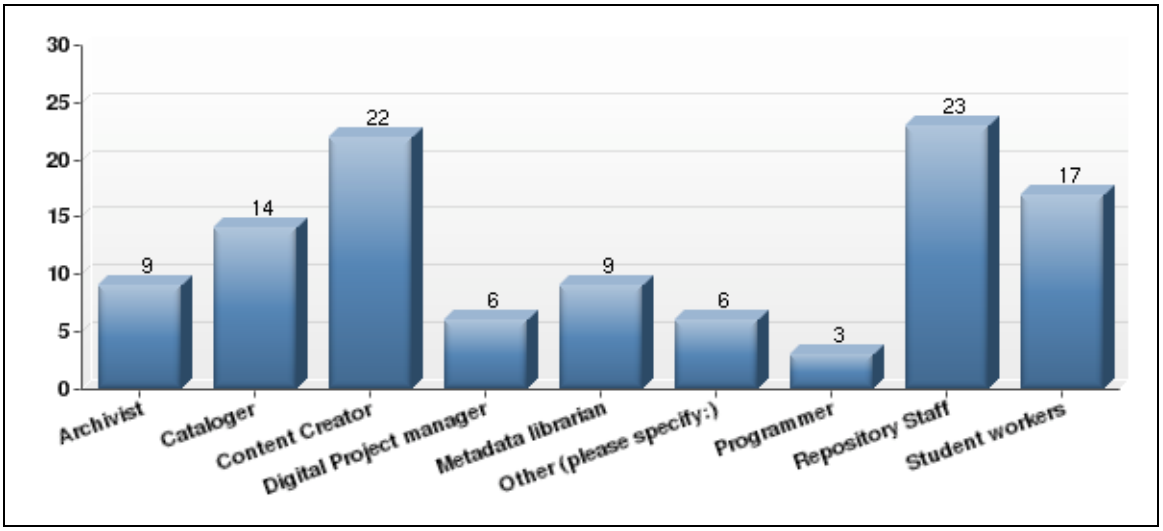


4.4 Metadata management

Thirty-six institutions responded to a question about how metadata was created for deposited objects in their respective IR. Again, respondents were allowed to select as many options as were applicable. The highest number of institutions (23, or 64%) reported that metadata was created by repository staff. The second most frequent metadata creator reported by institutions was the content creator of the submission (22, or 61%). There was significant overlap between these categories as well, with fifteen

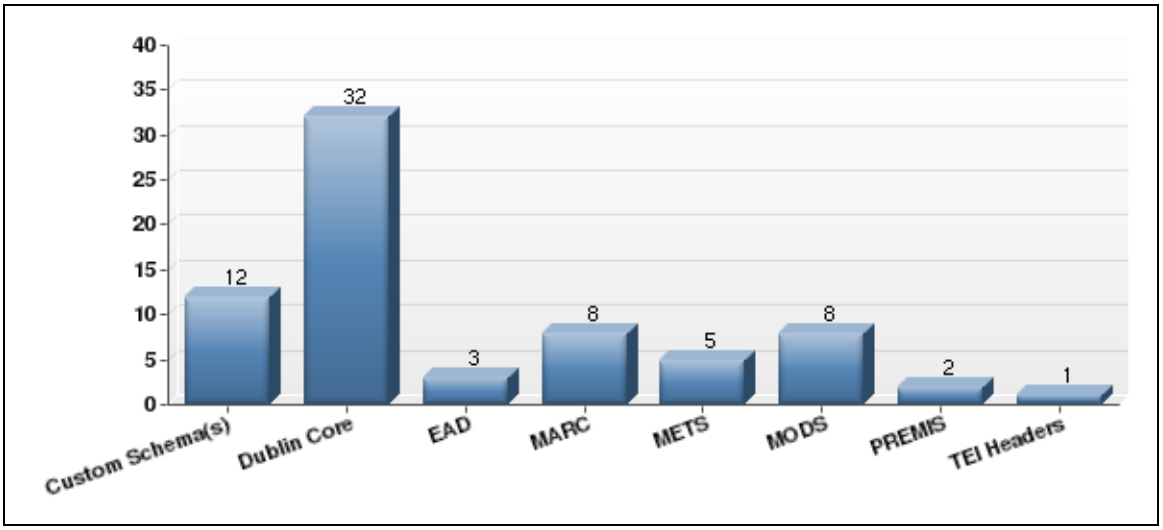
institutions reporting overall that metadata was created by both repository staff and content creators. In fact, most institutions reported that more than one entity was creating metadata for submissions. For example, out of those institutions that attributed metadata creation to a cataloger, at least half of those also reported the participation of an archivist, content creator, or student worker.

Figure 3: Metadata creator(s)



Thirty-seven respondents reported on the number and types of metadata schemas being utilized by their institutions. The most frequently reported schema was Dublin Core, with thirty-two institutions (86%) claiming use. Twelve institutions (or 32%) reporting using custom schemas. Since institutions could select multiple schemas, it is interesting to note that thirteen of the institutions using Dublin Core were using at least one additional schema; ten of these were custom schemas.

Figure 4: Metadata schema(s) usage



Twenty-seven institutions responded to a question about which controlled vocabularies they were using. Library of Congress Subject Headings (LCSH) were the most commonly used controlled vocabulary. As expected, many of those using LCSH were using another bibliographic classification tool, the Library of Congress Name Authority File (LCNAF). Most institutions that were using controlled vocabularies reported using at least two combinations. For example, all of the seven institutions using the Art and Architecture Thesaurus (AAT) also used LCSH and LCNAF.

Table 2: Controlled vocabularies

#	Answer	Response	%
1	Art and Architecture Thesaurus (AAT)	8	30%
2	Getty Thesaurus of Geographic Names	7	26%
3	Getty Union List of Artist Names (ULAN)	4	15%
6	Library of Congress Name Authority File (LCNAF)	13	48%
5	Library of Congress Subject Headings (LCSH)	20	74%
4	Library of Congress Thesaurus for Graphical Materials	7	26%
7	MARC relator codes	8	30%
8	Other (please specify:)	8	30%

Thirty-six respondents answered a question regarding quality control procedures at their institution. Respondents could select as many choices as were applicable; a total of eight (or 22%) reported that no quality control procedures were currently in place. Sixteen respondents (or 44%) reported that metadata are manually checked and approved before publishing, with half of those respondents indicating that metadata are also checked by librarians, catalogers, or other staff. Four institutions (or 11%) reported use of tool to check metadata consistency and accuracy.

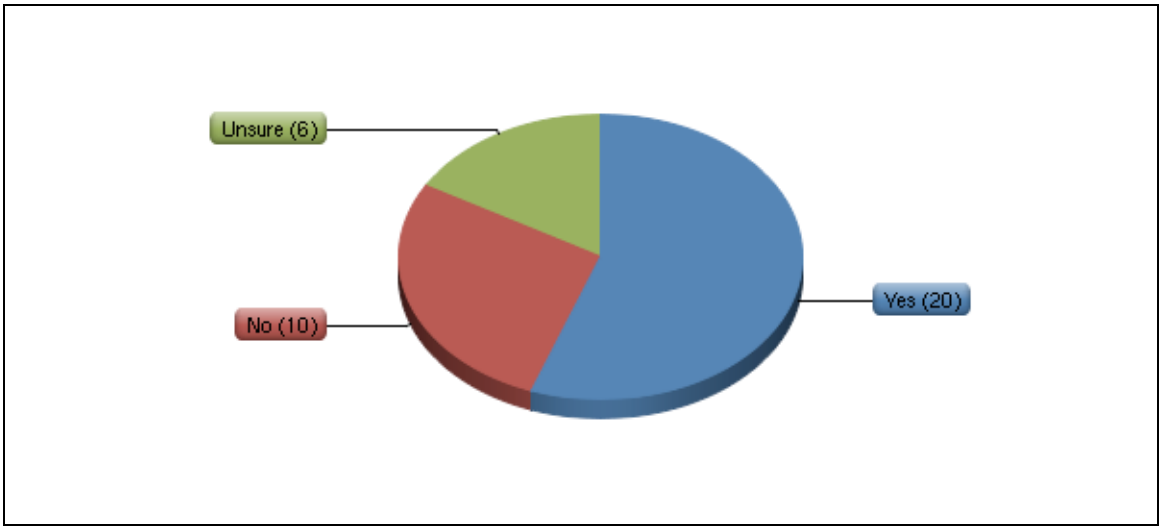
Table 3: Metadata quality control

#	Answer	Response	%
1	Metadata are manually checked and approved before publishing.	16	44%
2	Metadata created by users or content creators are checked and approved by metadata librarians, catalogers, or other staff.	16	44%
3	A tool is used to check metadata consistency and accuracy.	4	11%
4	No quality procedures are currently in place.	8	22%
5	Other (please specify:)	9	25%

4.5 Metadata documentation

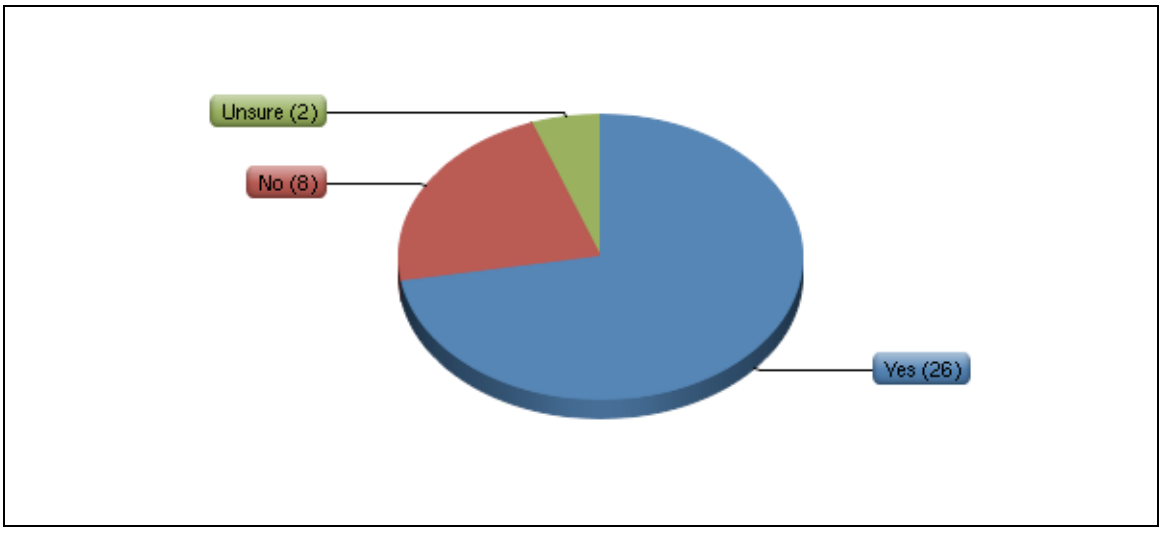
Thirty-six institutions responded to three survey questions seeking to explore the degree to which specific policies, guidelines, and documentation were in use at institutional repositories. Question eight asked whether internal guidelines on metadata quality existed at institutional repositories. Twenty institutions (55%) reported that their repository maintained some form of documentation, while ten (28%) reported they did not.

Figure 5: Internal documentation on metadata quality



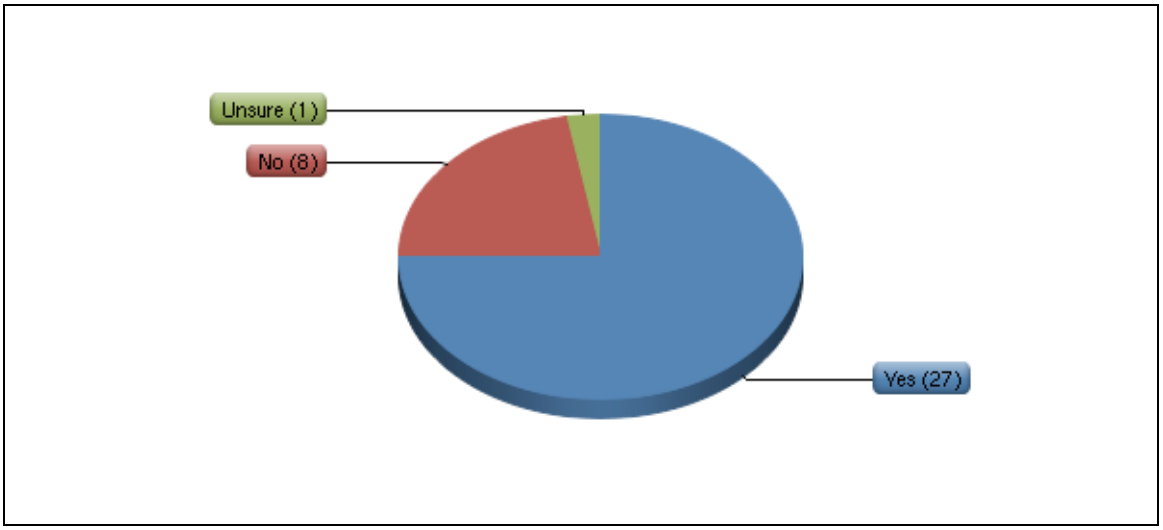
Question nine asked about whether repositories provided metadata guidelines for depositors. The majority (26, or 72%) responded that guidelines were provided.

Figure 6: Metadata guidelines for depositors



Question ten asked whether institutions had policies on minimum metadata requirements. Again, the majority (27, or 75%) of respondents answered that policies were in place.

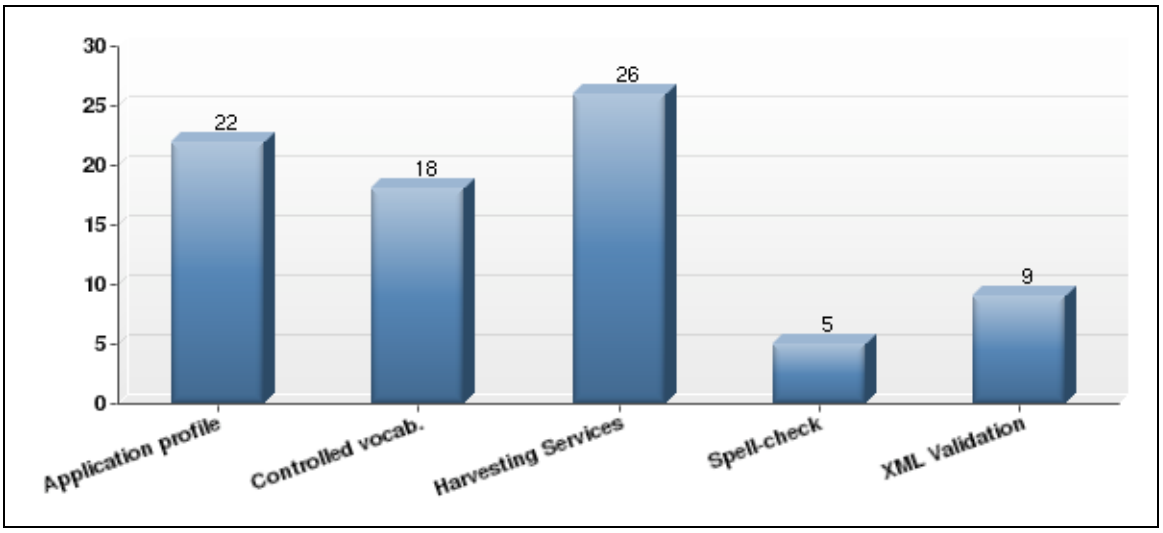
Figure 7: Minimum metadata requirements



4.6 Repository infrastructure

Thirty-two institutions responded to a question about infrastructure services currently in use at their repository, selecting all applicable answers. The majority of respondents (26, or 81%) used harvesting services such as the OAI-PMH. Respondents also reporting using application profiles (22, or 69%), controlled vocabularies (18, or 56%), XML validation (9, or 28%), and spell-check services (5, or 16%).

Figure 8: Repository infrastructure services

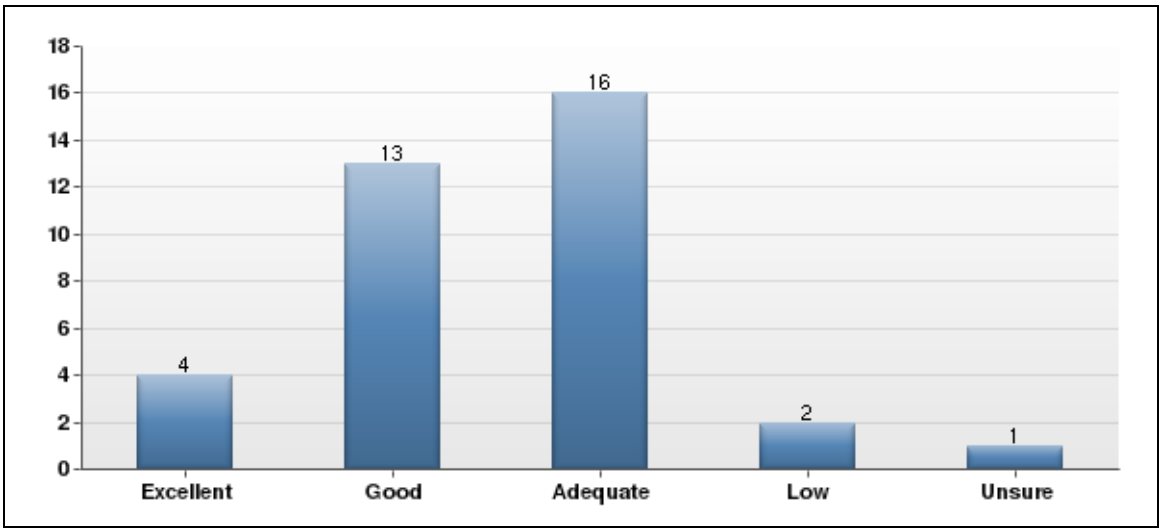


4.7 Metadata quality

Respondents were asked to describe how metadata quality problems were typically discovered in their repository. Thirty respondents provided write-in text detailing the process. In general, respondents seem to indicate that quality control is done on a reactive basis after deposit. As an institution noted, "If someone reports a problem, we correct it." Thirteen institutions mentioned that repository staff typically discovered errors, while eleven institutions noted that users or content creators found them. One respondent wrote, "All QC measures on the IR are intermittent at best! Some content providers are more conscientious than others." Two institutions mentioned that errors were discovered during batch upload fails. Only one institution is doing active data cleanup by exporting data. A few institutions reported that quality checking was done on an ad-hoc basis due to budgetary concerns, with certain collections receiving prioritized attention.

Question thirteen asked respondents for their overall impression of metadata quality in their repository. Thirty-six respondents gave their impression, with the highest number of institutions (16) describing their metadata quality as "Adequate." Thirteen institutions claimed their metadata was good, while four institutions claimed it was excellent.

Figure 9: Overall impression of metadata quality



Discussion

The results presented in the previous section provide a view of practices and procedures relating to metadata quality in institutional repositories at this time. They show that there is no set practice for quality control, and that many institutions have not had the opportunity to integrate metadata quality measures. In this section, further discussion of the results is covered.

The variance shown in job titles by survey respondents suggests that institutions may not have a unified vision of the required roles within an operational IR. For example, only four respondents held the same position title ("Metadata librarian"). While the majority of respondents self-identified themselves as "librarians" of some type, it is not clear from these results whether metadata activities are considered a primary responsibility for any respondents. The variance in job titles echoes previous survey research on IRs which has found similar discrepancies in respondents' positions. The MIRACLE Project's 2007 Census of Institutional Repositories in the United States recruited participation from individuals "most familiar with their institution's involvement with IRs;" yet respondents had at least six different self-identified position titles with nearly three-quarters of respondents (74%) reporting they were library directors (Markey, Rieh, St. Jean, Kim & Yakel, p.14). Ma (2007) found a similar distribution in her survey of ARL-member institutions; "metadata architect", "digital

content librarian", and "electronic resources librarian" were among the different positions mentioned as having responsibility for metadata activities (p. 13).

Institutional repositories appear to hold a wide range of digitized and born-digital materials, from electronic theses and dissertations to datasets. While the types of file formats in use were not explored in this study, the existence of multiple content types (e.g., text, images) almost certainly factors into the ability to effectively evaluate metadata quality. The absence of defined quality characteristics for differing content types combined with differing needs and requirements has been problematic. Moreover, there has been little consensus on determining what attributes should be preserved for specific content types. Indeed, previous research has suggested that metadata quality should be evaluated in terms of the functionality of its required use (Moen, Stewart & McClure, 1998; Stvilia and Gasser, 2008). As Stvilia, Gasser, Twidale, Shreeves, & Cole (2004) write, "Specific metadata quality problems arise when the existing quality level along some particular metadata dimension is lower than the required quality level, in the context of using this metadata to support a given activity" (p. 114).

This study also reported on the technical infrastructures in place at institutional repositories. The results indicate that institutions are typically using only one software platform rather than a combination of applications. Dspace is the most frequently used software (20 institutions), with Fedora being mentioned second (seven institutions). This finding offers promise for the possibility of establishing automated quality control measures, which must integrate with individual platforms in order to be effective.

The results of this study also found that IRs are typically using the Dublin Core to describe and manage their holdings. This makes sense considering that Dublin Core was one of the earliest metadata standards to emerge for digital repositories. Moreover, as noted above the majority of survey respondents are using Dspace, which has adopted a Dublin Core metadata schema. The small number of institutions using metadata schemas like MODS and PREMIS could be attributed to either their relative newness or the perceived complexity to implement them. Further work on the obstacles and barriers to use should be undertaken.

As the literature review demonstrated, there is a wide-range of quality problems found in metadata records. The process by which metadata is created, and who creates is, almost certainly has implications for its quality. The majority of institutions reported that metadata is created by either repository staff (23 institutions) or content creators (22 institutions). However, there is significant overlap among categories and most institutions attribute metadata creation to more than one entity. For example, only five out of the 23 institutions report that metadata creation is solely done by repository staff. Similarly, out of the 14 institutions that attributed metadata creation to a cataloger, at least half also reported the participation of an archivist, content creator, or student worker. The implication is that multiple people in different roles within an IR setting are creating metadata for deposited objects. This may be the result of organizational challenges or simply the well-intentioned effort of collaborators. Yet such endeavors would undoubtedly be problematic for institutions without standardized metadata guidelines in place.

The impact that quality-enforcing structures like controlled vocabularies and application profiles might have on metadata has yet to be fully explored; however, the results of this study offer a starting point for documenting the number of institutions utilizing them. Most respondents reported using some type of controlled vocabulary, with the highest number of institutions citing use of the Library of Congress Subject Headings. It appears that most institutions that are using controlled vocabularies tend to use more than one. Similarly, application profiles, which mix and match multiple metadata elements from various schemas, are perceived as beneficial structures that will positively impact metadata quality. As Heery and Patel (2000) argue in their influential paper, application profiles arose out of practical need within the community for greater support and management of digital objects. 22 institutions report using application profiles, with 16 of those claiming the use of controlled vocabularies as well.

An essential part of ensuring quality metadata for institutional repositories involves developing policies which outline metadata quality requirements and evaluation procedures. The results of this study indicate that institutional repositories are developing metadata policies for contributors and staff. For example, the majority of respondents (72%) reported that their institution provided metadata guidelines for depositors. Similarly, the majority of respondents (75%) reported that their institution had policies on minimum metadata requirements. The weakest area of documentation appears to be internal metadata guidelines. 20 institutions (or 55%) reported that their repository maintained some form of documentation.

Though most quality-checking of metadata fields remains a largely manual endeavor, the process by which institutions are finding problems with metadata seems to

vary. Institutions seem to rely equally on contributors and users to discover quality problems as they repository staff. As one institution explains, "They're discovered in a variety of way by different people. Users or collection owners may report them. One of the librarians may find them by chance when looking for something in our repository." Six institutions mentioned that metadata were manually reviewed by repository staff prior to publication.

There was some suggestion that certain metadata fields received higher priority for quality checking. When institutions made mention of specific metadata fields that received attention, the focus was on author or subject entries. One institution wrote, "We generally are not exerting a lot of control over metadata, but where we do try to look for issues is author's names; we want to enforce authority control when we can." Another institution reported that "different collections or types of material get differing levels of metadata checking." For example, this institution mentioned that while user-submitted data for ETDs is not screened, "extensive checking" is done for other types of materials.

It was interesting to note correlations between an institution's impression of metadata quality and the existence of quality-enforcing structures like controlled vocabularies at institutions. Out of the institutions that described their metadata as "good", the majority had a policy on minimum metadata guidelines (92%). Similarly, out of the twenty institutions that had internal documentation on quality, most (85%) also provided metadata guidelines to depositors. The implication is that *when* metadata documentation is developed, it tends to be created for both content creators and providers.

Conclusion

Though poor data quality inevitably impacts the usability and effectiveness of digital resources, thus far little empirical investigation has taken place into the current metadata quality practices that exist at institutional repositories. There has been little interest directed at the development of successful evaluation techniques and procedures. This study explored this knowledge gap by surveying current metadata practices and exploring quality control procedures, both of which could potentially be used to develop an evaluation framework for assessing metadata quality.

A literature review was performed to surface conceptual themes as well as possible quality indicators useful for evaluation. The complex domain of institutional repositories was discussed, including metadata quality issues that impact usage. Finally, practitioners were surveyed on applications of use in an attempt to (1) summarize current metadata practices and (2) explore any quality control procedures being used.

The results of this study suggest that metadata activities may not yet be streamlined into an institution's workflow and organizational structure. The number of varying job positions responding to this survey demonstrate that a range of positions with differing degrees of expertise are responsible for metadata practices. The lack of formalized quality control and procedures seem to indicate that metadata quality is an after-thought for most institutions. Metadata quality problems tend to be found on an ad-hoc basis, with users contributing as much to the discovery process as repository staff.

When metadata quality is checked, author and subject fields receive the most attention. The majority of institutional repositories are maintaining some form of documentation, either in the form of metadata guidelines for contributors or internal documentation for staff.

As institutional repositories continue to grow, distributed access and sharing of resources among contributors will increase. In this environment, the need to evaluate metadata quality in a meaningful, scalable manner will become even more critical. The establishment of shareable, good quality metadata is integral to the long-term health and sustainability of institutional repositories. For institutional repositories, the development of an application profile could go a long way towards formalizing community needs and modeling tasks and activities to evaluate.

A number of key challenges influence the development and successful deployment of metadata quality evaluation techniques. In particular, determining what criteria should be used to evaluate metadata quality is essential to understanding the level to which metadata records are compliant. Moreover, metadata quality evaluation should incorporate both the context in which the metadata was created as well as the functionality it is intended to support. Metadata quality evaluation should thus be viewed as an evolving process that necessarily relies on continuous refinement through assessment. Attention must be paid not only to the successful harvesting or output of data, but also to determining how good the output is in the context of specific, required activities.

References

Babbie, E. (2007). *The practice of social research (11th ed.)*. CA: Wadsworth.

Bailey, C. W., Jr., Coombs, K., Emery, J., Mitchell, A., Morris, C., Simons, S., & Wright, R. (2006). Institutional Repositories. *SPEC Kit 292*. Washington, D.C.: Association of Research Libraries.

Balci, O. (2003). Verification, validation, and certification of modeling and simulation applications. In *Proceedings of the 35th Conference on Winter Simulation: Driving innovation,* New Orleans, LA., December 07 - 10, 2003.

Barton, J., Currier, S., & Hey, J. (2003). Building quality assurance into metadata creation: an analysis based on the Learning Objects and e-Prints communities of practice." In *Proceedings 2003 Dublin Core Conference: Supporting Communities of Discourse and Practice - Metadata Research and Applications*. Seattle, WA., 2003.

Barton, J. & Robertson, R.J. (2005). Designing workflows for quality assured metadata. *CETIS Metadata & Digital Repositories SIG Meeting,* Edinburgh, March 10 2005.

California Digital Library. CDL guidelines for digital objects. Retrieved from: http://www.cdlib.org/inside/diglib/guidelines/

Cole, T. & Shreeves, S. (2004). Search and discovery across collections: The IMLS digital collections and content project. *Library Hi Tech, 22(3)*.

Creswell, J. W. (2003). *Research Design: Quantitative, Qualitative, and Mixed Methods Approaches*. SAGE. Thousand Oaks. USA.

Crow, R. (2002). The case for institutional repositories: A SPARC position paper." ARL Bimonthly Report 223. Retrieved from http://works.bepress.com/ir_research/7

Dushay, N. & Hillmann, D. (2003). Analyzing metadata for effective use and re-use. in *DC-2003: Proceedings of the International DCMI Metadata Conference and Workshop,* Seattle, WA., September 28-October 2 2003. Retrieved from http://dc2003.ischool.washington.edu/Archive-03/03dushay.pdf

Duval, E., Hodgins, W., Sutton, S., & Weibel, S. L. (2002). Metadata principles and

practicalities. *D-Lib Magazine*, Retrieved from http://www.dlib.org/dlib/april02/weibel/04weibel.html

Greenberg, J. (2001). A quantitative categorical analysis of metadata elements in image-applicable metadatas schemas. *Journal of the American Society for Information Science*, *52(11)*.

Heery, R., & Patel, M. (2000). Application profiles: mixing and matching metadata schemas. *Ariadne, 25.* Retrieved from http://www.ariadne.ac.uk/issue25/app-profiles/intro.html.

Hitchcock, S., Brody, T., Hey, J., Carr, L. (2007). Digital preservation service models for institutional repositories. *D-Lib Magazine.* Retrieved from http://www.dlib.org/dlib/may07/hitchcock/05hitchcock.html.

Hitchcock, S., Brody, T., Hey, J., Carr, L. (2007). Survey of repository preservation policy and activity. Preserv project, January 2007. Retrieved from http://preserv.eprints.org/papers/survey/survey-results.html.

Hodge, G. (2005). Metadata for electronic information resources: from variety to interoperability. *Information Services and Use, 25(1)*.

Hughes, B. (2004). Metadata quality evaluation: Experience from the Open Language Archives Community. *Digital Libraries: International Collaboration and Cross-Fertilization. 7th International Conference on Asian Digital Libraries, ICADL 2004,* Shanghai, China, December 13-17, 2004.

Lagoze, C. (2005). Guidelines for repository implementers. Implementation guidelines for the Open Archives Initiative Protocol for Metadata Harvesting. Retrieved from http://www.openarchives.org/OAI/2.0/guidelines-repository.htm.

Lynch, C. (2003). Institutional Repositories: Essential Infrastructure for Scholarship in the Digital Age. *ARL Bimonthly Report.* Retrieved from http://www.arl.org/resources/pubs/br/br226/br226ir.shtml

Lynch, C. & Lippincott, J. (2005). Institutional repository deployment in the United States as of early 2005. *D-Lib Magazine.* Retrieved from http://webdoc.sub.gwdg.de/edoc/aw/d-lib/dlib/september05/lynch/09lynch.html.

Ma, J. (2007). *Metadata*. Washington, D.C.: Association of Research Libraries.

Markey, K., Rieh, S. Y., St. Jean, B., Kim, J., & Yakel, E. (2007). Census of institutional repositories in the United States: *MIRACLE Project Research Findings*. Washington, D.C.: Council on Library and Information Resources. Retrieved from http://www.clir.org/pubs/reports/pub140/pub140.pdf.

McCord, A. (2003). Institutional repositories: Enhancing teaching, learning, and research.

EDUCAUSE Evolving Technologies Committee White Paper. Retrieved from http://net.educause.edu/ir/library/pdf/DEC0303.pdf.

McDowell, C. (2007). Evaluating institutional repository deployment in American academe since early 2005. *D-Lib Magazine.* Retrieved from http://www.dlib.org/dlib/september07/mcdowell/09mcdowell.html.

Moen, W.E., Stewart, E.L., & McClure, C.R. (1998). Assessing metadata quality: Findings and methodological considerations from an evaluation of the US government information locator service. *ADL '98: Proceedings of the Advances in Digital Libraries Conference,IEEE Computer Society,* Washington, DC, 1998.

Nichols, D.M., Paynter, G.W., Chan, C., Bainbridge, D., McKay, D., Twidale, M.B. & Blandford, A. (2008). Metadata tools for institutional repositories. (Working paper 10/2008). Hamilton, New Zealand: University of Waikato, Department of Computer Science.

Ochoa, X. & Duval, E. (2006). Towards automatic evaluation of learning object metadata quality. *Advances in Conceptual Modeling - Theory and Practice, ER 2006 Workshops BP-UML, CoMoGIS, COSS, ECDM, OIS, QoIS, SemWAT*, Springer, pp. 372-381.

Orr, K. (1998). Data quality and systems theory. *Communications of the ACM , 41(2)*. Retrieved from http://doi.acm.org.libproxy.lib.unc.edu/10.1145/269012.269023

Rothenberg, J. (1996). Metadata to support data quality and longevity. *Proceedings of the 1st IEEE Metadata Conference,* NOAA Complex, Silver Spring, MD., April 16-18 1996. Retrieved from http://web.archive.org/web/20000708182256/http://www.computer.org/conferences/meta96/rothenberg_paper/ieee.data-quality.html

Stvilia, B & Gasser, L. (2008). Value-based metadata quality assessment. *Library and Information Science Research, 3(1).*

Stvilia, B & Gasser, L., Twidale, M., Shreeves, S., Cole, T. (2004). Metadata quality for federated collections *Proceedings of the 9th International Conference on Information Quality,* Cambridge, MA. Retrieved from https://www.ideals.uiuc.edu/handle/2142/721

Tennant, R. (2001). Different paths to interoperability. *Library Journal*, *126(3)*.

Ward, J. (2003). A quantitative analysis of unqualified Dublin Core metadata element set usage within data providers registered with the Open Archives Initiative. *Joint Conference on Digital Libraries*, Houston,TX., 2003.

Appendix A: Survey

Metadata Quality Evaluation Practices in Institutional Repositories

1. Please provide your job title below.

2. Please describe the holdings of your institutional repository.

Check all that apply

- o Conference proceedings and presentations
- o Course content (syllabi, assignments, lectures)
- o Datasets
- o Digitized archival documents and university records (historical texts and primary sources)
- o Electronic theses and dissertations
- o E-journals and E-Books
- o E-Prints
- o Learning objects
- o Multimedia files (digital audio/video)
- o Non-scholarly institutional publications
- o Pictures (images)
- o Undergraduate student work
- o Graduate student work (non-ETD)
- o Working papers and technical reports

3. What software platform do you use?

Check all that apply

- o ContentDM
- o Custom-made IR
- o Digital Commons
- o DSpace
- o EPrints

- ExLibris
- Fedora
- Other (please specify):

4. Who is responsible for creating metadata for objects deposited into your institutional repository?

Check all that apply

- Cataloger
- Archivist
- Repository Staff/Manager
- Content Creator
- Metadata librarian/specialist
- Digital project manager
- Student workers
- Programmer
- Other (please specify):

5. What metadata schema has your institutional repository adopted?

Check all that apply

- Custom Schema(s)
- Dublin Core
- EAD (Encoded Archival Description)
- TEI Headers (Text Encoding Initiative)
- MARC (Machine Readable Cataloging)
- METS (Metadata Encoding and Transmission Standard)
- MODS (Metadata Object Description Schema)

- o PREMIS

6. Please indicate which of the following controlled vocabularies your institution applies to metadata.

Check all that apply

- o Art and Architecture Thesaurus
- o Getty Thesaurus of Geography
- o Getty Union List of Artists
- o Library of Congress Thesaurus
- o Library of Congress Subject Headings (LCSH)
- o Library of Congress Name Authority File (LCNAF)
- o MARC relator codes
- o UNESCO Thesaurus
- o Other (please specify):

7. How does your institution maintain quality control for metadata?

Check all that apply

- o Metadata are manually checked and approved before publishing.
- o Metadata created by users or content creators are checked and approved by metadata librarians, catalogers, or other staff.
- o A tool is used to check metadata consistency and accuracy.
- o No quality procedures are currently in place.
- o Other (please specify):

8. Does your institution maintain any internal documentation or guidelines on metadata quality?

- Yes
- No
- Unsure

9. Does your repository provide metadata guidelines for depositors?

- Yes
- No
- Unsure

10. Does your repository have a policy on minimum metadata requirements?

- Yes
- No
- Unsure

11. Does your repository use any of the following:

Please check all that apply

- Application profile or metadata schema
- Controlled vocabularies (e.g., Thesaurus of Graphic Materials)
- Spell-check on ingest or post-ingest
- XML Validation
- Harvesting Services (OAI-PMH, OAIster)

12. Please describe your impression of the overall quality of your metadata.

- Excellent
- Good
- Adequate
- Low
- Unsure

Appendix B:

Initial Recruitment Email Sent to Institutional Repository Managers/Staff

Dear Colleagues,

We would like to invite you to participate in an online survey entitled, "Metadata Quality Evaluation in Institutional Repositories: A Survey of Current Practices." This research study intends to investigate current practices of metadata quality evaluation at academic institutional repositories (IRs) in the United States.

We are interested in surveying participants from ARL-member institutions who are involved with metadata practices and operations at their respective repositories. Specifically, we are seeking participants who can provide information about their IR's creation and management of metadata, usage of metadata schema(s), guidelines and quality control procedures, and metadata policies. If you decide to be in this study, you will be one of the 123 institutions invited to participate in this research.

Participation in the survey is anonymous and voluntary. If you have any questions about the survey, please contact the Principal Investigator of this study at achass@email.unc.edu. The survey should take approximately 20 minutes to complete.

The survey, at <URL> is now open, and will remain open until <date>.

We will be sending a reminder email about the survey in one week. Your assistance in providing invaluable information about this topic is much appreciated.

Sincerely,

Alexandra Chassanoff

Appendix C: Initial Recruitment Email sent to ARL's SPAR-IR Listserve

Date _____

Dear Colleagues,

We would like to invite you to participate in an online survey entitled, "Metadata Quality Evaluation in Institutional Repositories: A Survey of Current Practices." This research study intends to explore current practices of metadata quality evaluation at academic institutional repositories (IRs) in the United States.

We are interested in surveying participants from ARL-member institutions who are involved with metadata practices and operations at their respective repositories. Specifically, we are seeking participants who can provide information about their IR's creation and management of metadata, usage of metadata schema(s), guidelines and quality control procedures, and metadata policies. If you decide to be in this study, you will be one of the 123 institutions invited to participate in this research.

Participation in the survey is anonymous and voluntary. If you have any questions about the survey, please contact the Principal Investigator of this study at achass@email.unc.edu. The survey should take approximately 20 minutes to complete.

The survey, at <URL> is now open, and will remain open until <date>.

We will be sending a reminder email about the survey in one week. Your assistance in providing invaluable information about this topic is much appreciated.

Sincerely,

Alexandra Chassanoff