

# RLG Programs Descriptive Metadata Practices Survey Results

**Karen Smith-Yoshimura**

**Program officer  
OCLC Programs and Research**



A publication of OCLC Programs and Research

RLG Programs Descriptive Metadata Practices Survey Results  
Karen Smith-Yoshimura, for OCLC Programs and Research

© 2007 OCLC Online Computer Library Center, Inc.  
All rights reserved  
November 2007

OCLC Programs & Research  
Dublin, Ohio 43017 USA  
[www.oclc.org](http://www.oclc.org)

ISBN: 1-55653-373-x (978-1-55663-373-0)  
OCLC (WorldCat): 182518599

Please direct correspondence to:  
Karen Smith-Yoshimura  
Program Officer  
[karen\\_smith-yoshimura@oclc.org](mailto:karen_smith-yoshimura@oclc.org)

Suggested citation:  
Smith-Yoshimura, Karen. 2007. RLG Programs Descriptive Metadata Practices Survey Results.  
Report produced by OCLC Programs and Research. Published online at:  
[www.oclc.org/programs/publications/reports/2007-03.pdf](http://www.oclc.org/programs/publications/reports/2007-03.pdf)

## Contents

Introduction .....	4
Background .....	4
Context.....	5
Metadata Description—Tools and Standards .....	6
The Broader Environment: Inward and Outward .....	8
Economic Considerations .....	10
Next Steps.....	11
Notes .....	13

## Introduction

As the first project in our program to [change metadata creation processes](#), RLG Programs surveyed 18 Partner institutions<sup>1</sup> in July and August 2007 to obtain a baseline understanding of their current descriptive metadata practices. Although we saw some expected variations in practice across libraries, archives and museums, we were struck by the high levels of customization and local tool development, the limited extent to which tools and practices are, or can be, shared (both within and across institutions), the lack of confidence institutions have in the effectiveness of their tools, and the disconnect between their interest in creating metadata to serve their primary audiences and the inability to serve that audience within the most commonly used discovery systems (such as Google, Yahoo, etc.).

Analyzing the survey results helped us identify where there are opportunities to simplify and integrate metadata practices as we seek to reduce costs and design network level services that are sufficiently modular to support local needs.

## Background

We selected the 18 survey participants from RLG partners in the United States and the United Kingdom because they had “multiple metadata creation centers” on campus that included libraries, archives, and museums and had some interaction among them.

Each institution was asked for a list of contacts responsible for creating metadata across its campus. We requested representatives from libraries, archives, special collections (if separate from a library), museums, institutional repositories, digital library programs, or any other site that creates metadata to describe information resources that a researcher, student, or teacher would want to access.

Four institutions chose just one or two people to respond on behalf of the entire institution; others had a half dozen or more responses representing different units. The breadth of responses we received suggest a wide range of metadata practices within and across institutions.

The charts and graphs in this report’s [companion document](#) are open to interpretation. This narrative represents the interpretation of RLG Programs and includes an analysis of results

according to how the respondents characterized their work environments: archival collections, digital libraries, institutional repositories, library technical services, or museum collections.

- Bullets flag questions and possible areas to pursue.

## Context

Most respondents work with both published and unpublished materials, and a small minority deals only with cultural or natural-history objects. Most respondents describe information resources of three types: those which have been reformatted into digital form, those which were born digital, and analog resources. The ratios among these three vary slightly depending on the respondent's workplace environment.

The types of materials respondents describe include still images, text, moving images, audio, cultural objects, computer files, Web sites, maps, and natural history objects. Half report describing cultural objects, indicating that materials often associated with museum collections are also described in other work environments.

A majority of respondents from archival and museum collections focus on specific types of materials. Half of the respondents include licensed resources in their collections, raising the question of how much redundant metadata creation effort this entails.

The number of staff dedicated to the process of describing resources ranged from zero (no fully dedicated staff) to sixty. Almost half describe resources with one to four dedicated staff. Library technical services, often perceived as “entrenched,” had the highest number of respondents who had been reorganized within the last two years. By contrast, archival and museum collections seem to undergo the least reorganization of the subgroups.

- What are the motivations, internal and external, that precipitate a reorganization?
- What sorts of changes constitute a “reorganization”?
- To what degree have units’ reorganizations affected the sharing of metadata across libraries, archives, and museums and the productivity or effectiveness of metadata creation?
- To what degree do the sources of licensed materials provide metadata—and could more be provided?

## Metadata Description—Tools and Standards

Respondents cited a great variety of metadata tools. These vary with the data structure and data content standards being used, and, to some degree, with the types of materials being described. The six categories of common metadata tools—archival management system, Collection Management System (CMS), Digital Asset Management System (DAMS), digital collections software, Integrated Library System (ILS), and Institutional Repository software—generate at least some standard-based records, but not all do.

Asked to name all the tools used to create, edit, and store metadata descriptions, respondents listed a total of 263 tools, which we normalized to 64 different ones. The majority of these tools are geared towards the libraries, archives, and museums markets for describing, editing, or providing access to content. The rest represent generic database systems, XML-related tools or technology, and text editors or spreadsheets. **The single most common response was “a customized tool,” cited by 69% of all respondents.** These include homegrown tools for creating, managing, or providing access to archival collections or finding aids, images, digital assets, and collection or content management.

Customization can inhibit the sharing of tools and metadata. The survey results suggest that the degree of customization correlates with the 44% of respondents who noted that only some of their systems can create standards-based records for export and sharing and the 50% who said they “sometimes share” their technological infrastructure.

- Why are there so many variations of metadata tools?
- What local needs are customized tools designed to meet that cannot be met with off-the-shelf tools?
- To what degree does customization result in redundant effort within and among institutions?

About 40% of respondents are able to generate some metadata automatically. We characterized about 60% of the metadata generated automatically as technical or administrative, and the remaining 40% as descriptive metadata.

- Can the tools used to generate descriptive metadata automatically be generalized?

Given the prominence of the Integrated Library System as a tool, it is no surprise that MARC is the most widely used data structure standard among the fourteen cited. But if we combine the Dublin Core Unqualified and Dublin Core Qualified responses, “Dublin Core” would be the second most common data structure, followed by Encoded Archival Description (EAD). Few institutions reported using any of the established profiles for Dublin Core. Combining CDWA (Categories for the

Descriptions of Works of Art) and CDWA Lite would represent just about half of those using VRA (Visual Resources Association) Core, designed to cover most visual materials. This correlates with still images being the most common type of material described by respondents.

The data content standards reported show a surprising number of respondents still using APPM (*Archives, Personal Papers, and Manuscripts*)<sup>2</sup> rather than DACS (*Describing Archives: A Content Standard*),<sup>3</sup> which was approved by the Society of American Archivists in 2004. Most are using **both** APPM and DACS, which indicates to us that it takes more than three years for one content standard to replace another. ISAD(G) (International Council on Archives, *General International Standard of Archival Description*)<sup>4</sup> and ISAAR (CPF) (International Council on Archives, *International Standard Archival Authority for Corporate Bodies, Persons and Families*)<sup>5</sup> are used in both the United Kingdom and the United States. The relatively new CCO (*Cataloging Cultural Objects*)<sup>6</sup> had a surprising uptake of 21%. The number one data content standard used is still AACR2 (*Anglo-American Cataloging Rules, Rev. 2*),<sup>7</sup> which correlates with the use of an Integrated Library System and MARC.

- How much longer will it take for the people still using APPM to migrate to DACS?
- How will RDA ([Resource Description and Access](#))<sup>8</sup> affect the data content standards used across libraries, archives, and museums?

Controlled vocabularies represent a significant investment. All controlled vocabularies used that had a response rate of over 30% come from just two organizations: the Library of Congress and the Getty. The three used by more than two-thirds of respondents are the Library of Congress Subject Headings, the Library of Congress Name Authority File, and the Art & Architecture Thesaurus. Use of other controlled vocabularies correlate roughly with the types of materials cataloged. **Nevertheless, about half the respondents build and maintain one or more local thesauri.**

Segmenting responses by workplace environment indicates that museum collections and digital libraries build and maintain local thesauri the most, and libraries do so the least. The types of information included in these local thesauri, in order of the response rate: genres of materials, topics, people and organization names, places, and time periods. Since a large majority of all respondents agreed or strongly agreed that “user-supplied tagging in addition to controlled vocabulary is the best for the resources we describe” and an even larger majority agreed or strongly agreed that “it is critical to provide controlled vocabulary to the resources we describe,” the relatively frequent use of local thesauri may suggest a need to make it easier to contribute to shared terminologies.

- Why do institutions develop local thesauri?
- Are there similarities between and among these local thesauri?

- Would terms in local thesauri make standard controlled vocabularies more robust and useful to wider audiences?

Only a small minority agreed or strongly agreed that user-supplied tagging “is the best option and will obviate the need for controlled vocabularies.” This contrasts with the importance respondents gave to the “intended audience of the metadata” as a leading factor in determining the practices used to describe information resources. Only museum collections and library technical services respondents gave slightly higher importance to material type over audience. Existing skills of staff and system limitations were relatively less important factors in determining descriptive metadata practices.

The importance accorded to audience contrasts with the later responses on effectiveness of the metadata tools used, where about one-third of all respondents didn’t know how effective their tools were, and another third thought they were only partly effective. One stated, “We do not really know if researchers outside the institution find what they want in or need in our catalog.”

- Given the importance of “audience” as a factor for descriptive practices, what user assessment tools could be deployed to measure their effectiveness?

## The Broader Environment: Inward and Outward

The RLG partners were selected for this survey partly because they were reported to already have established interactions among the libraries, archives, and museums within their campuses, so it is no surprise that almost all of the respondents noted that their staff works with other units within their library, archive, or museum and people outside their library, archive, or museum but within the institution. In addition, most respondents reported that other units describe the same or similar types of materials that they do. The staff who create metadata have the same or similar expertise as staff in other units within the institution for most respondents.

The survey responses indicate that there is less sharing when it comes to technical infrastructure, discovery environments, descriptive strategies, and metadata creation guidelines. Ratios of sharing or “sometimes sharing” vary by workplace environment.

Technological infrastructures seem to be shared most widely by archival collections and library technical services, but still by less than half of them. The discovery environment is “sometimes shared” by half or more of all workplace environments. Metadata creation guidelines are least likely to be shared across workplace environments.

- What are the characteristics of the institutions with the greatest amount of sharing descriptive strategies and metadata creation guidelines?

- Should those characteristics be fostered by other institutions?

The intended audiences for locally created metadata correlate with the type of institution. In general, 80% or more of respondents serve an affiliated population (students, faculty, visiting researchers, and academic staff) but even more also cite the need to serve the “interested public.” More than half of all respondents identified a “primary audience,” which we would not expect in a networked world. This tendency to “look inward” may be a factor in the degree to which institutions expose their metadata for use by others outside their local population.

A third of all respondents do not have MARC metadata; this includes a majority of the museum collections and Institutional Repository respondents. A majority of those that do create MARC metadata expose it, predominantly through a Z39.50 server; a quarter do so by using the [Open Archives Initiative—Protocol for Metadata Harvesting \(OAI-PMH\)](#). Only one respondent uses [SRU](#) (Search/Retrieval via URL) or [SRW](#) (Search/Retrieval via the Web). About a quarter of the different workplace environments do not expose their MARC metadata.

- Why do institutions that create MARC metadata not expose it?

About 40% of respondents expose some or all of their metadata to OAI harvesters. The responses vary according to workplace environment. Archival and museum collections seem to be the least likely to expose metadata to OAI harvesters; digital library and library technical services seem to be the most likely. More respondents—almost 60%—provide a Web interface for crawlers such as Google, Yahoo, and MSN, exposing at least some of their metadata on the Web via hypertext transfer protocol (http).

Again, workplace environment seems to influence the degree of such exposure. A majority of institutional repositories expose their data to crawlers, but only a minority of museum collections do. The 2006 OCLC publication, [College students’ perceptions of libraries and information resources](#),<sup>9</sup> notes that 89% of college students use search engines to begin an information search but only 2% begin an information search on a library Web site.

Since 90% of survey respondents viewed their “interested public” as an audience for the metadata they create, we would expect more efforts to expose metadata to the large-scale information hubs where the users (or “metadata audiences”) are found, both affiliated populations and the interested public.

- What constraints inhibit the exposure of metadata to OAI harvesters and Web crawlers?
- What descriptive metadata is really needed in an environment where users look first to search engines to fulfill their information needs?

- What strategies are being considered or pursued to ensure that descriptions of institutional holdings are optimized in search engines?"

Given the immense and increasing popularity of large-scale information hubs like Flickr and YouTube—which contrasts markedly with the *decreasing* popularity of library Web sites—we found it interesting that so few respondents push their metadata out to these hubs: only four do so.

- What can we learn from the experiences of the respondents who push their metadata out to large-scale information hubs?

## Economic Considerations

Most respondents measure productivity by counting “units,” predominantly using a record as the unit. The amount of time to complete one unit seems to vary by type of metadata tool used. A large majority of respondents reported backlogs, and less than half are able to keep up with additions to the information resources/collections they describe.

Backlogs and the inability to keep up with new additions mean some sizeable portions of collections have not been adequately described, and are unlikely to be described without additional resources, funding, or both.

In the aggregate, **almost half of all respondents estimated that 30% or more of their collections will remain inadequately described.** Looking at the results by workplace environment, it appears that museum collections are far more likely to have less than half of their collections adequately described. Archival collections are next likely to have “hidden collections.”

- What metadata tools could best facilitate museums and archival collections to adequately describe their collections?
- How do the descriptive strategies employed affect the ability to adequately describe a collection?

We found it interesting that 18% of all respondents do not have any criteria to measure the effectiveness of their metadata creation tools. Noted one: “*Since we don't have much choice in our selection of a metadata creation tool, there doesn't seem to be much point evaluating its effectiveness.*” Most cited some type of access metrics, and others subjective criteria such as “user feedback” or “ease of use; facilitation of standardization; consistency; accommodation of differing collection types; ease of output; ease of maintenance/editing; ease of linkage to described object; ease of linkage to other descriptive resources.”

Whatever the evaluation criteria used, we asked respondents to rate whether the tools used were effective. In the aggregate, only a quarter think the tools used are effective, a third thought they were partly effective, and a third didn't know. The comments accompanying the "partly effective" responses indicate that we need to know more about what constitutes an effective tool:

*The tools are all over the place. We do not have a cohesive, well thought-out much less well enforced metadata strategy.*

*We do not really know if researchers outside the Institution find what they want in or need in our catalog.*

*Need further automation for creation and additional conduits for exposure/sharing.*

*We use a variety of tools to produce a variety of records. Mature and established systems (such as our ILS) are generally effective. Tools for creation of XML are not as efficient - particularly EAD creation. Creation of EAD and ingest into our XML database is still a very manual process. Our tools are also generally not well integrated. Even when describing the same resource we use the ILS for creating MARC, home grown tools for creating EAD, and perhaps a third tool for creating item level descriptive metadata.*

*Better integration of metadata creation tools with vocabulary tools would improve efficiency. Metadata can be repurposed in different metadata systems, but not as smoothly and easily as necessary.*

## Next Steps

In analyzing the results of this survey, we were struck by an inward focus: the use of local tools to reach a generally local audience.

The customized systems in use are only shared to a limited degree with others within an institution, alongside a moderate sharing of discovery mechanisms, descriptive strategies, and metadata creation guidelines. Despite customization, the tools used are at best only "partly effective" by the respondents' own evaluation criteria, where these exist.

Arguably, one benchmark for effectiveness is to provide access to *all* of one's collections. With the prevalence of backlogs and inability to keep up with additions to one's collections, institutions will fall farther behind without some substantial changes in descriptive strategies, tools, or both.

While most count the public among the audience for their resources, respondents still see their primary audience as restricted to affiliated users (students, faculty, and staff). Arguably, both affiliated and unaffiliated audiences congregate in large-scale information hubs, which current disclosure strategies target only to a limited degree.

To reach users wherever they are, we as a community need to disclose more metadata to OAI harvesters, Web crawlers, and also push metadata out directly to information hubs. For disclosure to be effective, search engine optimization is crucial.

RLG Programs will be following up on the questions raised by the survey responses in other projects in the [Renovating Descriptive and Organizing Practices](#) theme of our work agenda. Our goals:

- Maximize resources to generate as much metadata as quickly as possible.
- Optimize descriptive data with Web information hub targets in mind.
- Share tools and descriptive strategies as much as possible.
- Leverage terminologies from as many sources as possible.

Meanwhile, take a look yourself at [the charts and graphs](#) and let us know what *you* think!

