Striking a Balance: Metadata Creation in Digital Library Projects

Holly Mercer

Holly Mercer recently began work at the University of Kansas Libraries, where she directs metadata creation for campus digital library initiatives. Prior to becoming Metadata Coordinator at KU, she was a Metadata Architect at the Stanford Graduate School of Business where she helped develop information management best practices including metadata standards and training.

Abstract

As institutions develop digital libraries, the work of describing these collections is increasingly taking place outside the library. Content creators and resource authors catalog digital content, yet may not have experience with descriptive metadata standards. One central library repository for metadata has been replaced by multiple collections managed within and outside the library. Librarians act as metadata consultants to digital projects, providing guidance in selecting standards, assistance in metadata creation, and provision of services for resource discovery. Issues affecting quality metadata generation and research in author-generated metadata are reviewed. Successful cases of collaborative and distributed metadata generation are presented, along with strategies implemented with success to achieve quality metadata (Name).

Introduction

Institutions desire to offer access to materials in digital format, whether they are locally developed collections or licensed electronic resources. Metadata harvesting, federated searching, and web services make searching possible across repositories and institutions to locate scholarly digital materials. Most universities do not maintain one Digital Library, but several repositories serving diverse communities and functions. Digital library initiatives include library and archival repositories. Increasingly, digital libraries are also developed as research projects or departmental programs, independent of any library or campus initiatives. Efforts to develop institutional repositories are gaining momentum, and learning management systems function with reusable learning objects. It is easy to see the advantages of having seamless access to the widest variety of resources for learning and research in this networked environment. More content is available digitally, but potential users may not know it exists. Among the challenges these developments present are the changing roles and responsibilities of faculty and staff, both internal and external to the library. New models place more responsibility for managing the resources on the creators of the content. Libraries do not manage these digital resources, yet are being called upon to provide access to these collections. While libraries may provide unified access to these resources, they likely will not catalog them. Librarians will provide leadership as consultants, trainers, and service providers. They will advocate for common standards, train others in metadata creation, and provide value-added services such as federated search capabilities and metadata record enhancement. Metadata is a key in the provision of digital resources.

Issues in Metadata Creation

Descriptive metadata assists in resource discovery and evaluation – searching and browsing. It acts as a surrogate for an actual resource; non-textual resources require textual metadata for queries (Lagoze). Descriptive metadata and subject schemes serve as a way to organize resources and provide a structure for browsing collections. Standards are necessary to ensure interoperability of diverse resources in distributed collections. Use of a common metadata standard brings order to collections of resources in various formats, or from different knowledge domains.

In order for metadata to be useful, it must meet a certain level of quality. Guy Tozer states it must be complete, accurate, and understood by the user (xxi). Standardizing metadata element values through the use of controlled vocabularies and encoding schemes can raise the level of quality, and also makes data input easier. Fewer choices can mean fewer chances for mistakes. Controlled vocabularies and thesauri also help establish relationships among resources. Incomplete metadata may not adequately identify a resource; it can hamper resource discovery.

Legacy library technologies rely on a centralized database of metadata records. A "disinterested group of information professionals (i.e., *librarians*)" (Brooks) catalogs resources represented in a central database. A widely adopted subject classification scheme accompanies the metadata scheme. Interoperability is not problematic because of the common framework. Providing subject analysis and authority control for resources are challenging and time consuming. Libraries therefore cooperatively share in the cataloging of non-unique materials.

Often resources in digital libraries are one of a kind, or are digital surrogates of unique resources. Original cataloging such as would be necessary for many digital resources is an expensive undertaking. Libraries strapped by budget and staffing limitations cannot catalog everything nor outsource everything to a third party. Libraries may not manage digital library collections that are distributed throughout a university. Librarians may not play a part in metadata generation for those resources. Still, libraries need to play an active role in digital library initiatives and in development of institutional repositories.

Who will perform subject analysis, practice authority control, and establish relationships among resources necessary for creating quality metadata? While computer programs may one day be able to perform semantic analysis, humans must make these connections now. Since it is not scalable for librarians to catalog all digital resources, the task falls to others. If disinterested information professionals cannot generate the metadata, then who better but the creators of the resources themselves?

Jane Greenberg has conducted extensive research in metadata creation. She describes four classes of human metadata creators: professional creators, or experts; technical creators; subject enthusiasts; and content creators, or authors (17). Experts are catalogers and indexers. They are advisors to technical creators and authors. Technical creators have some training in creating metadata, or some knowledge of the content being described. They are typically graduate assistants, program or project managers, or administrative assistants. They work with simple metadata schemes to create metadata, and edit or enhance metadata generated by content creators. Community or subject enthusiasts generally have not had formal training but do possess subject knowledge of, or interest in, a specific research area.

Content creators are authors, or those who have primary responsibility for the intellectual work (Greenberg 17). They have intimate knowledge of the resources they create as well as knowledge domain expertise. Content creators often possess information about the resources they have created that others would not know (Greenberg et al.). They may also possess a better understanding of relationships among resources that exist in digital collections.

Metadata generation by non-professionals is not new, but managing a variety of digital resources in an organized, integrated fashion is a challenge. Non-experts must understand the importance of standards and how their metadata fits into this new landscape. One obstacle to non-expert generated metadata may be the term *metadata* itself. People who are not library or information professionals may not understand the term. Persons who are experts in their field are often not familiar with data structures and standards.

If authors are not currently creating any metadata for their resources, buy-in to the new workflow is needed. Authors should understand the importance of metadata – that it is an integral step in the creation of the resource and in its publication. Authors can help others locate and evaluate the usefulness of their resources. Content creators have a personal stake in the availability of their work.

One concern authors may have is the perceived increased workload from creating metadata. There may be a sense that the administrative work of generating metadata takes away from the intellectual work of creating the resource in the first place. Counter-arguments include that by creating metadata that adheres to standards, federated searching and metadata harvesters will make faculty research available to a wider audience. Large scale metadata availability makes conducting research easier because resource materials are identified more quickly. Often content creators are already creating metadata; journals may require an abstract and keywords to accompany a submission, for example.

In one study conducted by Greenberg, expectations for quality metadata creation for non-experts were low, but research and anecdotal evidence did not support this assumption (Greenberg et al.). Her investigation suggested that authors see the value in metadata and believe they should be responsible for creating it. Authors want some assistance from professionals during metadata creation, and wish to be notified if another alters their metadata records (Greenberg and Robertson 49).

Assignment of subject terms is potentially a problem in creating quality metadata because “there are many questions about the author's ability to provide adequate subject access without being trained in the principles of subject analysis” (Greenberg et al.). Milstead reported that metadata created by or under the auspices of its creator is expected to predominate, “largely because the traditional third-party methods (a.k.a. cataloging and indexing) simply cannot cope with the massive and rapidly growing number of electronic objects in existence" (Milstead and Feldman). In one study, one third of participants indicated expert assistance in cataloging would have been useful, and more than half indicated they wanted assistance in assigning subjects and keywords (Greenberg and Robertson 48).

Studies show that metadata produced by non-experts can be as good, if not better than that of expert catalogers (Greenberg et al.). Tools used to produce quality metadata include people employed in the proper roles and workflows, standards and documentation, and devices such as metadata templates, editors, and generators (Greenberg 18).

Raym Crow details strategies to encourage widespread faculty participation in institutional repositories in SPARC Institutional Repository Checklist & Resource Guide. These strategies are applicable for encouraging participation in digital library initiatives in general. These tactics include:

- Produce a briefing paper
- Establish a project web site
- Identify existing problems the repository will solve
- Present the case at departmental and committee meetings
- Distribute literature
- Place articles, public service announcements, advertisements in publications
- Identify key persons to champion the project
- Develop an early adopter plan

Cases

The Stanford University – Graduate School of Business (GSB) began in 2000 BestWeb (Business Electronic Strategy and Technology Web), a portal development project including a major site redesign and repackaging of the School's Web-enabled content. One aspect of the project was deployment of a content management system to increase productivity, disseminate management knowledge, and foster communication. There were essentially three administrative roles in the content management system: producers, who had oversight in all areas of content delivery; web authors, who created the content and input into the system; and indexers (professional librarians), who assigned keywords from a thesaurus built in-house for the GSB student portal. There were many challenges in implementing and operating such a full-featured system; not all web authors had the same level of proficiency or comfort level with the new application, or the same understanding of a web publishing system. There was resistance to using the system because the user interface for staff was poor. Producers and authors handled content creation and entry of all descriptive and administrative metadata with the exception of keywords. As all users of the system became more comfortable with its intricacies, the web authors began assigning keywords to their own content. After a trial period where the indexers reviewed the author-generated keywords, that responsibility was turned over in full to the web authors. Indexers continue to collaborate with producers and web authors in maintaining the student portal thesaurus. The perception was that the web authors could not produce quality keywords that would aid in searching within the portal, but the reality was that the keywords produced by the authors were adequate. With training they understood the issues involved in assigning keyword terms to improve access to their content.

Another case from the GSB is an example of collaborative metadata generation among faculty and staff. The Graduate School of Business faculty, assisted by case writers, produce case studies for teaching. Case writers and faculty create metadata for each case, such as: title, creator, abstract, date and version, and subject and keywords. A coordinator manages access to and distribution of all submitted cases and research papers. The coordinator assigns the cases into broad subject categories created by GSB librarians. The coordinator then enters the metadata into a web-accessible database, adding additional metadata (such as rights information), and editing existing records when necessary.

The University of Kansas Libraries have taken an active part in developing the vision for the University's Digital Library Initiatives (KU-DLI). Members of the library have served on planning task forces and cross-functional work teams. They have produced documentation and served as advisors for adoption of metadata standards. The metadata coordinator works with other campus departments to integrate small digital collections into the Digital Library.

The University of Kansas Digital Library Initiatives (KU-DLI) awards grants to campus constituents to encourage development of scholarly digital collections. The projects funded by KU-DLI are accountable for adhering to standards for digitization and preservation as well as metadata. Faculty serve as principle investigators or sponsors for the digital projects. They develop the intellectual content of the collections. The faculty oversee the project, and graduate assistants and staff are the technical creators who generate the metadata for the digital resources. Each project has the flexibility to select a metadata framework for its repository; the metadata coordinator provides guidance and instruction in selecting and implementing standards, including syntax and semantics that are appropriate to the resources in question. By defining a metadata framework within the context of a local collection, communities of practice have more descriptive metadata available to them. For example, geospatial collections can use FGDC Content Standard for Digital Geospatial Metadata, and art image collections the VRA Core Categories. The libraries insure the interoperability by coordinating the integration of local collections in the Digital Library. The metadata coordinator monitors progress in the development of the collections, and coordinates mapping of individual repository formats to the main Dublin Core application used by the Digital Library.

Conclusion

Libraries will increasingly be service providers. They will offer valued-added services, such as spot-checking metadata records or metadata creation and enhancement. More than ever, librarians will be part of a team working to build digital collections. Cross-functional teams will consist of members from the library, information technology, instructional technology, networking, and academic departments. Some people will build content, while others will create policies and infrastructure. Metadata creation may occur in a distributed fashion, with authors, technical metadata creators, and experts working together.

Metadata generation is a collaborative effort. It is increasingly likely to occur outside of the library, but librarians will nonetheless maintain an active role in quality assurance and training. Librarians will instruct members of other campus communities in the creation of metadata. They will be the metadata experts and will work in conjunction with discipline experts in describing digital collections. Librarians will act as metadata consultants to digital projects and provide value-added services such as record enhancement. They will recommend appropriate standards and subject schemes for digital collections. They will maintain documentation, create crosswalks and data dictionaries, and be the institution authority in metadata standards. With the support of this metadata framework, libraries will provide integrated access to disparate, distributed collections through federated searching and metadata harvesting.

Works Cited

Brooks, Terrence A. Where is Meaning When Form is Gone? Knowledge Representation on the Web. Sep. 2000. 17 April 2003 <http://faculty.washington.edu/tabrooks/Documents/Meaning/M.html>.

Crow, Raym. 2002 SPARC Institutional Repository Checklist & Resource Guide. 2002. The Scholarly Publishing & Academic Resources Coalition. 11 March 2003 <http://www.arl.org/sparc/IR/IR_Guide.html>.

Greenberg, Jane. "Metadata Generation: Processes, People, and Tools." Bulletin of the American Society for Information Science and Technology 29.2 (2003): 16-19.

Greenberg, Jane, Maria Cristina Pattuelli, Bijan Parsia and W. Davenport Robertson. (2001). "Author-Generated Dublin Core Metadata for Web Resources: A Baseline Study in an Organization." Journal of Digital Information 2.2.(2001). 19 April 2003. <http://jodi.ecs.soton.ac.uk/Articles/v02/i02/Greenberg/>.

Greenberg, Jane, and W. Davenport Robertson. "Semantic Web Construction: An Inquiry of Authors' Views on Collaborative Metadata Generation." Proceedings for the International Conference on Dublin Core and Metadata for e-Communities, 2002: October 13-17, 2002, Florence, Italy. Firenze, Italy: Firenze University Press. 45-52. 10 Aug. 2003 <http://www.bncf.net/dc2002/program/ft/paper5.pdf>.

Lagoze, Carl. "From Static to Dynamic Surrogates: Resource Discovery in the Digital Age." D-Lib Magazine June 1997. 11 April 2003 <http://www.dlib.org/dlib/june97/06lagoze.html>.

Milstead, Jessica and Susan Feldman. "Metadata: Cataloging by Any Other Name…" ONLINE Jan. 1999. 16 October 2002 <http://www.infotoday.com/online/OL1999/milstead1.html>.

Tozer, Guy V. Metadata Management for Information Control and Business Success. Boston: Artech House, 1999.