Metadata Strategies to Address NSDL Objectives♣

© Dave Fulker

University Corp. for Atmospheric Research/NSDL Headquarters
fulker@ucar.edu

Abstract

There are many challenges surrounding the use of metadata to achieve bibliographic objectives in the National Science Digital Library (NSDL) and similar digital-age enterprises. Metadata uniformity versus descriptiveness, and metadata quality (as viewed by catalogers) versus collection size, both are typical of the tradeoffs to be faced. Furthermore, basic questions are being posed about the metadata-surrogate model. This paper describes how these challenges are being addressed in the NSDL, a distributed endeavor that is intended to engage and meet the needs of educators and learners across the entirety of science and mathematics.

1 Introduction

Current practices of knowledge organization have been developed and refined over a 150-year period. Four decades of this have included the application of computer technologies, including 10 years of research efforts in the United States that have been funded by the National Science Foundation (NSF) to develop and study digital libraries. Some of this research, such as studies by Mischo [6], explicitly addressed bibliographic practices.

Despite this rich history, technologies for creating and disseminating information have outstripped the capacities of bibliographic systems to 1) paint comprehensive views of the entire knowledge landscape or 2) fully address the needs of people who (in growing numbers) expect to navigate this landscape without the assistance of librarians. This is not a criticism; the challenges are profoundly difficult and changes continue to occur with astounding speed. This paper discusses these challenges from the perspective of a recent NSF-sponsored initiative: the National Science Digital Library or NSDL.

The scope of the NSDL is broad: all levels and forms of science, mathematics, engineering, and technology education. The NSF has awarded more than 100 grants to a wide variety of organizations that have proposed to build collections, offer services, or perform focused research as part of the NSDL initiative [12]. This author, with a number of partnering organizations, holds responsibility for "Core Integration" of the NSDL, which includes providing the infrastructure for joining (into a coherent whole) a multiplicity of diverse NSDL collections and other components, operated at various locations. (NSDL content is not held centrally.)

The basic functions of a bibliographic system are *finding, collocating, choosing, acquisition, and navigation* [10], and each is being affected by rapid technological change. We describe challenges associated with these functions, especially in the NSDL context, including tensions that pertain to creating and using metadata, the usual underpinnings for the needed functionality. In part because of the NSDL's distributed nature—it eventually will embrace a large number of rather independent collections and services—the associated bibliographic system must cope a variety of interesting problems borne of the digital age.

2 Some Digital-Age Realities

2.1 Scalability of Bibliographic Systems

There may be no greater problem facing contemporary digital libraries than scalability. The problem derives primarily from increased creation rates of materials, combined with user expectations that are driven by their experiences with Web search engines.

A tempting solution might be to rely entirely on Google-like approaches for all bibliographic functionality (which completely eliminates the bottleneck of human metadata creation), but NSDL studies [9] indicate that users expect levels of functionality that are beyond what content- and link-analysis systems can deliver, at least for now. Hence an important challenge is to chart a course for bibliographic systems that meet user expectations (even in specialized domains) and simultaneously exploit the scalability of automated approaches to information organization and discovery.

2.2 Dynamic Content and Atomicity

Resources that are commonplace in the computer age do not necessarily come packaged neatly as indivisible (atomic) units, and some of the most expressive media, such as real-time observational data, are highly dynamic. A challenge arises because items that change or have uncertain boundaries create identity problems.

Proceedings of the 5th Russian Conference on Digital Libraries RCDL2003, St.-Petersburg, Russia, 2003

Furthermore, it is difficult to describe, and thus to organize, anything that is hard to identify. In *The Intellectual Foundations of Information Organization*, Svenonius points out:

> "Documents with uncertain boundaries, which are ongoing, continually growing, or replacing parts of themselves, have identity problems. It is not possible to maintain identity through flux ('On cannot step twice into the same river' [4]). ... A snapshot cannot accurately describe information that is dynamic. This is not simply a philosophical matter, since what is difficult to identify is difficult to describe and therefore difficult to organize." [10]

Unfortunately for bibliographers, such documents are an increasingly common feature of the knowledge landscape, especially for library users in science and engineering, the primary foci of the NSDL. For reasons of practicality, the initial release of the NSDL simply employed the identity mechanism associated with the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) [5], as this was the means for acquiring essentially all NSDL metadata records.

Future versions of the NSDL infrastructure will implement means for indicating (various forms of) document *equivalence*. In addition to handling those cases where two or more distinctly identified metadata records or URLs point to identical documents, this mechanism will permit subtle notions of equivalence (such as multiple versions of the same document, or even significant amounts of identical text) to be documented for use in library services. The implication is that a multiplicity of equivalence relations potentially can enrich and formalize the traditional *collocating* function of bibliographic systems.

2.3 Universality vs Specialized Expressiveness

Some of the most important bibliographic-systems developments were motivated by ideals of universality as ambitious as a common organizational framework for all human knowledge. Such ideals also informed the development of digital metadata frameworks, especially those associated with traditional cataloguing systems, but another force also has been at play: the power of computers and networks to make accessible very large amounts of highly diverse materials has led to needs for greater *specificity* in many domains. Thus bibliographic system designers face tradeoffs between maximizing compatibility with other (ostensibly universal) systems and maximizing expressiveness, relative to the needs of a target audience needing access to specific collections.

Metadata frameworks, such as Dublin Core, address this problem through hierarchies that allow varying degrees of domain-sensitive expressiveness, linked to well-defined semantics and vocabularies [2]. Despite these advances, the use of such frameworks may not be increasing. Robert Wilensky expressed a somewhat typical view of these matters in an unpublished white paper prepared for an NSF workshop (Chatham, Massachusetts, June 2003) on the future of digital libraries. In a section titled "Living Without Metadata" he wrote:

> "Attempts to produce large, useful digital libraries run into key problems associated with metadata. Typically, objects are either missing key metadata or the metadata is difficult to interpret—it has been supplied according to a different schema or the interpretation with respect to a schema is unclear. We note that the most successful tools, e.g., web indices like Google, work as well as they do precisely because they do not rely on any metadata whatsoever. Of course, this approach undeniably limits usefulness."

The NSDL projects and other contemporary digital-library activities exhibit a wide range of responses to the universality/expressivity tradeoffs and the metadata problems that Wilensky articulates. One approach is manifest by the Alexandria Digital Library at the University of California in Santa Barbara, which features normalized search and discovery across geospatial collections having heterogeneous metadata. The approach employs middleware and distributed searches [8], and seems especially well suited to domains where strongly typed variables (e.g., latitude and longitude) are essential for characterizing the resources of interest. As discussed below, the NSDL plans—which must accommodate all scientific disciplines—will seek a synergistic relationship between text-based indexing (in the style of Google) and various forms of computer- and human-generated metadata, including annotations created by end-users (not cataloguers).

2.4 Mixed Modes of Metadata Creation

It is evident among the NSDL-funded projects and elsewhere that the number of methodologies applied to metadata creation is large. Though a common thread is to reduce human effort—especially by cataloguing experts—the approaches run the gamut from complete automation to a variety of leveraging strategies. Some systems make use of typed metadata fields (education grade level) while other successful efforts employ few data types beyond free text. It may be noted that this situation differs markedly from an era in which individuals that were explicitly trained for such work created nearly all bibliographic metadata.

Hence an important but open question for the digital age is how to fulfil the bibliographic functions of finding, collocating, choosing, acquisition, and navigation in contexts where the joining of multiple libraries and collections yields a huge mixture of metadata approaches. The NSDL strategy for meeting this challenge is discussed below.

3 An NSDL Strategy

3.1 Partners in Bibliographic Roles

The initial release of the NSDL (launched in December of 2002) embodied successful use of the OAI-PMH methodology for aggregating metadata records from multiple, geographically distributed sources into a centralized metadata repository [11] now containing over 200,000 entries. Further, the same protocol allows NSDL services to access the repository. Capabilities now supported by this framework include a search service, a preservation service (which generates weekly snapshots of all openly-accessible NSDL content), a user-interface at http://nsdl.org/, and several embedded services, including news, exhibits, collection-level "branding" and an analysis of broken-link patterns.

Recent work has eliminated several human steps from the workflow, so the foundations are in place for significant collection development. Emphasis will be placed on the engagement of key partners, including publishers, and professional societies. Partners will convey their Dublin Core (DC) metadata to the NSDL via OAI-PMH or some lighter-weight method for small, static collections. Means soon will be offered for also accepting metadata in ONYX form. More on these matters may be found in the NSDL Collection Development Policy, which is now being finalized and soon will be accessible at http://nsdl.org/.

The NSDL project soon will explore human-mediated automation of metadata creation, probably by seeking partners such as the INFOMINE project, with their iVia software [7].

However, to augment DC-dependent collection development, the CI team will develop a lightweight mechanism by which users voluntarily contribute and describe resources for NSDL. Actual inclusion of such resources will be guided by policy and hierarchical review, consistent with the aforementioned Collection Development Policy. Once the NSDL has engaged a significant number of users, this method should help the NSDL gain high-quality resources at a rapid rate.

3.2 Users in Bibliographic Roles

A framework for characterizing user-identified resources (i.e., a metadata schema and associated user interfaces) is currently being defined and the details are beyond the scope of this paper. However, the highest priority will be to gain descriptions that are oriented toward educational contextualization of NSDL resources, the area in which end-users (i.e., educators) bring the greatest experience.

Direct user involvement in collection-building and resource description will yield several key benefits: timely addition of relevant new resources (before they become well known); immediate data on user needs (to help maximize NSDL responsiveness); a heightened sense—among users—of ownership and pride in the NSDL; data on relationships between NSDL resources and the educational contexts in which they are applied, as viewed by those who apply them.

The last of these merits additional discussion, because it relates so directly to the educational purpose of the NSDL. A major dissatisfaction among teachers who use the Web is the difficulty of determining whether a specified resource is applicable to a given educational context [9], where "context" may refer to curriculum, grade-level, standards (district, state, and national), and even gender or other cultural factors. Furthermore, if NSDL causes teachers to think deeply about such contextual factors, it seems likely that their effectiveness as educators will be enhanced.

3.3 Emancipation from a Simple Surrogate Model

As reported by Arms [1], metadata harvesting as manifest in the NSDL initial release has set the stage for incorporating future output from small- and large-scale collection-development activities, including ones that leverage human effort with automation. However, Arms also describes important understandings about the limitations of DC item-level metadata surrogates as a foundation for library development and services:

- Variations among metadata providers.
- Immaturity, imprecision, and interpretation of standards.
- Scalability tradeoffs.

These problems are compounded by another: with growing complexity in the data to be represented, standard cataloguing models—those with one-to-one relationships between documents and metadata surrogates—become increasingly limiting. This is true whatever the semantics of the item-level record, such AARC/MARC, Dublin Core, IMS, etc. The precision and expressiveness of the item-level surrogates can be increased, but they will still fail to adequately model the rich relationships that exist between multiple entities and make environments like Amazon.com so attractive to users. In other words, even if one increased the expressive power of Dublin Core and addressed all the problems described in the preceding paragraph, the associated data model imposes unsatisfactory limitations on the provision of advanced services that meet expectations of modern users.

NSDL plans include embedding the present metadata repository in a more flexible data space, designed to support library services that require rich information about user and document *contexts* and about dynamic relationships among documents. Among other benefits, this will allow user-provided contextualization data—as discussed in the preceding subsection—to be represented and utilized to achieve enhanced library services. A major challenge (to be met through discourse with numerous stakeholders) will be to formalize and implement interfaces to this data space that are sufficient to support an increasingly rich and diverse array of NSDL services.

3.4 Architectural Implications

Though digital-library progress has been achieved in important areas (including OAI-PMH and Dublin Core metadata), a common architecture for digital libraries remains elusive. Yet to be realized fully on a large scale is a goal articulated by Besser for digital libraries: "*to deliver information to multiple clienteles, using the same collection to serve many different groups of users, each with its own level of knowledge and modality of learning and interacting"* [3]. It is hoped that the advances discussed above, especially the engagement of clients in contextualization of resources, will position the NSDL somewhat closer to Besser's goal.

However, the contextualization concept depends on easy and significant end-user participation in the NSDL, and it is unclear whether the NSDL as a Web site (at nsdl.org) can engender such participation. It may be noted that digital libraries, particularly those whose presence is manifest as a Web site, often loose track of clients as soon as they find what they want. This precludes some types of user interaction and may fundamentally limit the library's power to gain understanding about any user group's *level of knowledge and modality of learning and interacting*, reiterating Besser's goal. (Note: the potential gains from contextualization data clearly include improved understandings on these matters.)

Hence consideration is being given to other forms of NSDL presence. It is premature to speculate on conclusions, but interesting experiments are underway on how the NSDL might maintain a presence in browser extensions, similar to the Alexa toolbar at alexa.com.

A final point on architecture is that libraries now must address the implications (in the presentation of search results, for example) of utilizing multiple forms of metadata, as discussed in 2.3. An illustrative and open question is: when ordering search results, should those identified as relevant via conventional metadata be ranked higher or lower than those identified via full-text searching and link analysis? A possibility being considered for the NSDL is to employ Google (or other well-liked search mechanisms) directly for discerning overall relevancy, followed by metadata-based filtering to refine the results and gain precision in respect to the educational interests of NSDL users.

4 Conclusion

Metadata use in the digital age cannot simply be patterned after bibliographic systems from the pre-digital era, even though key principles from that era remain relevant in digital libraries. Matters of scalability, dynamic content, atomicity, universality, specialized expressiveness, and mixed modes of metadata creation are challenging builders of most digital libraries, including the NSDL. In the latter context, progress has been achieved, via OAI metadata harvesting, toward the engagement of partners that can contribute useful metadata and help address certain of these challenges, including scalability.

To complement this partnering strategy, the NSDL effort is now considering two additional steps. One is to have end-users playing bibliographic roles, especially in the educational contextualization of NSDL resources, and this may necessitate new forms of user interaction and new forms of NSDL presence, beyond the Web site at nsdl.org. The other is to expand upon the current metadata repository to implement a more flexible data space, adequate to characterize user and document *contexts* as well as dynamic relationships among documents in the NSDL.

The author hopes that the architectural and other implications of this NSDL work will be of relevance in other digital-library endeavours.

References

[1] Arms, William Y., Naomi Dushay, Dave Fulker, Carl Lagoze. A case study in metadata harvesting: the NSDL. *Library Hi Tech,* Vol. 21 No. 2, 2003.

[2] Baker, T., "A Grammar of Dublin Core", *D-Lib Magazine*, October 2000.

[3] Besser, Howard. The Next Stage: Moving from Isolated Digital Collections to Interoperable Digital Libraries. First Monday, Vol.7 No. 6, June 2002.

[4] Heraclitus. *The Encyclopedia of Philosophy.*

[5] Lagoze, Carl, Herbert Van de Sompel, Michael Nelson, and Simeon Warner. The Open Archives Initiative Protocol for Metadata Harvesting-Version 2.0, 2002. http://www.openarchives.org/OAI/openarchivesprotocol.html

[6] Mischo, William H., Thomas G. Habing, and Timothy W. Cole. Integration of Simultaneous Searching and Reference Linking across Bibliographic Resources on the Web. *Proceedings of the Second ACM/IEEE-CS Joint Conference on Digital Libraries*, (JCDL 2002).

[7] Mitchell, Steve, Margaret Mooney, Julie Mason, Gordon Paynter, Johannes Ruscheinski, Artur Kedzierski, and Keith Humphreys, iVia Open Source Virtual Library System, *D-Lib Magazine*, January 2003.

[8] Smith, T., G. Janee, J. Frew, and A. Coleman. The Alexandria Digital Earth Prototype System. *Proceedings of the First ACM+IEEE Joint Conference on Digital Libraries*. (JCDL 2001).

[9] Sumner, Tamara, Michael Khoo ,Mimi Recker ,Mary Marlino : Understanding Educator Perceptions of "Quality" in Digital Libraries. *Proceedings of the Third ACM+IEEE Joint Conference on Digital Libraries*. (JCDL 2003)

[10] Svenonius, Elaine. *The Intellectual Foundations of Information Organization.* MIT Press, 2000.

[11] Warner, Simeon. Report on the "OAI Metadata Harvesting Workshop" at JCDL 2003. "In Brief" column of *D-Lib Magazine,* July/August 2003.

[12] Zia, Lee L., The NSF National Science, Mathematics, Engineering, and Technology Education Digital Library (NSDL) Program : New Projects in Fiscal Year 2002. *D-Lib Magazine*, November 2002.

♣ The work described herein was supported by grants from the National Science Foundation of the USA.