Mixing and Mapping Metadata to Provide Integrated Access to Digital Library Collections: An Activity Report

Submitted by the Cornell University Library ENCompass Team
Karen Calhoun, ksc10@cornell.edu
Tom Turner, tpt2@cornell.edu
Meryl Brodsky, mb297@cornell.edu
George Kozak, gsk5@cornell.edu
Marty Kurth, mk168@cornell.edu
Fred Muratori, fmm1@cornell.edu
David Ruddy, dwr4@cornell.edu
Sarah Young, sy82@cornell.edu

Abstract

This paper provides a report of work in progress to implement integrated access to multiple digital collections that are described using a variety of metadata formats. Using the emerging resource discovery and digital library management system, ENCompass, a team at Cornell University Library is experimenting with a new discovery system model. The model uses simple, "pidgin" metadata at the collection management level, but combines this simple layer with other metadata for describing specific resources, to enable users not only to discover relevant collections, but also to conduct deep searches. The authors frame their ENCompass activity report to illustrate the principle of modularity—as described by Lagoze—in which a metadata format tailored for simplicity (Dublin Core) is used alongside other, more complex metadata formats.

Keywords: Metadata, digital library collections, integrated access, resource discovery, Dublin Core, ENCompass, Cornell University Library

1 Introduction

1.1 Background

Sarah Thomas, Cornell's university librarian, recently wrote "the world's information resources are abundant, but time is a scarce commodity. The ideal discovery tool, therefore, is one which consults omnivorously, but which returns a selection of relevant results in rapid sequence … Such a tool is still imaginary …" [1]. In a perfect world, searchers would find what they need quickly, without having to sort through masses of data stored and organized in multiple places, in multiple ways. A huge obstacle to attaining this goal is the endlessly diverse ways in which information management communities have described their resources. Nowhere is this obstacle more visibly at work than in the Internet Commons, but it is also at work within Cornell's digital library collections.

Within Cornell University's nineteen-library system, multiple digital collection and delivery systems have arisen over the past ten years. There is one unified library portal called the Library Gateway, but it exists alongside numerous other library-affiliated, searchable Web sites and archives. In addition, Cornell library staff members have been fortunate to gain much experience with digital library development through the receipt of grants. Some notable projects include the Making of America (with the

University of Michigan), the Fuertes Ornithological Collection, the Ezra Cornell papers, and the Core Historical Literature of Agriculture [2]. Each of these projects uses different methods and metadata formats to support resource discovery and digital collection management. For instance, the Fuertes Ornithological Collection uses a locally created metadata format, the Making of America and the Core Historical Literature of Agriculture use metadata formats derived from TEILite, and the Ezra Cornell papers are described using EAD. Other digital collections use other standards suited to the particular type of digital material being served up. Searchers must not only become aware that all of these different collections exist, and know where to find links to them, they must also cope with a variety of interfaces, searching protocols, and access methods for connecting to the digital materials themselves.

1.2 The notion of metadata mixing and mapping

At the DCMI conference in Ottawa, Thomas Baker referred to Dublin Core as "a metadata pidgin for digital tourists who must find their way in [a] linguistically diverse landscape … [DC] is well-suited to be an auxiliary language for digital libraries" [3]. Using the principles in his "grammar of Dublin Core," Baker proposes that implementors can combine simple Dublin Core elements and qualifiers with elements from other namespaces into rich vocabularies. Rachel Heery and Manjula Patel's paper at the Ottawa conference introduced application profiles, which consist of data elements drawn from multiple namespaces and "mixed and matched" by implementors for a particular application or community [4]. The concepts laid out by Baker, Heery and Patel have a great deal of merit for implementors who are developing domain-specific application profiles (for example, education or libraries) to be applied *prospectively* to describe the domain's resources. But what is to be done when the metadata that is available for pre-existing digital collections is already "linguistically diverse," as is the case at Cornell? How might the existing metadata be mixed and matched in a single integrated system for discovery and access?

In a January 2001 article, Carl Lagoze extended the concepts introduced by Baker, Heery and Patel at the DC Ottawa conference. Lagoze contrasted the twin goals of Dublin Core—cross-domain discovery and resource description [5] and suggested that DC should serve first and foremost the goal of cross-domain resource discovery. Introducing the "principle of modularity," he proposed that metadata formats tailored for simplicity, like DC, be used alongside others tailored for complexity. To do so, implementors might deploy simple DC metadata for cross-community interoperability, while at the same time offering parallel support for community-specific metadata.

The authors of this paper propose that the Lagoze approach, which has been a major theme of the Open Archives Initiative (OAI), has the effect of extending the Heery-Patel "mixing and matching" application profile model to one that is based on "mixing and mapping."

2 Metadata complexity

2.1 Issues and solutions

Despite the great strides made by the introduction of the Cornell University Library Gateway in 1998, Cornell's library continues to offer its users a confusing array of digital collections and delivery systems. The library's great success with obtaining grants for digital collection building adds to the need to offer an integrated entry point to the library's increasingly rich digital resources. The library faces two important questions:

- How can the library improve the methods patrons use to navigate the breadth and depth of its digital collections?
- How can the library streamline the management of the digital collections it creates or purchases?

Any solution needs to take into account not only prospective work but also the ability to support the current diverse metadata environment. ENCompass, a new product developed by Endeavor Information Systems Inc., offers a set of tools to begin to address these issues. It has the potential to provide simultaneous searching across the library's diverse collections through a single, user-friendly interface while supporting a wide variety of metadata types "behind the scenes."

2.2 The Cornell-ENCompass development partnership

The Cornell University Library, after being named a SUN Center for Excellence in Digital Libraries, was able to dramatically increase its computing resources. ENCompass offers a digital library management system that addresses the

issues of metadata diversity in our complex environment. About a year and a half ago, the Cornell library became the first ENCompass development partner. Other partners include the Kansas State University Library and the Getty Research Library.

As a development partner, Cornell's ENCompass Working Group worked with Endeavor on the system's design. The Cornell team reports to Tom Hickerson, the library's Associate University Librarian for Information Technology. Since the initial commercial release of the ENCompass software the team has suggested and tested improvements and enhancements. During the first phase of the team's work with Endeavor, Endeavor staff loaded two Cornell digital collections—the Fuertes bird images database and the Ezra Cornell papers—into the ENCompass environment. The Fuertes database is based on a locally-developed metadata format and the Ezra Cornell papers use the Encoded Archival Description (EAD). In need of a new digital project that could be created within ENCompass from the beginning—one with prior funding but without a pre-existing delivery system—the team chose Saving America's Treasures (SAT), an effort to conserve and digitize Cornell's extensive Samuel May Anti-Slavery Pamphlet Collection.

2.3 Metadata management in ENCompass

ENCompass offers a hierarchical method for arranging data and metadata. At Cornell, ENCompass collections, the highest level of intellectual organization in the hierarchy, use a base metadata type of simple Dublin Core. Collections can consist of other collections or containers. Containers are the intermediate grouping of digital objects, and they can contain other containers or objects. The metadata at the container level can be anything that the library desires to define. Objects represent the lowest level of information and the place where links are made. Again, the library defines the metadata for objects. For instance, the Fuertes collection object records contain fields for common as well as scientific names of birds.

One of the most important features of ENCompass is its ability to support a variety of metadata types. As libraries move beyond reliance on MARC and expand their digital collections, they need new tools to exploit the rich metadata being produced. The ENCompass design permits the use of pre-existing metadata formats while also allowing the creation of locally defined metadata structures.

There remains the difficulty of supporting cross-collection searching across a variety of metadata formats. This is where "mixing and mapping" comes in. The ENCompass design relies upon Dublin Core as a lingua franca. Existing, domain-specific, standard or non-standard, or newly defined metadata types are mapped to Dublin Core to support searching and resource discovery. In addition, the library catalog may be included so it becomes possible to search ENCompass collections and the catalog at the same time. This is true for Z39.50-enabled databases as well.

One of the most powerful features is what the Endeavor staff call "bubble-up"—this can also be thought of as "reverse metadata inheritance." This means that the selected metadata at the object level is passed up through the system to the containers of that object and then to the collections that contain those containers. This enables searchers to retrieve collections based on very deep searches. For instance, a searcher can discover that the Fuertes Ornithological Collection contains images of bobwhites. A metadata record describing the Fuertes collection could never accomplish that task.

3 Activity report: Saving America's Treasures (SAT)

The Cornell team is planning to make the 10,000 pamphlets in SAT available in digital form through ENCompass by first determining an appropriate data structure, creating scanned images for each page, and OCR'ing the text. The database will also contain collection-level descriptive information, pamphlet-level metadata (derived from MARC records), and page-level metadata. The team's goal is to facilitate searching the full-text of the documents and returning page images for pages containing the search terms. Searchers would then be able to navigate within the pamphlet that contains that page.

The team began by experimenting with a structure for storing SAT materials in ENCompass. Figure 1 illustrates what the team is attempting to do. The collection-level information describes the anti-slavery pamphlets as a group. The pamphlet-level metadata (stored as ENCompass containers) is derived from MARC records for the printed pamphlets and consists of basic bibliographic information. The page-level metadata is the page number, page feature

information (for example, table of contents, title page, illustration, etc.) and OCR text. Each page is stored as an ENCompass object. The OCR is a separate object so that it will only be viewed when requested.

The metadata structure for the Saving America's Treasures materials at the container and object levels is based on TEILite with some slight modifications. The TEIHeader was simplified and restricted, and the <PB> tag was modified to associate page images unambiguously with page level OCR. The element set used captures data analogous to Dublin Core concepts of title, author, publisher, description, subject/keywords, and identifier. The team's approach with this collection has been to view the page image as the primary textual object and the OCR as a secondary representation.

The team has defined title, subject, and OCR in object and/or container metadata records as "bubble up" or "reverse inheritance" elements. The effect of this structure and the "bubble up" definitions will mean, for example, that searchers can get a hit on the OCR data (which will be bubbled up to the upper levels). They will first see a result set that contains the collection with the term or phrase they have searched, and then they will be presented with an image of the exact page

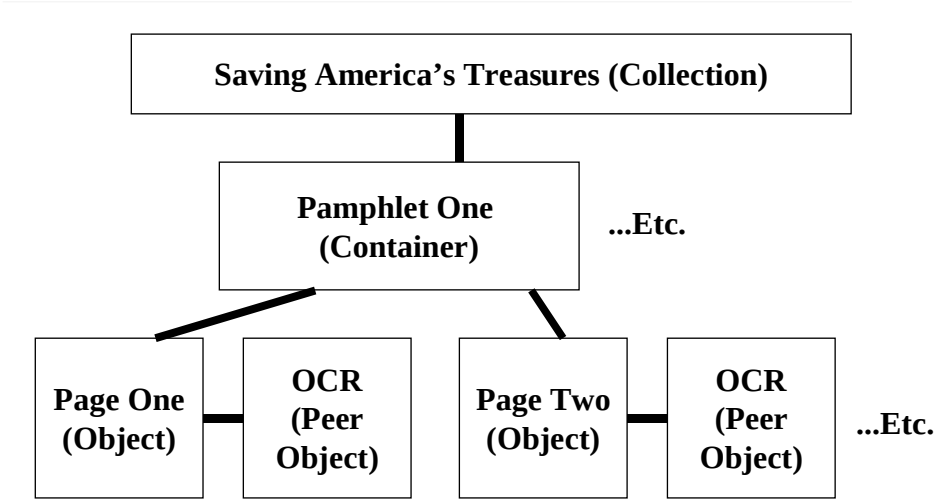

Figure 1. Saving America's Treasures Metadata Structure in ENCompass

that meets the search criteria, along with descriptive metadata for the pamphlet of which the page is a part. For example, assuming the searcher has an interest in slavery in the West Indies and types the query "West Indies," he or she will be led first to Saving America's Treasures as a relevant collection, then to the relevant anti-slavery pamphlets and pages.

The team's work with SAT and ENCompass has promise for allowing searchers to find the most appropriate materials from among the thousands of non-proprietary and proprietary resources that are available to the Cornell University community, and through one integrated interface. Further, through the "mixing and mapping" model of managing diverse metadata formats, the team anticipates the tools will facilitate not only resource discovery at the collection level, but also deep searching.

References

[1] Thomas, Sarah E. Abundance, attention, and access. ARL Newsletter 212: 1-3, October 2000.
[2] The URLs are: The Making of America: http://cdl.library.cornell.edu/moa/; The Fuertes

Ornithological Collection: http://rmc.library.cornell.edu/Fuertes2000/;
The Ezra Cornell Papers: http://cidc.library.cornell.edu/cornell/guide.htm;
The Core Historical Literature of Agriculture: http://chla.library.cornell.edu/.
[3] Baker, Thomas. A grammar of Dublin Core. D-Lib Magazine 6 (11), October 2000. Available: http://www.dlib.org/dlib/october00/baker/10baker.html
[4] Heery, Rachel and Manjula Patel. "Application profiles: mixing and matching metadata schemas." Ariadne 25, September 2000. Available: http://www.ariadne.ac.uk/issue25/app-profiles/
[5] Lagoze, Carl. Keeping Dublin Core simple: cross-domain discovery or resource description? D-Lib Magazine 7 (1), January 2001. Available: http://www.dlib.org/dlib/january01/lagoze/01lagoze.html

Acknowledgements

The authors thank former Cornell ENCompass team members Angi Faiks, Oya Reiger, and Edward Weissman for reviewing early drafts of this article. We also thank Tom Hickerson, Cornell's Associate University Librarian for Information Technologies, for his support of the team's work, together with the ENCompass project staff at Endeavor Information Systems, Inc.