A Training Program for Shareable Metadata

About the Project | Online Course Materials | Resources | Contact Us

01 - Introduction

A Look at Sharing
02 - The Current Sharing Environment

03 - Aggregators

Defining Shareable Metadata
04 - Foundations
05 - Content
06 - Coherence
07 - Context
08 - Activity 1
09 - Communication
10 - Consistency
11 - Conformance to Standards
12 - Activity 2

Shareable Metadata in Practice
13 - Crosswalking
14 - Workflow

15 - Looking Ahead

Metadata for You & Me - A Look at Sharing: The Current Sharing Environment

Module Content

- Screencast
- Powerpoint Slides and Other Resources
- Module Text

Screencast
Module Text

1. Questions to ask yourself

As you read through the content in this module, consider the following questions.

- How do you currently share?
- Who do you share with?
- What do you find hard about sharing/not sharing?
- What technical standards do you use for sharing?

2. Why share?

But perhaps we ought to back up and ask the question 'why share?' How does sharing our metadata benefit our users? How does sharing our metadata benefit us?

We believe that sharing has many benefits for users:

- The metadata can be where the user is. By providing your metadata openly in machine-readable form for any number of services to pick up and use, users can more easily find the resources that are hidden deep in your catalogs, databases, or collection management systems. Library catalogs can notify users of new books via RSS feeds. With the use of persistent URLs and identifiers, users can reliably cite, reference, and call on our digital objects.
• One-stop searching (maybe). The hope for many in the library community especially is that many of the protocols that allow us to share information about our resources (such as the OAI Protocol for Metadata Harvesting and Z39.50) would enable us to create a 'one stop' search portal for our users. It's unlikely we'll every really only have one search destination, but fewer destinations is certainly a reasonable goal.

• Ability to create customized searches of specific domains, subjects, as well as interdisciplinary searches. Perhaps more useful than the 'one stop' approach is how sharing protocols have enabled building of systems that allow users to find resources in specialized disciplines or formats, such as the Sheet Music Consortium and the Western Waters Digital Library.

Sharing metadata also has benefits for those organizations willing to share:

• Increases exposure of collections and broadens user base.
• Broadens user base
• Potentially adds collaboration opportunities
• We can no longer assume that users will come through the front door, sharing metadata gets us 'in the flow'. See Lorcan Dempsey's In the Flow blog entry.

Are there other benefits for users or our institutions to sharing?

3. An abridged history of metadata sharing

Some milestones in the history of sharing metadata:

• 1877 - Standardization of size of Catalog Cards
• 1901 - Library of Congress card distribution program
• 1967 - Museum Computer Network
• 1967 - Anglo-American Cataloging Rules (AACR)
• 1968 - MAdine Readable Catalog (MARC)
• 1969 - International Standard for Bibliographic Description (ISBD)
• 1971 - OCLC Union Catalog
• 1987 - Z39.50
• 1995 - Dublin Core
• 1996 - VRACore
• 1999 - First version of RSS developed
• 2001 - Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH)
• 2001 - Darwin Core
• 2004 - Search and Retrieve via URL (SRU)
• 2004 - Describing Archives: A Content Standard (DACS) approved by the Society of American Archivists
• 2005 - OpenSearch
• 2005 - CDWA Lite introduced
• 2005 - PBCore introduced
• 2006 - Cataloging Cultural Objects (CCO)

As this timeline shows, sharing information about our collections is far from a new idea. In fact, sharing information about what we hold in our collections has been an important activity since the founding of many cultural heritage organizations. Today we face new challenges about the methods we use to share, but they are not unlike those faced by our predecessors.

With the advent of printing technologies, many museums and libraries began publication of bibliographies and catalogues of their collections in the seventeenth-century. These catalogues often became important reference works for other repositories that allowed scholars to learn about collections located elsewhere.

For the library community, methods of sharing catalog records also have a more recent lineage. In 1877 Melvil Dewey standardized dimensions for catalog cards and established a service bureau to supply libraries with catalog cases (the hardware) and cards (software). In 1901 the Library of Congress was able to take advantage of this standardization by widely distributing catalog cards to other libraries based on its acquisitions.
With the advent of automation in the late 1960s and 70s, sharing of computer coded records was now possible. The Library of Congress was again at the forefront of this development when it received funding to develop the MAchine Readable Catalog (MARC) record format between 1966 and 1967. With approval from national and international library community the MARC format laid the foundations for collaborative cataloging efforts and distribution of MARC records on tape in 1969. By the early 1970s statewide collaboratives such as NELINET and the Ohio Computer Library Consortium (later the Online Computer Library Center - OCLC) developed online networks for shared cataloging and the production of catalog cards and eventually online access catalogs for public use.

Similar efforts were underway in the art museum community with the formation of the Museum Computer Network in 1967 and the application of computers to archeological and natural science collections. While these efforts did not result in a single large-scale catalog of museum collections, the profession had begun to explore the possibilities of creating shared networks. Natural history museums, in particular, have participated in growing repositories of biodiversity information.

With the growth of information resources on the Internet, OCLC and other organizations created the Dublin Core in 1995 as a lightweight format for describing online content. Other communities extended the concept of core metadata elements by creating profiles that were appropriate for their resources. Examples include the VRACore, CDWA Lite and the Darwin Core, as well as other Dublin Core Application Profile communities.

As users increasingly relied on mainstream search engines to find content on the Web, many content owners were concerned that their content was not being found. The Open Archives Initiative - Protocol for Metadata Harvesting (OAI-PMH) was released in 2001 in order to expose the "deep web" and provide a mechanism for interoperability and dissemination of content.

Sharing information about resources relies not only on data format standards, but also on content standards. The Joint Steering Committee for the Ango-American Cataloging Rules (AACR) has created a timeline outlining the development of bibliographic cataloging rules. Like data standards, other communities frequently adopted bibliographic rules for description for their own community, examples include Archives, Personal Papers, and Manuscripts (APPM), Describing Archives: A Content Standard (DACS), and Cataloging Cultural Objects (CCO).

Next we will cover two different ways that metadata can be shared: federated search and data aggregation. There are other ways that metadata can be shared (via RSS feeds for example), but we will cover more of these at the end of the course.

4. Federated Searching

In federated search methods, a users query is broadcast to a number of different databases that are registered with the system. Each individual database sends results back to the query service, which collates the results for the user. Z39.50, SRU and other metasearch technologies rely on federation for their searches.

Pros:

- Real-time, up-to-date search results
- No need to store large amounts of data

Cons:

- Speed of results relies on speed of the slowest responder
- Limited ability to modify and augment incoming metadata
- Results depend on remote services being available
5. Data Aggregation

Aggregation services routinely collect information from identified resources and create a new aggregated data source for users to search. Because aggregations provide a surrogate set of data for users, it is not dependent on multiple remote services and aggregation managers can provide value-added enhancements based on the entire set of data collected. The Open Archives Initiative-Protocol for Metadata Harvesting (OAI-PMH) and large search engines such as Google Scholar (and Google) rely on aggregated search.

Pros:

- Faster response time from single data source
- Able to normalize and augment data from heterogeneous sources

Cons:

- Data is only as up-to-date as the last collection activity
- Significant work required to transform, normalize or augment collected metadata
6. Safe Assumptions about Sharing Metadata

The metadata sharing landscape is constantly changing. However, given the current environment, there are a few assumptions it is safe to make, to guide you in preparing your metadata for sharing.

- Users often discover material through shared records, not through your front door. For example, many institutional repositories find that most of their downloads are coming through places like Google or through aggregations like OAIster. The following diagram gives a hint of the current proliferation of metadata from one institution out to multiple places.

- Users don't know about your collection or don't remember (or don't want) to search it separately. While the proliferation of digital content from museums, archives, libraries, and other organizations has been a boon to most of our users, the accompanying proliferation of interfaces and silos of content is a problem. Consider the example of a teacher collecting some primary source material on immigration in the early twentieth century. If we're lucky she might come to the reference desk at her local library, but more likely she'll turn to Google or to a teacher's guide to content. She might go to a high profile site like the Library of Congress's American Memory site or maybe even to something like the California Digital Library's Calisphere (which is, of course, an aggregation). But more than likely she'll miss the dozens and dozens of lower profile and smaller collections that contain relevant content.

- Shared records usually lead users to your local environment where you can provide the full context for the resource. This means that the URLs you provide in the shared record should be persistent!

- Because users typically enter through these 'deep links', they may bypass introductory information that provides the larger context for a collection. Many repositories of digital content were built with the assumption that users would arrive at content through a front door or a welcome area that provided context or general information about a collection. In reality, users are coming in - or expect to come in - directly at the resource. This means that the page the user arrives at should contain contextual and navigational information so a user can easily and quickly orient themselves to the new environment and understand quickly what the collection is. We'll talk more about context later in the course.

Are there other assumptions about the current environment of sharing records that should be listed here?

Next module: A Look at Sharing: Aggregators
Metadata for You & Me is a collaboration between the University of Illinois Library and Indiana University. This project is made possible by a grant from the U.S. Institute of Museum and Library Services (IMLS) through the Laura Bush 21st Century Librarian Program.

Metadata for You & Me Feed