Planning for Metadata: the Quick Tour

Jody Perkins
Miami University Libraries

Follow this and additional works at: http://docs.lib.purdue.edu/atg

Part of the Library and Information Science Commons

Recommended Citation
DOI: http://dx.doi.org/10.7771/2380-176X.5227

This document has been made available through Purdue e-Pubs, a service of the Purdue University Libraries. Please contact epubs@purdue.edu for additional information.
The Articles in This Issue

Collection development increasingly features digitization of hidden resources, unique collections, and rare materials. But digitization involves more than just scanning items in some Web-friendly format. It involves metadata, the key to making a digital collection easily searchable, compatible with local, consortial, and even global systems — and accessible into the future.

Contributors to this issue of Against the Grain emphasize the importance of coordinating with catalogers from the beginning of any digitization initiative. Doing so will save much backtracking and associated expense later. Thus collection decision makers and metadata catalogers/specialists should continue to forge strong relationships to bring the best product to the user.

Traditionally, collections librarians have chosen materials represented in the catalog by a MARC record. Raised on the ISBDs, firmly married to the content standard AACR2, and happily housed in your local ILS, MARC is a well established schema. Those collecting standard resources rarely had to wonder, “How will we provide access?” When selecting resources for digitization, however, collection development principles must be augmented by answers to a host of questions. How will digital assets be preserved? What schema will be used to describe them? What system will house them?

In this issue, we hope to answer these questions and others. First of all, Jody Perkins will give a conспектus of the essential matters that planners of a digital project need to take into consideration. Her excellent checklist includes sixteen vital points to consider when evaluating a collection. She discusses metadata design, choosing schemas and standards, and documenting decisions through the use of a data dictionary.

Reflecting further on schema selection, Jeffrey Beall enumerates twelve points of comparison to help one decide which of the many schemas available best suits one’s digital project. He addresses such concerns as interoperability, granularity, proven success, and level of community or domain specificity.

Next, a pair of case studies: James Bradley discusses the efficacy, for a digital image collection, of CONTENTdm and Dublin Core; and Jen Wolfe and Mark F. Anderson review the difficulties and decision-making involved in opting for DigiTools and METS to provide access to a collection of science fiction fanzines. These case studies cover crosswalking, the viability of existing schemes, copyright issues, and decisions about the depth and extent of metadata needed.

Finally, Arwen Hutt, Trish Rose-Sandler, and Bradley D. Westbrook share one library community’s successful approach to metadata preservation, a hot topic that the digital library community must concern itself with, especially complex problems of long-term usability. In their article, they describe creation of a digital asset management system that, ingeniously wrapping MODS in METS, converts different types of metadata from many diverse projects into one interoperable and manageable schema.

These essays offer a wealth of insight into some of the most important electronic resources issues currently facing collection development. As we digitize our unique holdings, preserve items in jeopardy, or offer our most popular collections to the broadest user base, we would do well to keep in mind that the important decisions are made at the beginning of the collection digitization project and are mission critical to current and future plans for interoperability.
Planning for Metadata ...
from page 20

has been written on everything from emerging standards and new applications to research on automated processes, interoperability and measures of quality. Unfortunately there is still relatively little information available that gets at the everyday issues many implementers or potential implementers are likely to face in the field. What follows is based largely on my own experience, the experience of colleagues, what I’ve learned from workshops and to a much lesser extent from an emerging practice based literature.

Since the quality and interoperability of metadata has a direct bearing on access to, and in the case of primary source materials (photographs, manuscripts, raw datasets etc.) comprehension of digitized resources, and since its creation is often the most time-consuming and thus most costly component of a digital library project, it’s critical that anyone proposing such a project have at least a fundamental understanding of what’s involved. When metadata is an afterthought, implementers can be forced into making ad-hoc decisions resulting in poor quality non-interoperable metadata. However, when metadata is part of a thoughtful planning process, obstacles can be anticipated and trade-offs either managed or avoided.

Planning

Metadata creation requires planning because for the most part it is carried out as part of a project and projects vary from one another in ways that can’t always be anticipated. Among the many project variables are the types of collections, hardware, software, required expertise, project team members, metadata creators and the source and extent of funding.

Planning for metadata is only one part of the larger digital project planning process. Project planning typically includes: clarifying the purpose and establishing the goals of the project, identifying stakeholders, planning for scanning and metadata, allocating resources, and designing workflow. Deliverables and criteria for a final evaluation should be specified where possible. All the decisions to follow should be made with project outcomes in mind.

In some cases a metadata specialist may be involved in setting project goals and at other times there will already be established. In either case it’s necessary to determine whether there is a match between the resources available and the extent to which you will be able to create metadata that complies with current best practices.

Metadata Design

Interpretation and negotiation of many different and at times conflicting standards is often required. One must also remain vigilant about interoperability issues and be sensitive to any unique requirements of the project in question. Compromises are inevitable and knowing where and how to cut corners without sacrificing quality is a vital part of the process. Final decisions will be based on the goals established for a specific project as well as the priorities of your particular institution or department at the time the project is underway.

Metadata design, the way I’ve come to define it, includes the following: an evaluation of project collection(s) and any associated metadata, a review of current standards, a review of other relevant collections, and documentation of decisions related to the selection and implementation of standards. The preparation of a crosswalk may also be required when migrating legacy data to a new schema.

Critical Decision Points — Metadata Planning at the Strategic Level

As a pre-requisite to making many of the smaller decisions that are part of metadata design at the project level other more critical decisions need to be made with regard to interoperability compliance, measures of quality and the breadth and depth of metadata. It’s a good idea to establish a set of minimal requirements that every project must meet. However, a discussion of all the issues that might best be addressed as part of an overall digital library program plan is beyond the scope of this discussion (see Agnew, 2003).

Evaluation of Project Collections

An evaluation of project collections is an important first step in the metadata planning process. It includes a review of representative items as well as any existing metadata or other information sources that could be converted into metadata. A thorough review of this type makes it possible to understand not only the content but also the context of the collection and how it relates to the desired project outcomes. Such an understanding is fundamental to the selection of appropriate content standards, schemas, controlled vocabularies and related value spaces and is also critical in establishing an efficient workflow, selecting project team members and guiding the training of metadata creators. For a more detailed treatment of issues to consider please refer to the collection evaluation checklist at the end of this article.

continued on page 24

Jody Perkins

Metadata Librarian, Miami University Libraries
306 King Library, Oxford, OH 45056
Phone: (513) 529-0135    Fax: (513) 529-1719

BORN: Cleveland, Ohio.
EARLY LIFE: Before I discovered libraries, museums, poetry, and politics, I spent most of my time reading, hanging out with horses, going to church, making art or traveling with my family.
FAMILY: My late father was a piano tuner/technician and instructor of the same, former Regional VP of the Piano Technicians Guild and an amateur tenor soloist. My mother taught dance for four years.
EDUCATION: BA Sociology, University of Akron (Alpha Kappa Delta); MLS, Kent State University (Beta Phi Mu).
FIRST JOB: Copy prep/graphic design, John S. Swift Publishers.
FIRST LIBRARY RELATED JOB: Librarian, Akron Art Museum.
IN MY SPARE TIME I LIKE TO: Read, ride and do yoga.
FAVORITE BOOKS: Wisdom, Information and Wonder: What is knowledge for? by Mary Midgley and Ambient Findability by Peter Morville.
HOW/WHERE DO I SEE THE INDUSTRY IN FIVE YEARS: In spite of our best efforts I believe the vast majority of digital information will continue to exist in a state of chaos, at least in the public sector. The private sector particularly the fields of finance and medicine may be the exception. Pockets of order (i.e., reliable access) will exist for those disciplines/industries that need access to the most comprehensive, accurate, and timely resources available. I tend to agree with the folks who have argued that the focus will begin to shift toward services that will increase metadata quality and add value at the post-creation stage. However, I also believe there are certain aspects of quality that can only be implemented at the point of creation.

Mary Midgley
The primary goal of reviewing existing metadata is to determine to what extent it’s possible to leverage existing information through mapping or other processes. The extent (how much), granularity (at what level of detail), quality (consistency, authority, reliability etc.) and format (paper files, spreadsheets, databases, finding aids, MARC records, etc.) of any existing metadata will have a direct impact on workflow and turnaround time. Even when repurposing existing metadata it is necessary to determine how much editing will be needed and whether additional elements might be required.

Standards Review and Environmental Scan

Like anything that requires a certain level of compatibility between systems, metadata is standards driven. Standards provide the foundation for interoperability. Anyone who wants to increase access to their digital collections — whether it’s through directly contributing to collaborative projects or indirectly through harvesting needs to be aware of a variety of metadata related standards. Relevant standards and related guidelines include: metadata schemas, content standards, controlled value schemes, best practices and application profiles. For a selective list of examples please see the standards referenced at the end of this article.

In the open access environment in which digital libraries operate it’s necessary to ensure that all collections meet at least the minimal requirements for interoperability. To a large extent metadata planning is about interoperability compliance. Interoperability has a direct impact on resource discovery both within and across networks. There are three types of interoperability to keep in mind: semantic (either at the element or value level), syntactic and structural, all of which may occur at many different levels (local, consortial, community of practice, etc.) at the same time depending on the project. For that reason it’s often prudent to do a quick scan of your environment in order to make sure you comply with the requirements of any digital repositories where you anticipate contributing metadata as well as reviewing metadata for existing digital collections within the community of practice most closely aligned with the project. (Shreeves et al)

Documenting Decisions

Unless you already have a system in place that mandates the use of certain standards you will need to make decisions about encoding or mark-up, metadata schema(s), which elements to use within a schema, content standards, controlled values and the addition of local or project specific elements. Many best practices and application profiles developed by consortia or other communities of practice have been designed with a certain level of flexibility in order to accommodate legacy data or other special needs of individual contributors. Their guidelines are typically suggestive rather than prescriptive and as a result many final decisions rest with local implementers. When that’s the case, decisions should be documented. In addition, a variety of local needs are not addressed by standards (e.g., whether or not a specific element is displayed or suppressed, use of customized labels and taxonomies, the addition of fields for tracking workflow and managing objects, etc.). A data dictionary is one method of documenting such decisions. Data dictionaries can be used in different contexts for different purposes. Please note that I am using the term in the informal sense (see Data dictionary — Wikipedia http://en.wikipedia.org/wiki/Data_dictionary).

A data dictionary for documenting metadata consists of the attributes and associated specifications for each element in the element set. Establishing a core set of OAI-PMH (Open Archives Initiative — Protocol for Metadata Harvesting) compliant elements that support interoperability is an important first step. Selection of additional elements is made with consideration given to the specific needs of the project in question. The data dictionaries I create typically include information related to the following element attributes: element name, field label, element definition, element obligation (mandatory or optional), element occurrence (repeatable or not), names of or direct links to controlled lists of values, data mapping information, responsibility for data entry or data source, input guidelines, data type, and display and search specifications. There’s really no official standard for developing data dictionaries though there are sections in ISO 11179 that are applicable.

Although data dictionary development is not included in any best practice documents I’ve reviewed it can be useful for a number of reasons:

- It can guide you through the decisions that need to be made for a specific project.
- It documents those decisions and the revisions that are made as you go along.
- It becomes part of the permanent project documentation, which can then be used with minor alterations for similar projects in the future.
- As a form of meta-metadata it supports metadata preservation. When it becomes necessary to move the metadata to a different system or to a different standard, all the information needed to map the data will be there.
- The completed data dictionary can also be used to guide the creation of data entry templates, design workflow, plan training and monitor quality as the metadata creation process progresses.

Concluding Thoughts

Metadata can be as simple or as complex as you want or need it to be. The key is in knowing precisely what outcomes are desired. Regardless of the level of complexity complying with standards and best practices for interoperability, quality and preservation is critical. The surest way to meet project outcomes and enforce standards is to adequately plan for and document the metadata component of digital library projects.

Sources Consulted


Standards

General Guidelines


Best Practices and Application Profiles

CDP Metadata Working Group. Dublin Core Metadata Best Practices, Version 2.1.1. Denver, CO: Collaborative Digitization Pro-


Metadata Schemas


Content Standards


Controlled Value Schemes


Data Dictionaries


COLLECTION EVALUATION CHECKLIST FOR METADATA PLANNING

What follows is a checklist of issues that I would address as part of a collection review. Some information can be derived from the direct inspection of items in the collection, while other details may require consultation with the collection’s administrator and other project participants.

1. How is “the collection” defined for the purposes of the project in question? What are its physical and conceptual boundaries?

2. How is an “item” defined for the purpose of metadata creation? Are the items compound objects having more than one part? If so, how will the parts be represented in the digital surrogate and documented in the metadata?

3. What’s the number of items in the collection? Is the collection finite or is it open-ended? If it’s open-ended what is known about the type of items that might be added in the future?

4. What is the origin of the collection and what is the history of ownership? To what extent does it need to be represented in the metadata?

5. Overall has the order of the original owner been maintained? If so, does it add context that should be preserved and incorporated into the metadata scheme? How will the order (or lack thereof) impact workflow?

6. Does the place of an item in the overall order of a collection carry any significance that should not be lost?

7. Are there existing or potential groupings or other patterns that might lend themselves to the inheritance of metadata for one or more large groups of items?

8. Is the subject matter heterogeneous or homogeneous? If it’s heterogeneous how does the community of practice for the purpose of researching existing metadata standards? If it’s heterogeneous what are the interoperability issues?

9. Are the formats represented heterogeneous or homogeneous? If they are heterogeneous what community of practice can be identified for the purpose of researching existing metadata standards? If they are heterogeneous what are the interoperability issues?

10. What attributes and other details can be identified to aid discovery or that need to be documented?

11. Are there attributes that share common values across all items in the collection that can be added as constant data or as part of a batch process?

12. Are there attributes where values would vary? Of those, which could make use of existing controlled lists of values and which might require a customized list?

13. Does the collection contain duplicates that could share the same metadata or alternately be passed over in the digitizing process?

14. Are the items identified located on site? In one place? and what impact might that have on workflow? Will metadata creators have access to the originals or will the cataloging need to be done from the digital surrogates?

15. Does the condition of the collection require special handling or need to be documented in the metadata? Is the digitization being done as part of a preservation strategy which will require additional metadata?

16. Who are the current primary users of the collection? Are there other potential primary users? What community of practice is most closely aligned with the users identified? Other than users (students and faculty) are there other stakeholders (e.g. collection administrators, subject specialists) that need to be taken into consideration when documenting the collection or creating access points?

Rumors

from page 16

marketing director at the University of Akron Press as well as Head of Acquisitions at the

Library — has promised to send us a report of the meeting. Stay tuned.

I just got Bryan M. Carson’s new book, The Law of Libraries and Archives (Scarecrow, 2007). Wow and Gosh! There are 12

continued on page 36