Against the Grain

Volume 19 | Issue 1 Article 7

February 2007

Discrete Criteria for Selecting and Comparing Metadata Schemes

Jeffrey Beall
University of Colorado at Denver and Health Sciences Center, Jeffrey.Beall@ucdenver.edu

Discrete Criteria for Selecting and Comparing Metadata Schemes

by **Jeffrey Beall** (Catalog Librarian / Assistant Professor, Auraria Library, University of Colorado at Denver and Health Sciences Center, Downtown Denver Campus, 1100 Lawrence St., Denver, CO 80204; Phone: 303-556-5936) <jeffrey.beall@cudenver.edu>

This article lists and describes the twelve chief points of comparison among the different metadata schemes available. Before implementing a metadata scheme, digital libraries or individual domains or organizations must decide on which one to use. Knowing the chief points of comparison among the schemes available can simplify this selection process. Some organizations have chosen to create new, home-grown schemes instead of implementing an existing one, when an existing scheme may have been adequate for their needs. However, an organization planning to create a new metadata scheme could also use the points described here as a guide for developing the specifications of the new scheme.

Knowing the points of comparison among existing metadata schemes is also valuable when an institution is evaluating the effectiveness of a scheme already in use. Because the metadata scheme landscape is still relatively new, and because some schemes are likely to increase or decrease in popularity or effectiveness in response to changes in information technology, libraries and organizations ought to regularly examine the schemes they have in use to determine whether the scheme is still meeting their needs. Libraries and organizations should use the criteria we describe here in terms of the needs of their particular application of the metadata, that is, the needs of the library or organization and the needs of the users of the data the metadata describes. The following is a list and description of twelve criteria for comparing metadata schemes.

1. Granularity and Formats of Description

Metadata schemes differ in the amount of specificity they provide for as well as their ability to describe data in different formats. For example, some schemes provide a way of differentiating among different types of authors (i.e., **MARC**, **VRA Core**), yet others do not (i.e., **Dublin Core**). Different types of authors include personal authors, corporate authors, and conference authors. Here the specificity is also often referred to as granularity.[1]

Schemes also differ in their ability to describe data that comes in different formats. For example, some schemes may only be designed to describe data in electronic form, and others can describe data in any form. The **MPEG-7** metadata scheme is used to describe multimedia, including digital photographs and videos. It is not designed for textual objects and would be a poor choice for this type of data.

2. Level of Connection to Content Standards

Some schemes, like **MARC**, are closely connected to content standards. **MARC** is often closely associated with the **Anglo-American Cataloguing Rules** and with the **Library of Congress Subject Headings**. Other schemes, **Dublin Core (DC)** for example, are much more autonomous from content standards, so selecting such a scheme may also involve the additional task of selecting content standards. On the other hand, selecting a scheme with a strong connection to a particular content standard usually means having also to adopt the content standards, ontologies, etc. that are associated with it.

Also, schemes may differ in their ability to encode different character sets, such as non-Roman scripts and Unicode, but this ability may also depend on the computer system being used to encode the data. Content standard selection is important because it can affect the ability to crosswalk data from one database into another.

3. Availability of Searching Systems

Metadata systems sometimes include software or applications that provide a search interface for metadata. Integrated library systems (ILSs) are an example of a system that searches **MARC** metadata. This can be problematic for less popular metadata schemes, as there is a lack of developed systems available to fully exploit the metadata and create a search platform for it.

Another aspect worth considering is how well the system can create metadata. Potential implementers should determine whether the scheme in question has systems available for metadata creation by humans or computers. An example of this is integrated library systems that have the functionality to create **MARC** records. Similarly, search systems differ in their ability to store and manipulate data created in a particular scheme. For example, one system could accommodate both **MARC** and **DC** data, but another system could be designed only to handle **DC** data.

The next few years will likely see a greater development of digital library management systems (DLMSs)[2] that will differ in their ability to accommodate different metadata schemes. These systems will be similar to integrated library systems but will be designed specifically for digital libraries. The process of selecting a particular metadata scheme will need to take into account the availability of systems for a given scheme, as well as the quality of each system.

4. Level of Community or Domain Specificity

Some metadata schemes are created for the specific needs of an individual community or domain. For example, the aforementioned **MPEG-7** scheme is designed for multimedia. The **ONIX** scheme is designed for the book trade industry, which is also referred to as the publishing domain. Other schemes are general in design, and can accommodate metadata from most fields of study. The desire for community specificity has led to an abundance of metadata schemes, but a scheme designed for a particular domain will likely be very efficient at meeting the metadata requirements of that domain.

Further, some metadata schemes are proprietary. That is to say, using the scheme or elements associated with the scheme requires membership in or payments to an organization. One example is the **Digital Object Identifier**, or **DOI**, system.

5. Interoperability

Interoperability encompasses several things. First, it describes how well-suited a scheme is for crosswalking data into other schemes. More practically, it involves whether those systems designers have created mappings and whether they are available. Designers have developed crosswalks from most of the more popular schemes to other schemes. For example, there is a crosswalk from **Dublin Core** to **EAD**. The **Getty Museum** has a crosswalk between eleven different standards on its Website.[3]

Interoperability also includes metadata harvesting. A scheme with high interoperability enables the harvesting and meta-searching of metadata encoded in it. To some degree, interoperability is related to a scheme's popularity: the more popular and widely used a scheme is, the more likely it is to have crosswalks to other schemes and harvesting standards.

6. Proven Success, Reputation, Popularity

Success and popularity of a scheme often weigh heavily for users deciding whether or not to adopt it. Users will likely prefer a scheme that has successfully left beta testing and has had several documented, successful implementations.

7. Amount of Training Required

Those selecting a scheme will need to take into account the amount of training that individuals will need to become proficient in encoding metadata in the scheme. For schemes that are closely connected to content standards, this training will also need to take into account the amount of training needed to gain proficiency in those standards, if necessary. There is likely a positive correlation between the amount of training needed to master a scheme and the richness of description it provides.

8. Viability of the Organization behind the Scheme

The stability and vibrancy of the organizations behind metadata schemes are crucial to their success. Potential implementers of a scheme should investigate the organization behind it to ensure that it keeps the scheme current with the latest developments and user needs. A related factor worth investigating is how open the organization is to receiving input

continued on page 30

Discrete Criteria ...
from page 28

and suggestions from implementers and users. Also, implementers will need to consider the amount, quality, and currency of documentation that is available for a particular scheme. Further, the availability of the documentation in other languages may be an issue if the implementers of the scheme use these other languages.

9. Ability of the Scheme to Handle a Particular Metadata Function

Metadata serves different purposes, from discovery and rights management to recording preservation data. But not all schemes are able to serve all of these various functions. Before implementing a scheme, users need to determine exactly what functions they want their metadata to serve, and they should then select a scheme that adequately handles these functions. Of course, some schemes can perform multiple functions, but potential users of the scheme must evaluate how well a scheme handles each function, for a scheme could perform well in one required function but poorly in others.

With the increased use and popularity of federated search engines, the de-duplication of individual metadata records has become crucial. A federated search engine may have difficulty in identifying duplicates even when all the records are in the same metadata format or scheme. This is because the records may have originated from various sources, leading to data that is slightly different in each. Federated search engines have an even tougher time in de-duplication when the records involved are encoded in different metadata schemes. Some schemes provide for unique identifiers, such as document numbers, ISBNs, etc. that help systems in the de-duplication process. So in any metadata scheme application that will involve de-duplication, it is important to evaluate how well each scheme accommodates automated de-duplication.

It is also useful to examine metadata schemes in terms of access versus description. Access involves metadata elements that help users discover or find desired data, including elements such as author, subject headings, etc. Description involves metadata that provides details about the characteristics of an individual resource, such as a summary or description of the number of pages in the resource. Sometimes we use a single metadata element, such as title, for both access and description. So when examining a particular metadata scheme, implementers should consider how each handles description and access and to what degree they are combined or separated in a scheme.

10. Adaptability of the Scheme to Local Needs

This relates to community specificity but is different in that some metadata schemes can be changed at the local level, such as by adding certain new fields or tags. Sometimes a modified scheme is also called a particular "flavor" of a scheme. For example, the **Collaborative Digitization Program** has created the **Western States Dublin Core**, which is a customized implementation of **Dublin Core**. Schemes that are more adaptable will have mechanisms for extensibility of the data elements so that they can be extended to better meet local needs.

A particular implementation of a metadata scheme (or elements from more than one scheme) is called an application profile. According to the **Dublin Core** glossary, an application profile is:

> A set of metadata elements, policies, and guidelines defined for a particular application. The elements may be from one or more element sets, thus allowing a given application to meet its functional requirements by using metadata from several element sets including locally defined sets. For example, a given application might choose a subset of the **Dublin Core** that meets its needs, or may include elements from the **Dublin Core**, another element set, and several locally defined elements, all combined in a single schema. An application profile is not complete without documentation that defines the policies and best practices appropriate to the application.[4]

A related concept to extensibility is metadata scheme modularity. This refers to how well different schemes lend themselves to having only particular elements being used and combined with elements from other schemes in a given metadata implementation. According to **Duval**, et al., "In a modular metadata world, data elements from different schemas as well as vocabularies and other building blocks can be combined in a syntactically and semantically interoperable way."[5]

11. Scalability

Scalability refers to how large a database of metadata the scheme and its retrieval system can handle successfully. For example, a scheme with only a few elements of description is not as scalable as a system with many elements because when one has millions of records using a "few-element" scheme, it becomes harder to generate precise search results. In general, the richer the description a scheme provides for, the more scalable it is. Also, the level of description or granularity within a particular element of description can also make a scheme more or less scalable. For example, a scheme that provides for precise geographical tagging by latitude and longitude is more scalable than a scheme that only allows for a single textual annotation of a geographical location.

12. Surrogacy

Surrogacy relates only to digital objects and describes whether the metadata is embedded in the object it describes or exists separately from it in a searchable database. **Howarth**[6]

continued on page 31

against the grain people profile

Catalog Librarian / Assistant Professor, Auraria Library
University of Colorado at Denver and Health Sciences Center
Downtown Denver Campus, 1100 Lawrence St., Denver, CO 80204
Phone: (303) 556-5936 <jeffrey.beall@cudenver.edu>

Jeffrey Beall

BORN & LIVED: Born in California; lived in four states and four countries.

EARLY LIFE: All over California.

FAMILY: Single.

EDUCATION: **California State University, Northridge**, 1982, B.A. Spanish. **Oklahoma State University**, 1987, M.A. English. **University of North Carolina**, 1990, M.S.L.S Library Science.

FIRST JOB: Cataloger, **Harvard University**.

IN MY SPARE TIME I LIKE TO: Take pictures and post them on my **Flickr** account; bicycle riding; read and observe astronomy and cosmology.

PET PEEVES/WHAT MAKES ME MAD: Typos.

MOST MEANINGFUL CAREER ACHIEVEMENT: Being invited to speak at **Library and Archives Canada** (in Sept. 2005).

GOAL I HOPE TO ACHIEVE FIVE YEARS FROM NOW: Saving cataloging from destruction by short-sighted library administrators.

HOW/WHERE DO I SEE THE INDUSTRY IN FIVE YEARS: In 2012, libraries will be re-thinking their decision to take away resources from metadata creation. Following a shortsighted move to a reliance on keyword searching, libraries in 2012 will be faced with general chaos in information discovery. They will finally understand the value and importance of rich metadata and will begin to re-invest resources in this important library function.

Discrete Criteria ...
from page 30

describes embedded metadata in the following way:

> "In general, a distinction can be made between simple format metadata — such as that represented in the syntax of a mark-up language (e.g., XML; HTML; SGML), and embedded within the structure of the digital object — and structured rich format metadata. For the former, Web crawlers or "bots" can harvest the specified metatags (e.g., <Title>) to extract particular values…"

Of course, some schemes can have the metadata exist within the data it describes and also as a surrogate separate from it. For example, a Web page can have its metadata embedded within its meta tags and also copied to a separate external database. Descriptive or technical metadata can also be embedded in image files. Metadata that is separate from the item it describes and that is created by someone other than the item's author is called third-party metadata.

Conclusion

As the number of metadata schemes available continues to grow, digital libraries will need clear points of comparison for selecting and evaluating from among the schemes available. The first step in selecting a metadata scheme is determining the local needs, that is, what functions the metadata needs to serve. The points listed here can serve as a comprehensive set of criteria for making an implementation decision or for evaluating an existing metadata scheme implementation.

Appendix 1

Appendix 1: A sample grid for use in comparing metadata schemes for a particular implementation. The criteria are in the left column. The five columns on the right are for five major metadata schemes. The notes in the boxes for each scheme show possible descriptions of each criterion for each of the five schemes. The notes represent the author's opinion and are for illustrative purposes only.

Scheme→ / Criteria ↓	MARC	Dublin Core	MODS	VRA	EAD
1. Granularity and formats of description	Rich granularity; can describe all formats	Most applications have low granularity; most formats ok	Rich granularity; can describe all formats	Rich granularity; describes visual resources	Rich granularity; for encoding archival finding aids
2. Level of connection to content standards	High, connected to AACR2 and LCSH, but flexible	Low	High, connected to AACR2 and LCSH, but flexible	High, connected to CDWA, ULAN, CCO, AAT	High, connected to ISAD(G) and DACS
3. Availability of searching systems	Many commercial systems available	Few, generally Web based	Few	Few	Few
4. Level of community or domain specificity	Associated with the library domain	Library and digital repository	Associated with the library domain	Museum, library and digital repository	Archives, special collections, libraries and digital repositories
5. Interoperability	Highly interoperable	Medium; granularity may be lost when crosswalking into Dublin Core	In theory, high, but not proven yet	Medium; some elements may be lost or merged in crosswalking	Medium; some structure lost in crosswalking
6. Proven success, reputation, popularity	Probably most successful scheme of all	Mixed	Too few implementations to determine	Website lists 61 large successful implementations	Proven success
7. Amount of training required	High, especially for creation of new records	Low to medium	Medium	High	High
8. Viability of the organization behind the scheme	Very stable but subject to financial pressure	DCMI; stable for a decade	Strong (Library of Congress)	VRA; stable since 1982 with 600 members	Strong; supported by Library of Congress and SAA
9. Ability to handle a particular metadata function	Can handle most	Can handle the main ones, e.g. discovery	Can handle most	Can handle most; focus is on description	Can handle most
10. Adaptability of the scheme to local needs	Highly adaptable, but this lessens interoperability	Highly adaptable	Highly adaptable, but this lessens interoperability	Highly adaptable	Highly adaptable
11. Scalability	Highly scalable	Not proven in very large databases	Also not proven, but should be high	Proven in some large databases	Proven in some large databases
12. Surrogacy	Generally surrogate	Surrogate or embedded	Surrogate or embedded	Generally surrogate	Generally surrogate

continued on page 32

Zen and the Digital Collection Librarian

by **James A. Bradley** (Head of Metadata and Digital Initiatives, Ball State University, University Libraries, BL-025, Ball State University, Muncie, IN 47306: Phone: 765-285-5718) <jabradley2@bsu.edu>

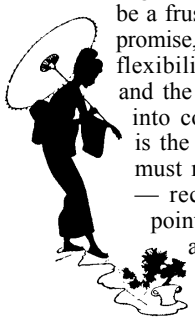
"The container tends to shape the contained."

It sounds a bit like an eastern proverb — and, to be completely honest, I'm not certain that I haven't unconsciously borrowed the phrase from one of the many poorly-dubbed Kung Fu films I indulge in from time to time.

Whatever the origins, the concept is simple: Flexible matter immediately assumes the shape of whatever you pour it into; and even a somewhat rigid object will, over time, succumb to the contours of its packaging. By the same token, a rigid item will simply break if forced into a container that is too foreign or restrictive.

I've found this maxim to be a true and useful analogy in the planning of digital collections.

Bringing an existing "real world" collection, and its accompanying metadata, across the digital threshold can sometimes be a frustrating process — full of promise, but also compromise. The flexibility of both the container and the contained must be taken into consideration; however, it is the collections librarian who must remain the most willowy — recognizing the shattering points of both and finding an appropriate fit.



Such was the case of the **Ball State University Architecture Image Collection**.

The challenge to **Ball State University Library**'s new Digital Initiatives program was to migrate the visual resources of the Architecture Library into a single online environment that would facilitate remote access, advanced searching capabilities, and image delivery at a resolution suitable for research and classroom instruction within the College of Architecture and Planning (CAP).

The first step in the conversion process was to assess and gather the characteristics of the materials to be digitized:

- Approximately 120,000 35mm slides
- Local call number for access purposes
- Group level **MARC** records that gather individual slides according to location or site.

The next step was to consider the characteristics of the desired online collection:

- Slides must be scanned and stored in accordance with archival standards.
- Derivative images must be created for online delivery.
- The "front end" metadata must be user-friendly, containing data fields and categorizations that CAP students and faculty would recognize.
- The "back end" metadata must conform to internationally recognized metadata standards and be suitable for Open Archives Initiative (OAI) harvesters.
- The online collection must be made available as widely as copyright will allow — so that outside educators and the general public may also utilize the collection.

With the above survey of existing materials, and list of collection goals we began our planning the collection and drafting workflows.

Content Management

The first task of any digital collection is to determine if one should develop or purchase a content management system (CMS) to house it. Fortunately, this decision had already been made: prior to the beginning of this project, **Ball State University Libraries** had purchased **CONTENTdm** to form the base of all collections in our **Digital Media Repository**.

As with any turnkey system, **CONTENTdm** has the disadvantage of already being a fully formed container. Homegrown systems are far more advantageous in this regard, and can be developed with a specific collection in mind for a tailor-made fit. This being said, however, **CONTENTdm** is a surprisingly flexible container, and has the added advantage of being ready to go practically out of the box.

Metadata

With our CMS in hand, we set about determining how to utilize the existing metadata.

As previously stated, the 35mm slides were already cataloged into group-level **MARC** records, with the title and call number of each individual image stored in the 505 field [See sample — Appendix A]. So, some programming was developed to extract the data from

continued on page 34

Appendix A — Sample group-level MARC record.
Art Institute of Chicago (Chicago, Ill.). Grant Park Garden [slide]

000:		: gm5 n FBU
007:		: gs uj jk
008:		: 910522s1987 waunnn 0sneng d
040:		: |cArch
049:		: IBSO
110:	2	: Skidmore, Owings & Merrill.
245:	10	: Art Institute of Chicago (Chicago, Ill.).|pGrant Park Garden|h[slide].
260:		: Seattle, Wash. :|bArt on File,|cc1987.
300:		: slides|bcol.
440:	0	: Place as art : pocket parks and gardens in the city (Series)
500:		: 1977
505:	0	: Art Inst.of Chicago. Grant Park Gdn. Overview: 2097-006 -- Plant bed: 2097-007 -- Arch: 2097-008, 2097-009.
596:		: 2
650:	0	: Gardens.
650:	0	: Urban parks.
650:	0	: Landscape architecture.
856:	40	: |uhttp://libx.bsu.edu/cdmlink.php?ckey=700018&coll=BSU_Arch SlidesCpght|yClick to view available images of this site or work.

Discrete Criteria ...
from page 31

Endotes

1. **Brian Kelly**, "Choosing a Metadata Standard for Resource Discovery." *http://www.ukoln.ac.uk/qa-focus/documents/briefings/briefing-63/html/*

2. **Yannis Ioannidis**, "Digital Libraries at a Crossroads," *International Journal on Digital Libraries 5, no. 4* (2005): 255-265.

3. **The Getty Museum**. "Metadata Standards Crosswalks" In Introduction to Metadata: Pathways to Digital Information. Online edition, version 2.1, *http://www.getty.edu/research/conducting_research/standards/intrometadata/metadata_element_sets.html*

4. **Dublin Core Metadata Initiative**. Glossary, s.v. "Application profile." *http://dublincore.org/documents/2001/04/12/usageguide/glossary.shtml*

5. **Erik Duval**, **Wayne Hodgins**, **Stuart Sutton**, and **Stuart L. Weibel**, "Metadata Principles and Practicalities," *D-Lib Magazine 8, no. 4* (2002). *http://www.dlib.org/dlib/april02/weibel/04weibel.html*

6. **Lynne C. Howarth**, "Metadata Schemas for Subject Gateways," *International Cataloguing and Bibliographic Control 33, no. 1*, (January/March 2004):8-12.