

Metadata Matters

Robby Robson, Eduworks Corporation & Oregon State University rrobson@eduworks.com

Forward

I was asked to write up some notes on the relevance of metadata to DML. My assumption, possibly false, is that the committee may know relatively little about metadata, so I wrote a bit of an introduction. The last section (entitled "Backward") has a very short list of my metadata-related concerns about DML.

Metadata Basics

Metadata are descriptive information associated with digital objects for cataloging, search and discovery, and delivery. All three uses are important, significant, and distinct. *Cataloging* refers to the act of sorting objects for storage and being able to retrieve them later. *Search and discovery* involves finding objects by matching syntactic or semantic criteria. *Delivery* includes causing a copy of an identified object to be available to a user.

Metadata Schemata

A system used to assign metadata to objects is called a *metadata schema*. A schema has several parts:

- *Elements* that serve as "labels" with agreed-upon interpretations and formats. (A typical element is "title" and a typical format is "string of characters not to exceed 1024 characters in length.")
- *Taxonomies* that define *controlled vocabulary* that can be used to populate the elements. (The MSC is a typical taxonomy.)
- *Obligations* that require certain elements to be used in any metadata instance.
- *Extension rules* that say how the schema can legitimately be expanded and altered. (For example, a rule might say that you cannot define a new element that has substantially the same meaning as an existing element.)

To illustrate the use of metadata, consider the problem of finding a paper written in French that references Bass' 1963 paper on the ubiquity of Gorenstein rings and is readable by a first or second year graduate student in mathematics. For this to happen, the hypothetical French paper must be properly classified and cataloged, not just with an MSC category but also with cross references and metadata indicating its language and difficulty level. Then the seeker of the paper must have access to a process that can search across all repositories that might contain the paper and that can submit a search query against the proper metadata fields. The repositories must be able to respond to such a search and the results federated and sorted if more than one result is generated. Finally, when the paper is found it must be selected and delivered. To do this it might be convenient to check whether the paper is in T_EX, PDF, Word, or some other format and on what platform and with what software the document can be read. This, too, must be reflected in the metadata and possibly interpreted automatically by the process that ultimately delivers the document to the searcher's desktop.

Metadata Misconceptions

Among the most common misconceptions about metadata are (a) that metadata is objective and (b) an object has a single metadata record associated to it. Neither is true. Much metadata is subjective and context dependent. For example, an animated fly-through of the graph of the Riemann Zeta function might be considered

- Advanced for use in a public mathematics lecture
- Moderate for use in a complex analysis course
- Beginning as an example of computer animation techniques

It makes sense to have different metadata records associated with the same fly-through graph that reflect the intended use of the graph by different communities. Note also that a metadata record is *not* part and parcel of a digital object. Metadata is often stored and used completely separately from the "physical" resource it describes.

Metadata Standards

Metadata can be used to describe self-contained collections, but its real applicaiton is in situations that cross cultural and contextual boundaries and that require data to be exchanged among cooperating technologies. The key concept in this case is *interoperability*. Standardize syntax is needed for machine interchange, and standardized schemata promote *semantic interoperability*. This is hard to achieve. For example, if you wanted to search a DML for a paper on Calculus, you would have to be aware of the fact that "Calculus" may not translate as such in another language or another time.

To achieve interoperability it is necessary to either use the same metadata schema or to clearly understand how two different schemata relate. (The process of mapping schema is called making a *cross-walk*.) This is what standards do. At the "data model" level, metadata standards specify element names and meanings, the structure of elements, the format of elements, controlled vocabulary and taxonomies, obligation, and rules for creating extensions. At the concrete level, data models are "bound" to programming or data interchange languages. Currently, XML is a popular choice, so metadata standards typically specify a way for metadata to be expressed in XML.

Standards are meant to be universally applicable and therefore purposely avoid being tailored to a particular community of practice. But that makes them useless in applications, so communities of practice tend to take standards, mix them, extend them, change the obligation for specific elements, and document their use so that they can be applied.

Existing Standards

The *Dublin Core Metadata* is the primary metadata standard used by the digital library community. Another related schema is *Learning Object Metadata*, which is a "real" standard in the sense that it is an accredited IEEE standard (as of June, 2002). There is a great deal of compatibility between the two and there are real life examples of situations where *both* are used at the same time.

Applications to DML

It is natural to ask why metadata matters for digital libraries. Here is a short list of reasons:

1. *Digital Libraries are just collections of metadata anyway.*
2. *The DML will want to support searching across multiple collections. It will want to be multiple collections. Not everyone may agree with that, but that's my view.*
3. *Data without metadata is worthless. The point of a DML is hopefully to make content accessible and worthwhile That's what metadata does.*

Backward

To close, I would like to express the following concerns:

- Settling on a schema, or to be more precise, an application profile of a schema. Experience shows this will not be easy, and I don't believe a simple author/title/subject classification is enough.
- Internationalization and localization are always of concern when dealing with metadata. It often seems hard for Americans to "get it" and (with apologies) even harder for Europeans to understand that they don't get it either!
- It takes a commitment of time and money to go to the effort of assigning good metadata. I would argue that without metadata, the data is worthless, yet experience also shows that metadata generation is seldom part of the budget.
- This is not a metadata concern *per se*, but creating a DML requires choosing underlying technology. Hopefully this will be an enterprise grade content management system, and hopefully the collection will be distributed, not centralized. I won't go into the arguments here, but I am concerned that there may not be proper respect for the challenges of managing a large collection of digital content on a global scale.

Meta-Metadata-References

- 1 International Federation of Library Associations and Institutions, Digital Libraries: Metadata Resources <http://www.ifla.org/II/metadata.htm>
- 2 Taxonomy, Classification and Metadata Resources (Centre for Educational Technology Interoperability Standards) http://www.sesdl.scotcit.ac.uk:8082/taxonomy_links.html
- 3 Metadata, Schmetadata references: <http://www.eduworks.com/metadatarefs.htm>