

# **Metadata As A Component Of The DML**

## ***DML Metadata Working Group – Interim Report***

### **4 August 2003 version**

WG Co-Chairs:

Tim Cole, Mathematics Library, University of Illinois at Urbana-Champaign  
Heike Neuroth, State and University Library Göttingen

WG Member:

Robbie Robson, Eduworks Corporation

Contributors:

Stefan Farrenkopf, State and University Library Göttingen  
Marty Kurth, Cornell University Library, Ithaca, NY  
Jinfang Niu, Tsinghua University Library, Beijing

Editing:

Kizer Walker, Cornell University Library, Ithaca, NY

*This report describes the role that metadata will play in a comprehensive Digital Mathematics Library and outlines metadata-related issues and tasks that must be addressed as part of the DML undertaking. It recommends that a first priority of any effort to coordinate implementation of a comprehensive Digital Mathematics Library be the selection or creation of a preferred DML metadata schema and development of one or more DML Metadata Application Profiles, and provides a starting point for such work.*

## **I. Introduction**

In the context of the DML, metadata will serve a variety of purposes. From a technological perspective, objects in the DML will be metadata records representing primary resources that are digitally available from one or more servers. This is what makes it possible to construct distributed repositories and also what makes the identification of a metadata schema and development of an application profile among the first concrete design decisions that must be made. From an end-user perspective, metadata facilitates the processes of search and discovery (which includes research based on tracing references and ideas) and the sharing of objects across multiple application domains. Because so much is built upon metadata, the cost of reversing a decision about metadata may be second only to the cost of reversing a decision about the format used for encoding the actual objects.

Metadata records, defined generally as data about data (where “data” is interpreted broadly and very generally and encompasses such things as books, journal articles, dissertations, working papers, and even more abstract objects such as mathlets), come in a variety of “flavors” and differ considerably in scope and focus according to intended use and purpose. Multiple metadata records are often associated with a single object or manifestation of an intellectual work (e.g., the online edition of a particular journal article). Metadata may also be associated with higher level abstractions of information resources—e.g., borrowing terminology from the hierarchy developed by the IFLA Study Group on the Functional Requirements for Bibliographic Records (FRBR), metadata may be used to describe more abstract entities such as a work or a particular expression of a work which in turn may have multiple manifestations in both print and digital form. The depth, richness, and specific content of a given metadata record varies according to the purpose for which it is intended and the kind of object with which it is associated. Metadata records may be generated for a very narrow, specific purpose (generally less expensive to do), or may be generated in greater detail and breadth to serve multiple roles simultaneously (generally more expensive to do). There are ways to mix and match together multiple different, single-purpose metadata

records describing the same or related intellectual objects in order to create one, more complete, multi-faceted record that can serve multiple purposes at once.

In a project such as DML, which includes a focus on the retrospective digitization of existing scholarly print resources, there is the opportunity to leverage existing metadata resources to considerable benefit. Expressions of scholarly works in mathematics published as print monographs or journal articles are well described, especially in a bibliographic sense, in the databases of such services as MathSciNet, Zentralblatt MATH, the Jahrbuch Project, OCLC WorldCat, etc. Viewed as repositories of metadata, these implementations are well vetted and contain much value-added content. They do a comprehensive job describing the bibliographic characteristics of expressions of works and print manifestations of those expression entities. They are particularly useful as tools for discovery.

However, they do not provide the full range of metadata required to manage, coordinate, and use new digital manifestations of a work, even when derived from existing print manifestations. For instance, MathSciNet does not presently include (in any structured way) information about how a manifestation was digitized or the intellectual property rights associated with a particular digital manifestation of a work, information that may be necessary for informed use of such manifestations. There is typically little metadata provided useful for choosing between two different digital manifestations of a work derived from the same print manifestation. Linkages to some kinds of prior or successive digital-only manifestations of a related work (e.g., an online preprint or a post-publication annotation page or Website) are generally not represented. Metadata fields necessary for preservation and archival functions are not all present. Additionally, access to these services (and presumably the metadata they maintain) is often limited to paying subscribers, potentially limiting the use of metadata from these services in the open DML architecture.

The DML will need to develop and implement a strategy that at once takes maximum advantage of existing metadata repositories and provides guidance to digitizing agents as to additional metadata required or considered desirable. Bi-directional interfaces at the metadata level between existing mathematics-oriented metadata services and DML could offer potential advantages for both. Strategies for acquiring and facilitating creation of metadata will need to be developed in coordination with the selection or creation of an appropriate metadata schema and work done on one or more DML-specific application profiles.

## **II. Role and Purpose of Metadata in DML**

In building large, distributed repositories, metadata has several useful features that facilitate the integration of diverse and often widely distributed primary source objects. Though the primary source objects may be quite diverse in encoding schemes and formats, metadata records typically will be encoded in a more uniform way (e.g., XML). Semantics, ontologies, and taxonomies are usually defined to facilitate interoperability and interchange of metadata records at some common level. Metadata records are generally much smaller than the intellectual resources they describe, again making them easier to transport, manipulate, and integrate (e.g., into a single index) than primary source items themselves, which often are best maintained in a more distributed fashion. Metadata records often include content derived from the primary sources they describe, but the use of standard syntax and semantics means that this content is exposed in a more normalized fashion, facilitating implementation of search and discovery services and other aggregation and maintenance services such as preservation and archiving. Metadata records may also contain information not included in primary sources, such as added reference information, context information, provenance information, terms of use information, information needed to manage intellectual property rights and values to assure fixity. Administrative, structural, technical, and preservation metadata is essential to smooth operation and long-term maintenance of large distributed information resource repositories.

We anticipate that the DML will make use of metadata to support the following services (a less than comprehensive list):

1. A central registry (catalog) of available content for browsing, collection development, collection identification, etc. The first draft of the Content WG report <<http://www.library.cornell.edu/dmlib/internal/dml-content-2002oct.html>> suggests that it will be important early on to “install a global registry for all items detected and handled so far” in order to develop and maintain a complete picture of what is “in” the DML at any point in time. This knowledge is essential for ongoing DML Collection Development and also will be helpful to end users wishing to browse the resources of the DML and/or discover collection components of the DML of particular interest.
2. Bibliographic Description for Search and Discovery of Items. Because of the diversity of content and formats of content that will be included in the DML, it will not always be possible to search full text of all primary sources in the library. Eventually born-digital content will live side-by-side with retrospectively scanned and OCR’d content. Various classes of information resources will be included in the DML (e.g., journal articles, dissertations, monographs, textbooks, preprints, working papers, etc.). Additionally, as described above, metadata records will have content not in the primary sources. So in many instances it will be preferable to search at the metadata or bibliographic description level rather than trying to search full-text across the complete universe of content.  
Note: “Discovery” is not limited to searching references. With good metadata, temporal information, classifications, information about the nature of resources (e.g., that it stems from a graduate thesis) can provide valuable tools for tracing the development of ideas and techniques.
3. Descriptive information for utilization of items. While a standard for inclusion will define certain minimal technical quality criteria, it’s likely that the size and viewing requirements associated with a particular resource may influence its suitability for some purposes. Terms of use may also be an issue for some content in the DML. Metadata is a way to let end users determine whether they can satisfy authorization and technical requirements necessary to use a primary source in the DML.
4. Interoperability and metadata to support management and archiving. The DML will serve not only as a portal for direct use by end users who want to access the universe of scholarly and research mathematics information, but also as a centralized point for interoperation with allied projects (e.g., the USA’s National Science Digital Library). Additionally it will be a goal of the DML to insure long-term preservation of important and useful content. The administrative and archiving functions of the DML will rely heavily on metadata.
5. Intellectual property rights management. The goal of making resources available freely or at no cost should not be construed as obviating the need to manage copyright, attribution, and rights and conditions. As standards emerge for “rights management” (see MPEG-21), it will be required to incorporate these into metadata.
6. Authority control. To enable good search recall and precision, and to facilitate tying specific manifestations to related expression and work entities, it’s essential to implement authority control. This can best be done when constructing values for metadata fields such as names, subjects, identifiers.

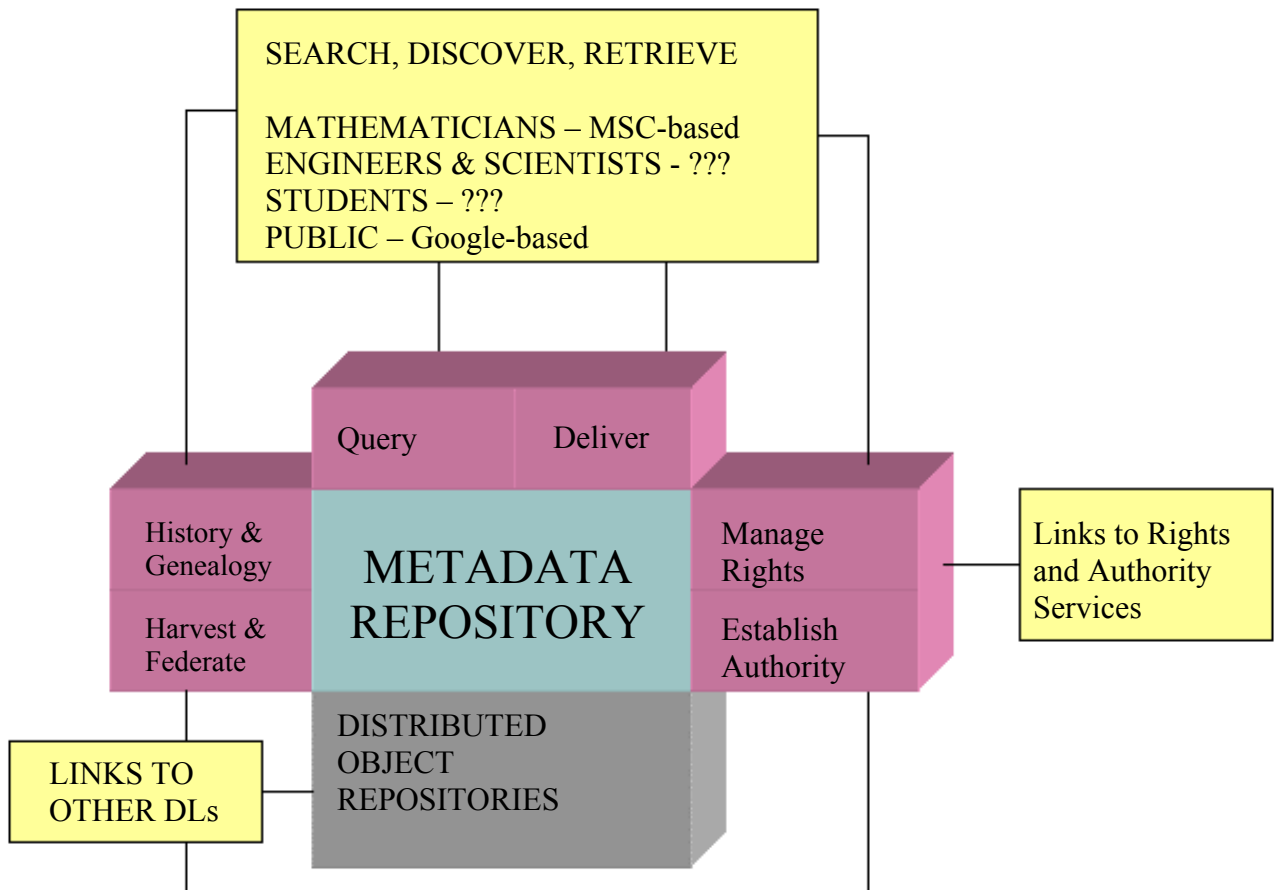


Figure shows a model of DML services that will be built on top of object metadata aggregation.

### III. Metadata Schemas and Application Profiles Relevant to DML

In order to encourage the creation and dissemination of metadata useful for constructing the services outlined above, the DML will want to publish and maintain DML-specific metadata schema standards and a DML application profile (or set of tightly coupled metadata application profiles specific to each class of content subsumed in the DML). Metadata schema and application profile(s) will present de facto standards or best practices for creation of metadata associated with all collections and information objects to be included in the DML. They will clearly define metadata record fields of interest to DML and will describe assumptions and rules that should be observed when populating those fields. They will identify fields required and/or recommended for specific purposes and may set minimum standards for metadata and metadata encoding that must be met for inclusion of content in the DML.

While a specific metadata schema and application profile(s) will need to be developed and maintained for the DML, these will be developed and maintained side-by-side with other community-specific metadata application profiles, many of which will overlap in detail and purpose with DML application profiles. Because metadata creation can represent a significant cost in the creation of new digital content or the translation of analog content to digital form, making sure that any DML metadata application profiles build on and make use of other existing metadata application profiles is especially important. The first

step in constructing a DML Metadata Application Profile is to survey relevant digital metadata application profiles in use or under development currently.

Below is a partial list of relevant published metadata application profiles and schemas, along with some relevant projects and organizations with well-defined metadata schemes, even if those schemes are not yet published in the format of a formal metadata application profile. The working group proposes that effort be directed to developing and maintaining a more complete listing of relevant metadata schemes.

1. EMANI Project (application profile forthcoming from SUB Göttingen)
2. DC-Lib (Dublin Core Library Application Profile)  
<<http://www.dublincore.org/documents/2002/09/24/library-application-profile/>>
3. Qualified Dublin Core Schema (and variants like NSDL\_DC  
<[http://ns.nsdlib.org/schemas/nsdl\\_dc/nsdl\\_dc\\_v1.00.xsd](http://ns.nsdlib.org/schemas/nsdl_dc/nsdl_dc_v1.00.xsd)>
4. Learning Object Metadata (IEEE). Draft Standard, 15 July 2002  
<[http://ltsc.ieee.org/doc/wg12/LOM\\_1484\\_12\\_1\\_v1\\_Final\\_Draft.pdf](http://ltsc.ieee.org/doc/wg12/LOM_1484_12_1_v1_Final_Draft.pdf)>
5. Metadata Object Description Schema (MODS) <<http://www.loc.gov/standards/mods/>> and  
MARC-XML <<http://www.loc.gov/standards/marcxml/>> (Library of Congress)
6. Tsinghua University Library Metadata Framework (see Jinfang Niu. "A Metadata Framework  
Developed at the Tsinghua University Library to Aid in the Preservation of Digital  
Resources." *D-Lib Magazine* 8.11 [Nov. 2002]).  
<<http://www.dlib.org/dlib/november02/niu/11niu.html>>
7. Container Schemas, e.g., Metadata Encoding and Transmission Standard (METS)  
<<http://www.loc.gov/standards/mets/>>, Resource Description Framework (RDF)  
<<http://www.w3.org/RDF/>>
8. MPEG-7 (formally "Multimedia Content Description Interface")  
<<http://ipsi.fraunhofer.de/delite/Projects/MPEG7/index.html>>
9. D-Lib Magazine
10. Project Euclid (SPARC / Cornell University)
11. GDZ (Göttinger Digitalisierungs-Zentrum)
12. Springer-Verlag

#### **IV. Unique Features of DML Metadata Schema and Application Profile**

The continuing DML metadata discussion will need to take the following questions into account:

- To what degree will DML need special IP Rights metadata and to what degree will it be able to or need to make use of IP Rights metadata work currently underway in the wider community (e.g., Creative Commons, Project RoMEO, IP Rights extensions being considered for OAI-PMH)?
- How can DML retrospective digitization efforts leverage existing math metadata resources? Can basic bibliographic fields from Math Reviews, Zentralblatt MATH, the Jahrbuch Project, OCLC be made available for DML?
- What are the options for prioritizing metadata fields for particular metadata-related services?
- What sort of centralized DML services can be constructed to facilitate metadata creation? These might include a name authority file lookup and matching service, subject controlled vocabulary switching, encoding and translation services, etc.

- To what extent would metadata field value inheritance be useful? For instance, metadata field values assigned to a repository or set of digitized content could by default be inherited for all items within that repository or set.
- What issues are raised by potential reuse of DML materials (possibly at sub-article level)? By the use of DML by non-mathematicians?

## V. Metadata Encoding Schemas and Standards to Be Used

The following provisional points should guide the ongoing discussion of DML metadata:

- DML metadata should conform to relevant metadata schemas and namespaces:
  - Dublin Core (DC) and Qualified Dublin Core (DCTERMS)
  - Dublin Core Variants (e.g., nsdl\_dc)
  - MODS or other XML-MARC encoding standard
  - MIX – NISO
  - RLG / OCLC Preservation Metadata schemas
- DML projects should consider associating descriptive metadata with the content they digitize that will facilitate the use of the materials in appropriate educational contexts; IEEE Learning Object Metadata represents one example of an educational metadata schema.
- Subject classification taxonomies are preferred.
- While it remains an open question whether DML projects will be required to express metadata in XML, they will at the very least be strongly urged to do so in order to facilitate interoperability, the implementation of metadata scheme namespaces in XML, and other useful technologies.
- The continuing DML metadata discussion should aim to estimate the impact of requirements on digitization projects and publishers.

## VI. Metadata Implications for DML Architecture Design

The DML metadata discussion must inform the development of an information architecture for the DML. Decisions on exchange and transport protocols have important metadata implications. Outstanding issues include:

- The question of federated search vs. harvesting models
- The implication for DML of Open Archive Initiative (OAI) reliance on HTTP and XML schema
- DML use of XML and Z39.50/ZING

A thorough consideration of the metadata implications of indexing and archiving models adopted by DML will be required.