Digital Mathematics Library
Final Report
October 2004

Principal Investigator:
Sarah E. Thomas, University Librarian, Cornell University

Co-Principal Investigators:
R. Keith Dennis, Professor of Mathematics, Cornell University
Jean Poland, Associate University Librarian for Engineering, Mathematics, and Physical Sciences, Cornell University

Contents

Digital Mathematics Library

NSF Award Number: DUE-0206640

2002-2003 Steering Committee

Keith Dennis	Cornell	dennis@rkd.math.cornell.edu
Jean Poland	Cornell	jp126@cornell.edu
Hans Becker	Göttingen	becker@mail.sub.uni-goettingen.de
Pierre Bérard	Grenoble	Pierre.Berard@ujf-grenoble.fr
Bernd Wegner	TU Berlin / Zentralblatt	wegner@math.tu-berlin.de

2002-2003 IMU Liaison Committee

Rolf Jeltsch	EMS / ETH Zurich	jeltsch@math.ethz.ch
David Mumford	Brown	David_Mumford@brown.edu

2002-2003 Working Groups

Content

Co-chairs:

Keith Dennis	Cornell	dennis@rkd.math.cornell.edu
Bernd Wegner	TU Berlin / Zentralblatt	wegner@math.tu-berlin.de

Member:

Steve Rockey	Cornell	swr1@cornell.edu

Technical Standards

Co-chairs:

Thierry Bouche	Grenoble / NUMDAM	thierry.bouche@ujf-grenoble.fr
Ulf Rehmann	Bielefeld	rehmann@mathematik.uni-bielefeld.de

Metadata

Co-chairs:

Tim Cole	Illinois	t-cole3@uiuc.edu
Heike Neuroth	Göttingen	neuroth@mail.sub.uni-goettingen.de

Member:

Robbie Robson	Eduworks Corporation	rrobson@eduworks.com

Rights and Licenses

Co-chairs:

Pierre Bérard	Grenoble / NUMDAM	Pierre.Berard@ujf-grenoble.fr
David Tranah	Cambridge UP	dtranah@cup.cam.ac.uk

Archiving

Co-chairs:

Hans Becker	Göttingen	becker@mail.sub.uni-goettingen.de
Kizer Walker	Cornell	kw33@cornell.edu

Economic Model

Chair:

James Crowley	SIAM	jcrowley@siam.org

Members:

Jonathan Borwein	Simon Fraser U	jborwein@cecm.sfu.ca
Arnoud de Kemp	Springer-Verlag	DeKemp@Springer.de
John Ewing	AMS	jhe@math.ams.org
David Tranah	Cambridge UP	dtranah@cambridge.org

Digital Mathematics Library
Final Report
October 2004

Principal Investigator:
Sarah E. Thomas, University Librarian, Cornell University

Co-Principal Investigators:
R. Keith Dennis, Professor of Mathematics, Cornell University
Jean Poland, Associate University Librarian for Engineering, Mathematics, and Physical Sciences, Cornell University

In July 2002, an international group of mathematicians, scholarly publishers, technical experts, and academic librarians convened in Washington, D.C., to initiate development of a framework for a comprehensive, international, distributed collection of digital information and published knowledge in mathematics. The one-year planning phase of the Digital Mathematics Library (DML) project was coordinated by Cornell University Library (CUL) and funded by the U.S. National Science Foundation (NSF). The DML will promote the digitization of mathematics content by establishing standards and guidelines for digitization, by providing tools, by providing a stable and global window to digital archives of mathematics literature.

Expert working groups were formed for the duration of the planning phase to investigate DML content parameters, rights and licenses, economic models, technical standards, metadata standards, and archiving. Their recommendations were delivered in individual interim reports, which were discussed and amended in a meeting of the DML Steering Committee in Grenoble, France, in March 2003, and at the second meeting of the larger planning group in Göttingen, Germany, in May 2003.[1] The present report summarizes the conclusions agreed on by the DML planning group with respect to the various working group areas; the individual working group reports appear as appendices. Further documentation of the activities of the working groups, along with other information related to the project, is available at the DML project website: http://www.library.cornell.edu/dmlib.

The initial DML planning group completed its work with the close of the May 2003 meeting in Göttingen and disbanded. The Committee on Electronic Information and Communication (CEIC) of the International Mathematics Union (IMU) has assumed coordination of the next phase of the project.[2] This global effort is named World Digital Mathematics Library (WDML), to differentiate it from national and regional DML initiatives, such as the multinational DML-EU project that has applied for funding from the European Union. In July 2003 a WDML Steering Committee was formed, consisting of:

---

[1] Allyn Jackson, "The Digital Mathematics Library," *Notices of the AMS* 50.8 (Sep. 2003)*:* 918-23. http://www.ams.org/notices/200308/comm-jackson.pdf.
[2] See Summarised Minutes of the Sixth Meeting of the Committee on Electronic Information and Communication (CEIC) of the International Mathematical Union held at the Konrad-Zuse-Zentrum (ZIB) Berlin, May 24–25, 2003. http://www.ceic.math.ca/filemgmt_data/files/minutesBerlinS.pdf

- Alf van der Poorten, Macquarie University, Sydney, Australia (Chair)
- Pierre Bérard, University of Grenoble, France
- Thierry Bouche, University of Grenoble, France
- Gertraud Griepke, Springer-Verlag, Heidelberg, Germany
- Rolf Jeltsch, Seminar for Applied Mathematics, ETH Zurich, Switzerland
- David Mumford, Department of Mathematics, Brown University, Providence, RI
- Jean Poland, Cornell University Library, Ithaca, New York
- Bernd Wegner, Zentralblatt für Mathematik and Department of Mathematics, Technical University Berlin, Germany

The Steering Committee organized a WDML workshop in connection with the Fourth European Congress of Mathematics held in Stockholm in June 2004.

The NSF extended the grant period for the original DML planning project to October 31, 2004. The extension has facilitated the project's transition to IMU leadership and supported additional planning for continued interaction among digitization projects. The no-cost extension has also allowed CUL to apply remaining grant funds, along with a contribution from the Library of approximately $30,000, toward digitizing the backrun of an important journal title as a proof-of-concept for DML standards. In the context of the Electronic Mathematical Archiving Network Initiative (EMANI) – a collaborative endeavor of the scientific publisher Springer, three major academic libraries, and leading mathematical societies and projects – CUL reached an agreement with Springer to digitize the backfiles of Communications in Mathematical Physics and make them freely available online to scholars worldwide. Springer offers online (subscription-based) access beginning with volume 183 (January 1997); CUL has digitized volumes 1-182 (1965-1996). Beginning in December 2004, the retrodigitized volumes will be offered on an open access basis via Project Euclid (http://projecteuclid.org), CUL's non-profit initiative to advance effective and affordable scholarly communication in mathematics and statistics. The CMP backfiles project thus leverages multiple distinct efforts toward the fulfillment of the DML vision. CUL is grateful to Springer for permitting redistribution of the CMP content, as well as to Professor Keith Dennis (Mathematics, Cornell), the American Mathematical Society library, and the Purdue University Libraries for donating the set used in scanning.

What is the DML?

The DML is a collective effort of mathematicians, scholarly publishers, technical experts, and librarians to greatly broaden access to the scientific and cultural heritage represented in published mathematics and to preserve it for the long term. Each stakeholder has much to gain from the present and future accessibility of the mathematics literature. The DML will be a distributed, interoperable collection of digital mathematical content. It is also to be understood as a loosely structured voluntary association, an international "club" entailing benefits as well as obligations for its members. The DML organization will be linked to the IMU to ensure that it provides fair and balanced international representation for mathematicians and mathematics. It will be headed by a steering committee drawn from participating digitization projects. This decentralized organization will be funded on a contributory model; for instance, travel and attendant costs for annual meetings will be incorporated into the grant proposals of the individual projects.

DML Content

The DML has set as the scope of its collections "the entirety of past mathematics scholarship." But even a goal of comprehensive coverage requires selection criteria. Building the DML is a

long-term process; a time schedule for retrodigitization efforts and collection building is needed, which implies prioritization, and hence selection. One of the important tasks for developing a complete picture of the content of the (global) DML is the creation of a global registry that will record all items processed to date along with a prioritized list of items targeted for digitization and inclusion in the collection.

Document types
The DML planning group has distinguished the following document types in mathematics for inclusion in the collection:

Journal material
  1. Articles in refereed journals (in the narrower sense)
  2. Articles in Newsletters and other non-refereed journals
  3. Articles in series of collections of publications

Monograph material
  4. Articles in conference proceedings and other non-periodic collections of publications
  5. Series of advanced level monographs
  6. Single advanced level monographs
  7. Series of textbooks
  8. Single textbooks
  9. Collected works, handbooks, encyclopedias, bibliographies and similar publications

Special cases
  10. Publications dealing with education in mathematics (including curricula), popularization of mathematics
  11. Dissertations
  12. Articles deposited in repositories like those at VINITI

Subject scope
There is no question that the DML must cover publications and serials dealing with pure and applied mathematics in a narrow sense. But there are serials with mixed content, which contain a relevant amount of mathematical literature, as well as serials publishing articles on the boundaries of mathematics with applications to other sciences like statistics, logic, theoretical computer science, theoretical physics, theoretical mechanics, and many other areas in the application of mathematics. On a practical level it has to be decided whether such serials should be digitized and stored as a whole, or if the important mathematical articles should be selected and put into DML case by case.

Action parameters
The planning group proposes five dimensions for determining DML coverage and prioritizing digitization and collection development efforts. These parameters may change over time:

T - Time: The first question is that of coverage in terms of publication year. What is the most recent material to be retrodigitized and how far back should the effort go? Note that all current and future electronic publications in mathematics will belong to the DML by definition.

M – Side to side: The scope of the literature to be covered has been addressed above. As the experience of the reviewing services shows, the problem of determining the boundaries between mathematics and economics, statistics, physics, etc., can have no precise solution. An agreement

must be reached as to the proportion of mathematical content required for material to be included in the DML.

L - Top to bottom: What is the range of mathematics in terms of levels of complexity to be included? Decisive factors include impact on research, potential interest on the part of different user communities (research, education, applications, history, etc.), quality, availability.

G - Back to front: A geographical dimension may affect priorities for DML efforts. The DML must broaden its scope to cover content from all over the world.

C - Free to locked: The key aim of the DML is to make content available "at reasonable cost." How is this to be defined in particular instances? Parameters T and C are strongly linked by what is called the "moving wall". This indicates the period after which the content may move from the charged state to the free access state.

Access to content
The demand to have content available as quickly as possible may also impact the selection procedure. In the interest of the user community, the DML could favor digitization projects that can deliver quick access. (A project that would link to digitized content via the current mathematics reviewing databases is one possible example.)

Status of the content
The content covered by the major reviewing services is well defined and may be taken as a core for the potential content of the DML. The DML should determine current status of this set of publications with regard to availability in digital form. These publications (and any others) can be roughly classification as follows:

A. Digitally published articles in mathematics.
B. Digital versions of articles initially available only in print, i.e., retrodigitized articles.
C. Printed articles without a digital version, already adopted by a retrodigitization project.
D. Printed articles without a digital version, not yet adopted by a retrodigitization project.

A principle task of the DML is to facilitate the swift migration from D to C and from C to B.

Rights and Licenses

In light of the complexity of intellectual property law and in particular the differences from one country to the next, it is impossible to give a single, concise set of recommendations for a digitization project of international scope.

In order for the DML to be successful in digitizing the legacy literature, it will be important to create a plan that will attract content owners/providers who currently hold the legacy literature. These include societies, not-for-profit publishers (including university presses), and commercial publishers. The question of who holds the relevant rights on the collection's content is a fundamental one. The planning group has concluded nonexclusive rights to articles will be retained by the original owner (e.g., the publishing journal) and that material will be included in the DML by agreement between the original owner and the DML or one of its member programs. This cautious approach vis-à-vis the traditional author-publisher relationship is most likely to meet with acceptance from both sides. It is also in keeping with the loose structure planned for the DML organization, which will not be capable of owning rights.

This approach to copyright aligns with the following model for building the DML: DML member projects digitize material for publishers, then return it to them; in exchange, publishers agree to make the material available on their sites via the DML framework. The DML will not require transfer of ownership of content to it from current owners; the key is rather making binding agreements that detail the publisher's responsibilities, as well as remedies in case those are not met. The DML need not host content for delivery purposes. It must host a large database with rich metadata that is freely searchable. Publishers' participation in the DML will require that they provide metadata and allow full access to it without charge. Archival copies of DML content will reside in digital repositories operated by participating libraries.

Economic model

Overall DML efforts will be distributed internationally among many smaller projects, to be funded in most cases by national funding agencies, or other appropriate agencies.

Content owners/providers will be willing to join in a venture to digitize legacy material provided their interests are protected and they perceive a benefit in cooperating.

Features that content owners/providers might wish to see in an agreement include:
- Financial support for digitization
- Retention of digitized material for posting on organization's web site
- Minimal standards for digitization that provide interoperability, but are flexible enough to fit within workflow and existing digital archives.

In exchange, content owners/providers would be expected to:
- Host digitized material on their own site
- Provide access to digitized material
- Link articles (and references) through Math Reviews and Zentralblatt, as well as CrossRef
- Upgrade formats in future years to ensure continued operability

The DML will manage the agreements among the content providers to ensure continued access to the information under the agreed terms. Contracts should stipulate that libraries participating in the DML will maintain a backup copy of files with the right to post the backup material should the content provider cease operation or fail to live up to the terms of the agreement. This provision ensures that materials produced with public funding will not vanish.

Archiving

Along with comprehensive online access to the mathematics literature, the long-term preservation of the literature is a core aim of the DML, which is envisioned as an "authoritative and *enduring* digital collection." A framework for preservation planning and provision for archival storage must be a DML priority from the outset. The DML will approach archiving and preservation by way of the generic archive framework proposed in the *Reference Model for an Open Archival Information System (OAIS)*. First published in 2001, the OAIS Reference Model has quickly become the prevailing model internationally for planning and implementing digital archives and has recently been adopted as an ISO standard. OAIS was developed for the aerospace domain, with participation of major space agencies in North America and Europe, as well as the library and archiving communities. It is valued for its rigor, as well as for its high level of abstraction, which makes it scalable and adaptable to various domains. OAIS represents the basic framework for digital archives planning at the DML partner libraries.

OAIS provides an abstract model for managing the preparation and transfer of digital objects into the digital archive, their long-term storage, and their dissemination. Functions such as disaster preparedness and recovery, and the periodic migration of data from older storage media and older digital formats to current ones are laid out in OAIS. The OAIS model is metadata-driven. The importance of standard metadata for digital archiving and for interoperability of a distributed archive cannot be overemphasized.

Metadata

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. 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.

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. 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. Administrative, structural, technical, and preservation metadata is essential to smooth operation and long-term maintenance of large distributed information resource repositories. 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.

Role and Purpose of Metadata in DML

The planning group anticipates that the DML will make use of metadata to support the following services (a less than comprehensive list):

1. A central registry (catalog) listing available content as well as a prioritized list of content to be digitized and included in the collection, for browsing, collection identification, collection development, etc.

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. Even where full-text searching is available, in many instances it will be preferable to search at the metadata or bibliographic description level rather than across the complete universe of content.

3. Descriptive information for utilization of items. While a standard for inclusion will define certain minimal technical quality criteria, it is 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 low 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 is essential to implement authority control. This can best be done when constructing values for metadata fields such as names, subjects, identifiers.

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. 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.

Technical standards

Building the DML will be an effort of many geographically dispersed projects over many years. To build and sustain a distributed digital library that is interoperable, searchable, and capable of collaborative archival custodianship of the mathematics literature, it is essential that participating digitization projects adhere to a basic set of shared technical standards. Projects that are already underway and that wish to participate in the DML should work toward adoption of the following standards; new projects should adopt them at the outset.

1. Scanning Quality:

600 dpi bitonal represents the minimum quality level for scanning. In special cases and in the long run, higher resolutions, grayscale, or even color may be more suitable. This minimum resolution is appropriate for archival storage; a lower resolution may be adequate for purposes of delivery in the present technological environment.

Obvious flaws of the printing like skewed printing areas should be corrected during the scanning process. The printing area of each page should be positioned at the same place for all pages of a given object, possibly reflecting the differences for "right" and "left" pages. Page jumping, rotations, and varying the margins and dimensions of images are discouraged.

2. Archiving Formats
For scanned raw data use PNM or TIFF. CCIT G4 for bitonal, lossless compressed, LZW, ZIP for gray, color.

3. File Name and URL Conventions:

Among other things, the following should be guaranteed:
  * a unique and meaningful name for all files;
  * stable URLs for all documents;
  * uniform appearance of web pages for all DML servers;
  * uniform methods of access for all documents.

4. Delivery Formats:

Two file formats are recommended for delivery: PDF and DJVU. PDF and DJVU files should be made searchable with an underlying text layer. Non-ASCII letters such as accents and diaereses should be encoded using unicode.

Links to MathSciNet or Zentralblatt Math should be added to the references.

5. Download Units:

	primary	*secondary*
Journals	single articles	(annual) volumes
Books	whole book	chapters?

Browsing tables of content is desirable.
Download of single pages only is discouraged; download of page ranges is desirable.

6. Server Techniques:

Linearized PDF files should be delivered; server should be configured for "byte serving." This allows users to view one page at a time without downloading an entire book.

Promoting the DML

Too few mathematicians know about or use the currently digitized mathematics literature. Still tied to print, they cannot take advantage of even simple, but extremely useful things, such as electronic searching. Today these mathematicians are not in the position to see the value of an international archive such as the DML represents, and so are not demanding support for it. Crucial to changing this state of affairs is a concerted effort on the part of all DML stakeholders to publicize the DML effort and its achievements. This must take the form of publications and formal and informal discussions in the scholarly societies and professional associations. It is essential that the progress of the DML be highly visible, and coordination of DML efforts by the IMU/CEIC will help ensure this visibility. The DML must maintain a registry of digitized mathematics and a listing with up-to-date information on the digitization projects. As digitization of the literature progresses, it must be disseminated and linked through the networks upon which mathematicians rely. DML digitization projects must offer exportable records for monographs and serials in standard formats to libraries for their online catalogs. Records for journal articles should be provided to MathSciNet and Zentralblatt Math, along with records for reference linking. Support for these efforts will naturally grow as the literature on which mathematics scholarship is built becomes available from the "shelves" of every library in the world.