

**DML ARCHIVING WORKING GROUP – INTERIM REPORT**

Co-chairs:

Hans Becker – SUB Göttingen

Kizer Walker – Cornell University Library

Nancy McGovern and Bill Kehoe (Cornell University Library) also consulted on this report.

The Digital Mathematics Library initiative has set as its core aims comprehensive online access to the mathematics literature coupled with the long-term preservation of the literature – the DML is envisioned as an “authoritative and *enduring* digital collection.” Libraries and archives have historically carried out the twin tasks of access provision and preservation of the cultural record. As that record has been increasingly set down in digital form, the library and archives communities have invested heavily in addressing what has been called “the digital preservation problem”: “how can a digital resource retain intelligible meaning in the long-term?”<sup>1</sup> A 1996 report proposing a network of digital archives defined the archival function in terms of both preservation of content and preservation of access. The archives would be “repositories of digital information that are collectively responsible for ensuring, through the exercise of various migration strategies, the integrity and long-term accessibility of the . . . social, economic, cultural and intellectual heritage instantiated in digital form.”<sup>2</sup>

**Long-term access**

Though preservation strategies for data in electronic form have developed over four decades, the relative newness of the digital media means that the *long-term* effectiveness of these strategies remains insufficiently tested by time. Indeed, in the context of digital archiving, the “long term” has been usefully described precisely in terms of this uncertainty about preservability, as a “period long enough to raise concern about the effect of changing technologies . . . and of a changing user community.”<sup>3</sup>

Objects can lose digital integrity over time as data or the medium on which they are stored deteriorates. But perhaps a more pernicious problem stems from the rapid evolution of digital media, which shows no sign of slowing: the storage medium of the digital object, as well as both the hardware and software required to interpret it are prone to obsolescence [see Cedars 4-5]. Discussion of digital preservation strategies in the face of the problem of obsolescence has tended to polarize around the concepts of data *migration* and technology *emulation*. The practice of periodically migrating data from older storage media and older digital formats to current ones emerged early as a strategy for preserving information in digital form.<sup>4</sup> It is not as yet established that migration strategies are effective and reliable for preserving the full range of digital formats and media over the long term. Critics argue that data migration is too labor-intensive and propose as an alternative the emulation of older technological environments to allow materials to be viewed in their original formats.<sup>5</sup> Recently it has been suggested that migration and emulation should be viewed as interrelated approaches to preservation, rather than as competing models.<sup>6</sup> Tools are, in any case, needed to automate the migration process. Since the DML will, at least initially, handle relatively homogenous materials – books and journals retrodigitized to agreed standards – technology emulation strategies will be of minimal relevance to this project for the foreseeable future.

Redundancy of storage is a key component of digital preservation and disaster recovery planning. Greenstein and Marcum provide these guidelines for viable archival use of redundant storage:

At a minimum, repositories will need to operate as part of a network to achieve a satisfactory degree of redundancy for their holdings. Although an appropriate level of

redundancy is difficult to quantify (let alone mandate), it will ideally extend for any single data to three archival sites, at least one of which is located off shore.<sup>7</sup>

As an international, inter-institutional project, the DML is in an excellent position to realize archival redundancy.

Peer-to-peer approaches have been put forward as cost-effective means of ensuring redundancy.<sup>8</sup> The LOCKSS (“Lots of Copies Keep Stuff Safe”) project at Stanford University (<http://lockss.stanford.edu/>) has developed a software mechanism to support redundancy through the creation of “low-cost, persistent digital ‘caches’ of e-journal content housed locally at institutions that have authorized access to that content and actively choose to preserve it.”<sup>9</sup> The LOCKSS model aims to uphold publishers’ rights while carrying libraries’ tradition of custodianship into the digital environment.

### **Administering an enduring collection**

As digital archiving practices have progressed and matured, the need has become clear for overarching programs backed by firm institutional commitment to manage long-term storage and preservation. RLG/OCLC’s 2001 and 2002 documents on the *Trusted Digital Repository* delineate the place of the digital archive within the framework of a research organization and have become key texts in the field.<sup>10</sup> RLG and OCLC define the attributes of the “trusted repository” as follows:

1. *Administrative responsibility*: In all of its operations, the repository demonstrates commitment to best practices as established by the archiving community and accountability to stakeholders.
2. *Organizational viability*: In its stated mission, legal status, staff expertise, etc., the repository establishes its commitment to long-term preservation and access.
3. *Financial sustainability*: The repository has adequate budget and reserves, has a sound business plan, and follows good business practices.
4. *Technological and procedural suitability*: The repository follows rigorous review procedures regarding preservation strategies, staff expertise, suitability of hardware and software, etc.
5. *System security*: The repository protects the security of its holdings with appropriately designed technologies and clearly articulated procedures; procedures for disaster preparedness and response are in place.
6. *Procedural accountability*: the repository will document its practices and will be accountable to its stakeholders for its strategies, procedures, and functions [RLG/OCLC 2002; 13-15].

The viability of the DML requires that such criteria be met by the DML organization as well as the partner repositories that make up the distributed digital archives, embedded as these latter are in their parent institutions.

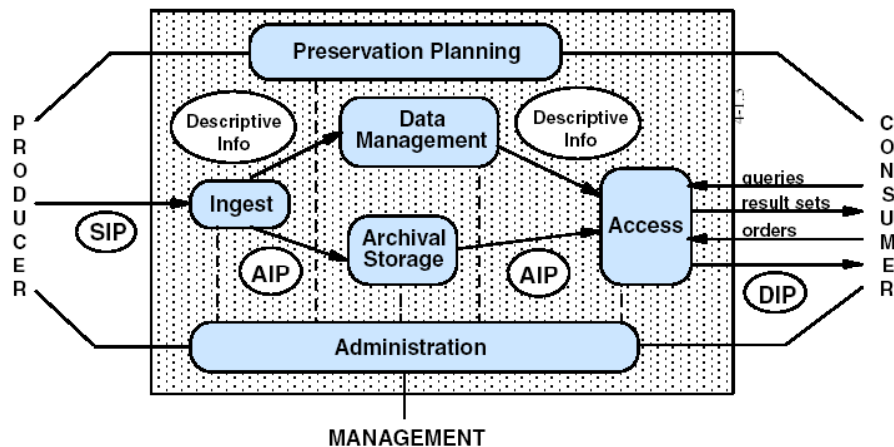
RLG and OCLC’s holistic picture of digital archives in the context of organizations builds on the generic archive framework proposed in the *Reference Model for an Open Archival Information System (OAIS)*.<sup>11</sup> First published in 2001, the OAIS Reference Model has quickly become the prevailing 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 provides the basic framework for

digital archives planning at the DML partner libraries at Cornell and Göttingen. The 2002 revision of RLG/OCLC *Trusted Digital Repository* document places a seventh attribute of the trusted repository at the top of its list: “Compliance with the *Reference Model for an Open Archival Information System (OAIS)*.”

The OAIS model posits six basic functions in the management of the long-term storage and retrieval of digital content: *Ingest*, *Archival Storage*, *Data Management*, *Administration*, *Preservation Planning*, and *Access*. Three of these – Ingest, Archival Storage, and Access – may be seen as stages through which content passes in the transfer between producer and archives, and between archives and user. These three stages correspond to distinct forms of *Information Package* in which content is held. The Ingest function governs the archives’ receipt of content from the information producer and the preparation of the materials for storage. In Ingest, content is received in the form of the Submission Information Package (SIP) and prepared for storage as Archival Information Package (AIP). The Archival Storage function manages storage and retrieval of content in the form of the AIP. The Access function entails dissemination of content to users in the form of the Dissemination Information Package (DIP). The Information Packages that are key to OAIS functions are constructed of metadata. The importance of standard metadata for digital archiving and for interoperability of a distributed archive cannot be overemphasized – the topic will be given due treatment in the report of the Metadata Working Group.

The remaining OAIS functions – Administration, Preservation Planning, and Data Management – encompass all stages of the archiving process.

### *OAIS functions*



### **DML and the OAIS model**

The DML will be a *distributed* digital collection of past mathematics literature; the digital archives at the heart of the DML will be “curated by a network of institutions,” according to the stated project vision. Interoperability of the components of the distributed collection and dissemination system will depend on accountability of the partner institutions to an agreed archive structure. Toward that end, in what follows, we summarize the OAIS functional requirements and propose to “map” them, in a preliminary fashion, to the needs of the DML.

## 1. Administration

The governance structure of the DML is still under discussion. But whether the DML emerges as a quasi-centralized organization or a looser confederation of autonomous projects, a coherent administrative entity will need to be in place to coordinate the operations of the distributed depository. This Administration function must entail close, long-term collaboration among personnel at the distributed nodes: the participating libraries. In the OAIS model, the archives Administration answers to a Management entity that is external to the archive and “set[s] overall OAIS policy as one component in a broader policy domain” [CCSDS 1-11]. For the DML, that relationship is doubled to the extent that the archive Administration answers to both the DML governing body and the institutions in which the archives’ distributed nodes are housed. The relationship among these bodies must be carefully negotiated.

A recent intervention by Kenney and McGovern concisely lays out options and basic requirements for inter-institutional distributed repositories:

[R]esponsibilities may be centralized or distributed, replicated in each member organization or provided through specialized assignments and modular development by each member, and managed in a micro or macro style, but the roles and responsibilities of the members must be explicit, accepted, current, feasible, effective, and coherent for such an amalgamation to be successful.<sup>12</sup>

It must be determined to what extent the DML will divide labor among the partner repositories – will each reproduce all the various archive functions, or will one partner specialize in Ingest, another in Storage, another in Dissemination, etc. This will no doubt depend in part on the outcomes of local efforts to secure needed funding. It may be that a more modular, specialized structure will develop after an initial period of many cooperative, but autonomous repositories.

Policies concerning collection scope, pricing, resource utilization, etc., will be set by the emerging DML governing body; the archives Administration would establish these in practice. Administration is charged with negotiating submissions agreements with information producers. It maintains a user service function that includes negotiating any user access agreements, billing, etc. Greenstein and Marcum propose the following minimum requirements with regard producer-archive agreements:

[A] repository will at a minimum require licenses that allow it sufficient control to accession, describe, manage, even transform deposited data (and accompanying metadata) for the sake of their preservation. Publishers may want to negotiate re-depositing when migration occurs. In any event, publishers must have the right to audit the contents of their deposited data. Where repositories act in association with one another (e.g. to ensure sufficient redundancy in the preservation process), they may also require rights allowing them to mirror or deposit data with other associated archives [Greenstein and Marcum 1].

The archives Administration establishes and must be empowered to enforce standards for metadata, format, documentation, etc. Administration maintains systems configurations for hardware and software. The specific here will flow from the recommendations of the DML Working Groups on metadata and technical standards.

## 2. Preservation Planning

Preservation Planning should be understood as a layer covering all phases and functions of the archiving process, rather than as a preliminary stage. OAIS designates monitoring of the user

community as a Preservation Planning function. For the DML, the user community overlaps with the governing body and to a certain extent with the archive Administration. An advisory board or similar entity should be charged with ongoing evaluation of the DML and assessment of the needs of mathematics researchers and other users. Preservation Planning also entails monitoring relevant technological developments and the ongoing development of preservation strategies and standards. Plans for data migration and disaster response are developed as part of this function.

### 3. Submission and pre-Ingest

Though not identified in the OAIS document, “pre-Ingest” functions have been elaborated in documentation from library and archive implementations of the OAIS model. RLG/OCLC’s *Trusted Digital Repository* report construes various processes that precede the initial acquisition of material as archival functions. These include the establishment of the content parameters of the digital archive as well as negotiations and determinations surrounding intellectual property rights.

Pre-Ingest processes also include validation of the integrity of the object and of the documentation/metadata accompanying the submission, as well as any preparation of the object necessary for Ingest. Digitization and OCR of print materials should be understood as pre-Ingest functions; since retrodigitization is a core function of the DML project, the pre-Ingest phase is of particular importance here. In the case of the DML, producer-archive negotiation over terms of storage, rights, and access takes place in preparation for digitization.<sup>13</sup> The fact that the DML will handle both retrodigitization and archiving means that archival and preservation standards can (and must) be brought to bear at the moment of the digital object’s production. In keeping with the project’s decentralized structure and the distributed nature of the proposed collections, the digitization phase of the DML will be handled by multiple local initiatives. It is likely that the individual partner institutions will further outsource the work of digitization and OCR in many cases. This added gradation in the division of labor will make uniform monitoring and enforcement of standards even more critical.

### 4. Ingest

The Ingest process prepares the retrodigitized object for archival storage. Presumably the partner institution that has arranged for the digitization of a particular digital object will in most cases be the one that will provide storage for that item. The movement from digitization to Ingest will thus be between units internal to an organization or between an outsourcing service and the institution that contracted the labor.

Preparation for archival storage entails a quality check of the received object. A unique identifier is assigned, along with other technical, descriptive, and preservation metadata established as standard for the archive. The OAIS model stipulates maintenance of *Preservation Description Information* consisting of the following<sup>14</sup>:

- Reference Information (indicates the scheme used to generate the object’s unique identifier)
- Context Information (records relationship of the object to associated objects in the archives)
- Provenance Information (documents history of digital object, including digitization history, alterations to object or metadata, changes in custody, etc.)
- Fixity Information (verifies authenticity of object, e.g. via digital watermark)

These pieces, together with digital object itself, constitute the Archival Information Package.

The Ingest process also prepares for the Access function by extracting descriptive metadata for the submitted object for use in searching the collection.

### 5. Archival Storage

The Archival Information Package is transferred from Ingest to permanent storage in the archives. Archival Storage involves maintenance of the archived objects through routine checks of data integrity, and data migration and refreshment of storage media, following a prescribed schedule. Archival Storage requires ongoing disaster recovery practices, including duplication of files. For the DML, a system of redundant storage of digital holdings among the partner repositories is part of a disaster recovery plan.

### 6. Data management

The data management function maintains the holdings records and catalog of the DML. This function includes control of user access to the collection, billing where appropriate, usage tracking, generation of statistics, etc.

### 7. Access / Dissemination

Access to the mathematics literature is the primary purpose of the DML. Dissemination of materials from archival storage entails the preparation of a Dissemination Information Package (DIP), which includes a copy of digital object, along with the appropriate metadata, and any necessary software. Integrity of the object is checked upon the generation of the DIP.

### **Cost and sustainability**

The costs of long-term digital archiving should not be underestimated. However, because of the rapid and unpredictable evolution of electronic media, those costs are difficult to calculate. In *Trusted Digital Repositories*, RLG/OCLC concedes that “not a great deal is known about the costs of preserving complex digital objects over time,” but emphasizes that “digital preservation will require ongoing resource commitments—potentially more than for traditional materials, but certainly different.”<sup>15</sup> The study goes on to cite a 2000 publication by Jones and Beagrie proposing four interrelated factors in digital preservation cost:

1. “The need to actively manage inevitable changes in technology at regular intervals and over a potentially infinite period”
2. “The lack of standardization in both the resources themselves and the licensing agreements with publishers and other data producers, making economies of scale difficult to achieve”
3. “The as yet unresolved means of reliably rendering certain digital publications so that they do not lose essential information after technology changes”
4. “That, for some time to come, the costs of digital preservation may be added to the costs for traditional collections, unless cost savings can be realized”<sup>16</sup>

Because it will create the digital objects that it archives in the retrodigitization phase, and inasmuch as the project is a collaborative venture with mathematics publishers, the DML may solve some of the problems associated with Jones and Beagrie’s cost factor 2. Although costs are linked to the volatility of digital formats and technologies, one should resist the temptation to postpone action on digital preservation, or to settle for half-measures, in the hope for a future stabilization of the technology. If the DML is to proceed, it must be with a commitment to long-term preservation from the beginning.

---

NOTES:

<sup>1</sup> Cedars Project, *Cedars Guide to the Distributed Digital Archiving Prototype*, March 2002: <http://www.leeds.ac.uk/cedars/guideto/cdap/>.

<sup>2</sup> CPA/RLG, *Preserving Digital Information: Report of the Task Force on Archiving of Digital Information*, 1 May 1996: <ftp://ftp.rlg.org/pub/archtf/final-report.pdf>.

<sup>3</sup> Research Libraries Group (RLG), *Trusted Digital Repositories: Attributes and Responsibilities* (Mountain View, CA: RLG, 2002) <http://www.rlg.org/longterm/repositories.pdf>.

<sup>4</sup> See, for example, Paul Wheatley, "Migration - a CAMiLEON discussion paper," *Ariadne* 29 (Sep. 2001): <http://www.ariadne.ac.uk/issue29/camileon/>. CUL performed an assessment of the risks of loss of information in the migration process: Gregory W. Lawrence, William R. Kehoe, Oya Y. Rieger, William H. Walters, Anne R. Kenney, *Risk Management of Digital Information: A File Format Investigation* (Washington, D.C.: Council on Library and Information Resources, June 2000) <http://www.clir.org/pubs/reports/pub93/pub93.pdf>.

<sup>5</sup> Jeff Rothenberg, "Avoiding Technological Quicksand: Finding a Viable Technical Foundation for Digital Preservation," *CLIR Publications* 77 (Jan 1999): <http://www.clir.org/pubs/reports/rothenberg/contents.html>; Stewart Granger, "Emulation as a Digital Preservation Strategy," *D-Lib Magazine* 6.10 (Oct. 2000): <http://www.dlib.org/dlib/october00/granger/10granger.html>.

<sup>6</sup> RLG 2002; 60.

<sup>7</sup> D. Greenstein and D. Marcum. "Minimum criteria for an archival repository of digital scholarly journals," version 1.2, 15 May 2000: <http://www.diglib.org/preserve/criteria.htm>.

<sup>8</sup> See, for example, Brian Cooper and Hector Garcia-Molina, "Creating Trading Networks of Digital Archives," *Proceedings of the First ACM/IEEE-CS Joint Conference on Digital Libraries* (New York: Association for Computing Machinery, 2001) <http://www-db.stanford.edu/~cooperb/pubs/dltrading.ps>.

<sup>9</sup> Victoria A. Reich, "Lots of Copies Keep Stuff Safe as a Cooperative Archiving Solution for E-Journals" 17 December 2002, LOCKSS, Stanford University: <http://www.istl.org/02-fall/article1.html>.

<sup>10</sup> The 2002 *Trusted Digital Repositories: Attributes and Responsibilities* was preceded by the Draft for Public of August 2001, *Attributes of a Trusted Digital Repository: Meeting the Needs of Research Resources*. (Mountain View, CA: Research Libraries Group) <http://www.rlg.org/longterm/attributes01.pdf>.

<sup>11</sup> Consultative Committee for Space Data Systems (CCSDS), Reference Model for an Open Archival Information System (OAIS), CCSDS 650.0-B-1, Blue Book, Issue 1, January 2002. <http://www.ccsds.org/documents/650x0b1.pdf>.

<sup>12</sup> Anne R. Kenney and Nancy Y. McGovern, "The Five Organizational Stages of Digital Preservation," forthcoming in 2003 a *Festschrift* for Wendy Lougee.

<sup>13</sup> On producer-archive negotiation see CCSDS, *Draft Recommendation for Space Data System Standards: Producer-Archive Interface Methodology Abstract Standard*, CCSDS –651.0-W-1, White Book, 28 December 2001: <http://ssdoo.gsfc.nasa.gov/nost/isoas/CCSDS-651.0-W-1.pdf>. Also, William R. Kehoe, "A Draft Of The E-Journal Archives Ingest Process," Cornell University Library, March 2002: <http://www.library.cornell.edu/iris/dpo/E-Archive-Ingest-process-draft3.pdf>.

<sup>14</sup> CCSDS 4-27–4-29. See also discussion in CPA/RLG 11-18.

<sup>15</sup> RLG/OCLC 2002

<sup>16</sup> Neil Beagrie and Maggie Jones, *Preservation Management of Digital Materials Workbook: a pre-publication draft*, October 2000, cited in RLG/OCLC 2002. The current version of Beagrie and Jones's handbook is at <http://www.dpconline.org/graphics/handbook/index.html> . For a discussion of cost modeling, see Arturo Crespo and Hector Garcia-Molina, "Cost-Driven Design for Archival Repositories," *Proceedings of the First ACM/IEEE-CS Joint Conference on Digital Libraries* (New York: Association for Computing Machinery, 2001) <http://www-db.stanford.edu/~crespo/publications/cost.ps> .