Chapter 3

Foundations and Standards

Preserving Digital Information

In 1994, the Commission on Preservation and Access (which later merged with the Council on Library Resources to form the Council on Library and Information Resources) and the Research Libraries Group (which later combined with OCLC) created a task force on digital archiving co-chaired by John Garrett and Donald Waters. The final report of the task force, *Preserving Digital Information,* was issued in 1996 and immediately recognized as a landmark document.

In addition to raising awareness of digital preservation as an imperative, the report spawned a number of short-term projects, and more important, introduced ideas which continue to be important, such as the need for a distributed network of preservation archives. The report is also worth singling out for its direct influence on the development of subsequent theoretical models. The discussion of the integrity of digital objects in relation to content, fixity, context, provenance, and reference clearly shaped the OAIS information model and propagated from there to current preservation metadata initiatives. The discussion of the need for repository certification to create trust in digital preservation led directly to subsequent efforts to define and measure characteristics of trusted digital repositories.

Reading

- Donald Waters and John Garrett, "Preserving Digital Information, Report of the Task Force on Archiving of Digital Information" (Commission on Preservation and Access, 1996) www.oclc.org/programs/ourwork/past/digpresstudy/final-report.pdf.

OAIS

The Open Archival Information Systems (OAIS) reference model is the core standard that provides the context for most work in digital preservation today. Originally developed by the Consultative Committee for Space Data Systems (CCSDS), OAIS became an ISO standard in 2003 (ISO 14721:2003) and has been embraced by the cultural heritage community. The term "OAIS" is generally used to refer to the standard, while "an OAIS" refers to a repository that conforms to the OAIS standard.

Essentially, OAIS does three things. First, it defines a common vocabulary for preservation-related concepts that anyone working in the field should know and understand. Second, it defines an information model for objects and metadata that an OAIS should support. Third, it defines a functional model for the activities that an OAIS should perform. The component responsibilities of six functional entities—Ingest, Data Management, Archival Storage, Dissemination, Preservation Planning, and Administration—are itemized in some detail. OAIS does not, however, prescribe how either the information model or the functional model should be instantiated in an operational repository. It is a high-level reference model that allows a good bit of interpretation in its actual implementation; so much, in fact, that nearly all preservation repositories claim OAIS conformance.

OAIS is too rich to be summarized here, but two good introductions are cited below. Here we will note only a couple of interesting points. First, OAIS introduced the requirement that an OAIS must define its particular user group (the "designated community") and take responsibility for making preserved content understandable to that community. This is not a traditional function of libraries, which tend to have broad, ill-defined, and heterogeneous

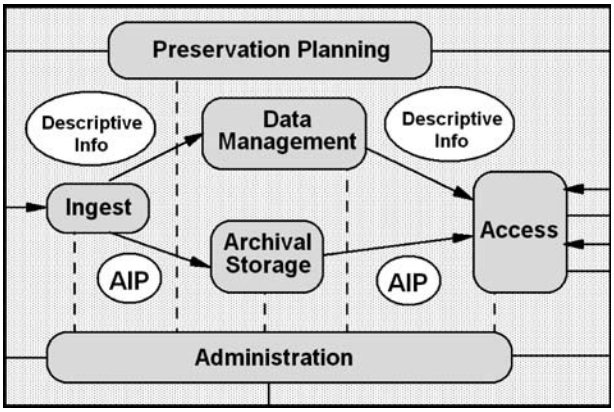

Figure 2
The OAIS functional model.

user communities. If a library stores geometry texts or French literature on its shelves, it does not thereby take responsibility for ensuring its card holders understand math or read French. The OAIS, however, must store and preserve as much information as necessary to ensure that its primary content will be understandable to its designated user community, given what that demographic can reasonably be expected to know now and in the future.

Second, the OAIS operates in an environment in which information packages are created, stored, and disseminated. A producer creates a Submission Information Package (SIP), which is the bundle of information the OAIS is given to archive. The OAIS reformats the SIP into an Archival Information Package (AIP) suitable for long-term storage and preservation. The AIP is accessed from outside the OAIS in the form of a Dissemination Information Package (DIP) created by the OAIS from the AIP. It is a useful model for repository design because the transformations between these three forms of information package make up a large part of the work of the repository. However, it does not fit particularly well within the life-cycle approach to preservation management, where many relevant preservation actions precede the creation of the formal SIP. Archives in particular have had trouble integrating OAIS into their own information life-cycle frameworks.

Anyone with more than a passing interest in digital preservation should take the time to get acquainted with OAIS. Nearly all subsequent work in the field uses OAIS terminology and presupposes some familiarity with the reference model.

Readings

- Consultative Committee for Space Data Systems, *Reference Model for an Open Archival Information System (OAIS),* Blue Book, Issue 1, Jan. 2002, http://public.ccsds.org/publications/archive/650x0b1.pdf. Also available as *ISO 14721:2003,* www.iso.org/iso/en/CatalogueDetailPage.CatalogueDetail?CSNUMBER=24683&ICS1=49&ICS2=140&ICS3.
- Alex Ball, "Briefing Paper: The OAIS Reference Model," Feb. 2006, www.ukoln.ac.uk/projects/grand-challenge/papers/oaisBriefing.pdf. A brief summary of OAIS that includes its influence on packaging languages, preservation metadata, and repository audit and certification.
- Brian F. Lavoie, "The Open Archival Information System Reference Model: Introductory Guide," *Digital Preservation Coalition Technology Watch Series Report 04-01,* Jan. 2004, www.dpconline.org/docs/lavoie_OAIS.pdf. A thorough overview, somewhat lengthier than Ball.

Preservation Metadata

Preservation metadata is information that supports and documents activities related to digital preservation. As such, it can be further defined as information that supports the process of ensuring the availability, identity, understandability, authenticity, viability, and renderability of digital materials. Although these activities require some descriptive and structural metadata, most preservation metadata falls into the category of administrative metadata.

In the library community, there have been two major streams of influence on the evolution of preservation metadata schemes. One, primarily theoretical, comes from OAIS, and the other, primarily practical, comes out of preservation research and development.

The OAIS information model provides a framework for defining preservation metadata. In OAIS an information package of any type consists of *content information* and *preservation description information.* Content information includes both the object to be preserved and the information necessary to interpret and understand it, which is called *representation information.* Preservation description information has four sub-types: *digital provenance,* which documents the origin and history of the object; *reference information,* which includes identifiers and other bibliographic description; *context information,* which includes information about the creation of the object and its relationships to other objects; and *fixity information* such as checksums. In this model, preservation metadata would comprise representation information and preservation description information.

The OAIS information model continues to influence metadata initiatives, especially in its detailed requirements for comprehensive representation information. It provides the theoretical basis for current projects such as the

Registry and Repository of Representation Information, which aims to implement a database of OAIS-conformant representation information, and the European CASPAR project, which focuses on the capture and use of representation information (see Chapter 7). At the same time, working preservation projects have found it useful to organize metadata not according to the category of information as in OAIS, but according to the type of entity it pertains to. This can be seen in the two dominant preservation metadata schemes in use today, PREMIS and LMER.

PREMIS is a data dictionary of "core" preservation metadata, where "core" is defined quite practically as what most preservation repositories are likely to need to know most of the time. PREMIS metadata is organized around four entity types: Objects, Agents, Events, and Rights. This means, for example, that PREMIS defines not one identifier but four: object identifier, agent identifier, event identifier, and rights (statement) identifier. Objects are further broken down into three types: files, bitstreams, and representations. Bitstreams are defined as data within a file that have common properties meaningful for preservation, while representations are sets of files needed to render a complete intellectual entity (for example, all the text and graphics files making up a Web page). By requiring the repository to associate each metadata element with the appropriate type of entity, PREMIS attempts to enforce a certain intellectual rigor.

While PREMIS is in use in most English-speaking countries, in Germany LMER (Long Term Preservation Metadata for Electronic Resources) is preferred. LMER is a standard of the German National Library and is based on a data model developed by the National Library of New Zealand. As with PREMIS, each metadata element is associated with a particular type of entity, which in LMER are objects, processes, files, and (the act of) metadata modification.

While a great deal of progress has been made in defining preservation metadata requirements over the last few years, at this point there are several important concerns about preservation metadata. First, nobody knows whether it works. That is, there has not been enough experience applying preservation strategies to know whether today's preservation metadata schemes actually support the process of long-term preservation. Second, neither PREMIS nor LMER define format-specific technical metadata, which is assumed be crucial. Only technical metadata for digital still images is formally standardized; specifications for audio, video, text, vector graphics, and other formats are in various stages of development (or not).[1]

Third, it is important that the values of preservation metadata elements can be supplied and processed automatically, as many preservation projects will be very large scale. Hand-entered, natural-language descriptions do not scale. However, there are few standard code lists or controlled vocabularies for the values of even the most important preservation metadata elements.

The OAIS information model requires packaging information to describe the various types of information package (SIP, DIP, and AIP). Most repositories use one of a number of standard container formats which allow various types of metadata and (optionally) content files to be bundled together. In the cultural heritage community, the most commonly used standard is the Metadata Encoding and Transmission Standard (METS).[2] METS is an XML schema that defines the structure of a digital object and has places for inserting descriptive and administrative metadata. METS distinguishes four types of administrative metadata: source, digital provenance, technical file information, and rights. It is possible to bundle PREMIS metadata in METS, but because the two standards are overlapping and somewhat orthogonal, it isn't straightforward. The Library of Congress and the Digital Library Federation have been working together to define best practices for using PREMIS and METS together. A best-practice guide, when available, will be published on the PREMIS Maintenance Activity Web site.[3]

Readings

- *Data Dictionary for Preservation Metadata: Final Report of the PREMIS Working Group,* May 2005, www.oclc.org/research/projects/pmwg/premis-final.pdf.
- "LMER: Long-Term Preservation Metadata for Electronic Resources," Deutsche Nationalbibliothek Web site, www.ddb.de/eng/standards/lmer/lmer.htm.
- Priscilla Caplan, "Preservation Metadata," *DCC Digital Curation Manual,* July 2006, www.dcc.ac.uk/resource/curation-manual/chapters/preservation-metadata. A little tedious but includes the archival stream of preservation metadata theory as well as the library side.

Trustworthy Repositories and Repository Certification

As noted above, one of the recommendations of *Preserving Digital Information* was to institute a dialogue "on the standards, criteria and mechanisms needed to certify repositories of digital information as archives." Four years later, OCLC and RLG convened a work group to define the attributes and operational responsibilities of a trusted digital repository. The group's 2002 report, *Trusted Digital Repositories: Attributes and Responsibilities,* spelled out a set of general characteristics and operational responsibilities based on the OAIS framework.[4] This report in turn

led directly to the creation in 2003 of a task force jointly sponsored by RLG and the (U.S.) National Archives and Records Administration (NARA) to refine these general responsibilities into specific goals and metrics that could be used in digital repository certification.

The RLG/NARA task force report, *An Audit Checklist to Support Digital Preservation,* was issued in a draft for public comment in 2005. The report itemized audit and certification criteria in four broad areas: organization; repository function, processes, and procedures; the designated community and the usability of information; and technologies and technical infrastructure.

While the draft was open for comment, the Andrew W. Mellon Foundation funded the Center for Research Libraries (CRL) to attempt to actually use the criteria in the *Audit Checklist* to evaluate a small number of preservation repositories run by third-party organizations. These included the National Library of the Netherlands, Portico, and the Inter-university Consortium for Political and Social Research (ICPSR). The project also looked at the LOCKSS system software. At the same time, the United Kingdom's Digital Curation Centre (DCC) used the *Audit Checklist* along with the *Catalog of Criteria for Trusted Digital Repositories,* a similar checklist developed by the German nestor project, as the basis for its own series of pilot audits of repositories in Europe, Australia, and the United States.[5] The DCC project took what it called an "evidence based approach" to the audit process and attempted to define the documentation that would support a repository's claims to meeting a particular goal.

Public comment, the CRL and DCC pilot projects, and other international input led to a revision and re-issue of the *Audit Checklist* in 2007 as *Trustworthy Repositories Audit and Certification: Criteria and Checklist* (TRAC).[6] According to TRAC:

> In determining trustworthiness, one must look at the entire system in which the digital information is managed, including the organization running the repository: its governance; organizational structure and staffing; policies and procedures; financial fitness and sustainability; the contracts, licenses, and liabilities under which it must operate; and trusted inheritors of data, as applicable. Additionally, the digital object management practices, technological infrastructure, and data security in place must be reasonable and adequate to fulfill the mission and commitments of the repository.

TRAC tightened up and reorganized audit criteria from the earlier draft and includes suggestions for evidence supporting compliance taken from the DCC audit process. It also includes a version of the criteria formatted as a genuine checklist for evaluation.

Meanwhile the DCC, in partnership with DigitalPreservationEurope, used its experience with the pilot audits to develop the Digital Repository Audit Method Based on Risk Assessment (DRAMBORA) tool kit.[7] Unlike TRAC, DRAMBORA does not itemize audit criteria but describes a methodology that repositories can use to identify their own objectives and assess the risks associated with them. The end result is a risk register which can be used to measure the repository's success in anticipating, avoiding, mitigating, and handling risks.

A working group of the International Organization for Standardization (ISO) has been formed under the auspices of CCSDS to produce an international standard on which a full audit and certification program for digital repositories can be based. As a first step, the group must determine the level of effort such a standard would require, and whether sufficient support for a standard exists. However, representatives of the U.S., U.K., and German efforts have acknowledged there is unlikely to be a single international certification process and that national variations in standards will exist. They have, however, agreed to ten points of broad common criteria to which all digital repositories should adhere:

- The repository commits to continuing maintenance of digital objects for identified community/communities.
- Demonstrates organizational fitness (including financial, staffing structure, and processes) to fulfill its commitment.
- Acquires and maintains requisite contractual and legal rights and fulfills responsibilities.
- Has an effective and efficient policy framework.
- Acquires and ingests digital objects based upon stated criteria that correspond to its commitments and capabilities.
- Maintains/ensures the integrity, authenticity and usability of digital objects it holds over time.
- Creates and maintains requisite metadata about actions taken on digital objects during preservation as well as about the relevant production, access support, and usage process contexts before preservation.
- Fulfills requisite dissemination requirements.
- Has a strategic program for preservation planning and action.
- Has technical infrastructure adequate to continuing maintenance and security of its digital objects.[8]

In the absence of a formal certification process, the Center for Research Libraries has taken responsibility for auditing digital repositories in the United States using

criteria appropriate to the institution and type of repository. Audits will focus on ascertaining that the repository actually does what it professes to and that it satisfies the expectations of its designated community.

Notes

1. ANSI/NISO Z39.87-2006, Data Dictionary—Technical Metadata for Digital Still Images Web site, www.niso.org/standards/standard_detail.cfm?std_id=731 (accessed Nov. 17, 2007).
2. Metadata Encoding and Transmission Standard (METS) Official Web Site, www.loc.gov/standards/mets (accessed Nov. 17, 2007).
3. PREMIS Preservation Metadata Maintenance Activity—Library of Congress Web site, www.loc.gov/standards/premis (accessed Nov. 17, 2007).
4. "Trusted Digital Repositories: Attributes and Responsibilities, An RLG-OCLC Report," May 2002 www.oclc.org/programs/ourwork/past/trustedrep/repositories.pdf (accessed Nov. 17, 2007)
5. nestor Working Group on Trusted Repositories Certification, "Catalogue of Criteria for Trusted Digital Repositories," version 1 draft for public comment, Dec. 2006, http://edoc.hu-berlin.de/series/nestor-materialien/8/PDF/8.pdf (accessed Nov. 17, 2007).
6. "Trustworthy Repositories Audit and Certification: Criteria and Checklist" (TRAC), version 1.0, Feb. 2007, www.crl.edu/PDF/trac.pdf (accessed Nov. 17, 2007).
7. DRAMBURA (Digital Repository Audit Method Based on Risk Assessment) Web site, www.repositoryaudit.eu (accessed Nov. 17, 2007).
8. Center for Research Libraries, Auditing and Certification of Digital Archives, Core Requirements for Digital Archives Web page, www.crl.edu/content.asp?l1=13&l2=58&l3=162&l4=92 (accessed Nov. 17, 2007).