Attributes of a Trusted Digital Repository:

Meeting the Needs of Research Resources

An RLG-OCLC Report

DRAFT FOR PUBLIC COMMENT

RLG

Mountain View, CA

August 2001

Executive Summary

In 1994, the joint RLG-Commission on Preservation & Access Task Force on Archiving of Digital Information began exploring the nature of a reliable repository for preserving digital materials. One of the recommendations of the 1996 CPA/RLG report was for the establishment of a certification program for digital archives or repositories.[1] Since then, key initiatives in digital preservation have advanced thinking and provided the necessary experience to begin to articulate the attributes of a trusted repository for digital research resources. In March 2000, RLG and OCLC began work on establishing these attributes, building on the soon to be international standard of the Open Archival Information Systems (OAIS) Reference Model. RLG and OCLC recognized that, despite emerging OAIS-related initiatives in Europe and Australia, consensus on the characteristics of a sustainable digital repository for large-scale, heterogeneous collections held by research repositories (e.g. research libraries and archives) was still needed.

The Digital Archive Attributes Working Group, convened in 2000, was charged with "defining the characteristics of reliable archiving services for heterogeneous research collections." Kelly Russell, former manager of the Cedars Project, drafted an initial document and the working group and invited experts contributed helpful, thoughtful suggestions. These were incorporated into this draft report, which aims to provide a framework for the implementation of reliable digital repositories. Building on work done since 1996, it focuses on the issues surrounding the long-term preservation of digital materials.

In this report, long-term preservation means two *distinct but equally important* functions: long-term maintenance of a bytestream and continuing access to its contents through time and changing technology. Given current understanding of how responsibility for long-term preservation will be assumed by research repositories, discussions and recommendations are based on a distributed model where different functions (e.g. deposit, storage, or access) can be provided across federated organizations.

The CPA/RLG report recommended "a dialogue among the appropriate organizations and individuals on the standards, criteria and mechanisms needed to certify repositories of digital information as archives."[2] This report answers that directive and:

- proposes a definition of a trusted digital repository (for community response/agreement);
- identifies the primary attributes of a trusted digital repository;
- articulates a framework for the development of a certification program;
- identifies the responsibilities of an OAIS-compliant digital repository; and
- makes several recommendations for follow-on work.

Interested parties are encouraged to read this report and contribute to the effort to build community consensus on the attributes of trusted repositories for digital research resources. Please send your comments to Robin Dale at Robin.Dale@notes.rlg.org by October 8, 2001.

Robin Dale, RLG

Meg Bellinger, Preservation Resources, OCLC

[1] John Garrett and Donald Waters, *Preserving Digital Information: Report of the Task Force on Archiving of Digital Information* (Commission on Preservation and Access and RLG, 1996) www.rlg.org/ArchTF/index.html.
[2] Ibid.

1 Introduction

In 1994, the joint Commission on Preservation & Access/RLG Task Force on Archiving of Digital Information began work to describe and explore the nature of a reliable repository for digital materials. The major findings of the CPA/RLG report included these key points:

- Long-term preservation of digital information on a scale adequate for the demands of future research and scholarship will require a deep infrastructure capable of supporting a distributed system of digital archives.
- A critical component of the digital archiving infrastructure is the existence of a sufficient number of trusted organizations capable of storing, migrating and providing access to digital collections.
- A process of certification for digital archives is needed to create an overall climate of trust about the prospects of preserving digital information.[3]

Work in these areas has been advanced by the Consultative Committee for Space Data Systems in its Reference Model for an Open Archival Information System (OAIS) and by the many groups and individual institutions that are designing their own digital repository systems. Since 1995, when the CPA/RLG report first appeared in draft, a great deal of work has been done and it is now possible to consider some of the outcomes and synthesize the learning from those recent efforts.

In March 2000, RLG and OCLC began a collaboration to establish attributes of a digital repository for research organizations, building on and incorporating the soon to be international standard of the OAIS model. RLG and OCLC recognized that, despite emerging OAIS-related projects and initiatives in Europe and Australia, a definition and consensus on the characteristics of a sustainable digital repository for large-scale, heterogeneous collections held by research libraries and archives was still needed. They formed a working group comprising digital preservation experts from around the globe who represented key research organizations involved in the long-term maintenance of digital materials (see Appendix C, Roster of the RLG/OCLC Working Group).

This report describes a framework for reliable repositories for digital content capable of handling the size, range, and type of materials held by research repositories. It builds on the foundations laid down in the CPA/RLG report, including its concept of a "deep infrastructure" and on the more recent work on the OAIS Reference Model, which provides a high-level, generic model for the environment, producers, users, data types, and information flows of a digital repository. This report applies this work to the specific situation of libraries and archives and identifies resources to assist institutions seeking or building repository services. It

- proposes a definition of a trusted digital repository (for community response/agreement);
- identifies the primary attributes of a trusted digital repository;
- articulates a framework for the development of a certification program;
- identifies the responsibilities of an OAIS-compliant digital repository;
- informs the RLG/OCLC communities of other developments necessary to implement a reliable repository; and
- provides formal recommendations for future work.

[3] Ibid.

Intended Audience

The guidance and recommendations included here are applicable to any organization interested in the long-term maintenance of and continuing access to digital materials. They are primarily intended for research organizations such as libraries and archives and are specifically aimed at research repositories. Although the report highlights some key strategic issues, its main focus is practical so that it can be useful to senior administrators as well as to those implementing digital archiving services. And while this report is concerned with technology, it is not a technical document. **It does not assume detailed technical knowledge on the part of the reader, but some basic level of technical understanding and a general awareness of digital preservation and related issues will be necessary.**

This report provides guidance relevant to local, regional, national, and, international efforts—digital preservation is not limited by geography. The coordination of digital archiving activities will be of paramount importance to ensure convergent rather than divergent development; future scholarship and research will depend heavily on interoperability and collaboration.

Terminology

Digital preservation interests a range of different communities, each with a distinct vocabulary and local definitions for key terms. It will be important to understand the meaning of a number of terms within the context of this document.

For the purposes of this report, "digital preservation" is defined as the managed activities necessary for ensuring *both* the long-term maintenance of a bytestream and continued accessibility of its contents. If discussion pertains specifically to one, a more precise term is used.

For the purposes of this report, an organization responsible for digital preservation activities is referred to as a "digital repository" or simply as a "repository." In practice, a digital repository may be a separate, independent organization, but this report assumes that it is more likely to be managed by an existing library or archive.

The term "archive" can mean different things to different people. To the frustration of many archivists, the term archive has been appropriated by others to mean many different things. In this report, "archives" is used to describe the professional community in general, that is, the library and archive community. The OAIS Reference Model uses "digital archive" to mean the organization responsible for digital preservation; this paper uses "archive" in place of "repository" *only* when "archive" is taken directly from the OAIS model.

Appendix B provides a glossary to clarify the exact meaning of terms in this paper.

2 Attributes of Trusted Digital Repositories

This section describes the attributes of a trusted digital repository and offers a series of increasingly granular illustrations, definitions, and explications. It provides scenarios of three digital repositories to help illustrate the ways that cultural repositories will respond to the need for a digital archiving service. Following those, it discusses notions of reliability and trust as applied to such organizations and the systems they manage. Next, it proposes a set of attributes that characterize reliable repositories and makes suggestions for the development of certification programs. Section 5 itemizes and analyzes the specific responsibilities of such repositories.

The focus here is to consider generally the needs of research repositories (e.g., libraries, historical societies, archives, museums, and the parent organizations within which these repositories operate), including a range of institutions holding, serving, and preserving a multiplicity of digital materials. The experiences of the research repository community serves as the first target for review; however, expertise from other communities such as computer science and government/military computing is also assessed and used to enrich and strengthen the case made for a proposed definition of a trusted digital repository.

What is a Digital Repository?

A digital repository is an organization that has responsibility for the long-term maintenance of digital resources, as well as for making them available to communities agreed on by the depositor and the repository. A few examples might help define a repository:

> **Scenario 1:** A national library responsible for ensuring long-term accessibility to large, diverse, and growing collections of digital resources, including online publications, complex multimedia products, the digital output of large imaging programs, and a range of special databases. The community it intends (or is mandated) to serve is extremely diverse and may be defined as anyone, anywhere, anytime with access to a contemporary, lowest-common-denominator personal computer.
>
> This repository may operate as part of a legal deposit environment, and the producer/creator community may include almost anyone: large commercial publishers that already supply print-based resources to the library; new commercial publishers; individuals engaged in vanity publishing; research networks establishing scholarly journals; digitization contractors; the institution's own staff; writers depositing papers, including computer files; etc.
>
> The national library may be also establishing collaborative distributed archiving of some classes of digital collections, such as online publications, with regional and higher education libraries and other memory institutions, as well as publishers, with different partners exercising different levels of archiving function and responsibility over varying periods.
>
> The digital repository is being built in-house using Artesia software and implemented both selectively and incrementally. The library also recognizes a need to take more comprehensive views of digital resources at specified points in their lifecycle.
>
> **Scenario 2:** A large university library with a growing collection of digital materials to support teaching and research, including online databases, electronic journals, digitized materials, digital output of university staff and students (e.g., theses and dissertations), digital course materials, and institutional records in electronic form. This repository serves primarily the university's faculty, staff, and students, but secondary users include the wider academic community and local community members who purchase library privileges. The university assumes that users will gain

access to digital collections on campus via the LAN or via the Internet using the high-speed academic network and a personal computer.

The producer/creator community for the university library, like the national library, may include almost anyone, from an individual scholar to a large commercial publisher. Locally, the university will have some control over the creation of digital materials, but most of it comes from producers over whom the library exercises little or no influence.

The digital repository is hosted through the university computer service, which is contracted to provide this service to the library. The system, currently under review, is being redesigned using IBM technology and is based on the OAIS Reference Model. Access to archived resources is available through the library management system.

Scenario 3: A museum with a growing collection of digital materials, including surrogates of museum objects, surrogates created for online exhibitions, and original digital art. The community the museum serves is very diverse and comprises students, researchers, artists, the general public, and organizations seeking digital material for commercial use.

The producer/creator community for the museum, unlike that of the national and university libraries, is generally individual artists over whom the museum has little control. The museum does have control in the creation of some digital surrogates (those created at the museum's request), though not all. The responsibility for archiving the digital materials for the long-term—regardless of creator—belongs to the museum alone.

Because the museum lacks technical infrastructure and qualified staff, it will contract with a third-party archiving service so that its materials will be professionally managed, controlled, and backed up. The commercial service is based on an Oracle platform and is OAIS-compliant. Access to materials in the repository is possible through seamless links between the museum's access management system and the repository.

These scenarios provide just three examples of the many different approaches that national agencies, research institutions, and cultural organizations will take to establish digital repositories. The infrastructure of the institution (a large university or national repository versus a small library, archive, or museum) will be a determining factor in the structure of the overall digital repository, but another factor will be equally important: the repository's **designated community**—its identified group of potential users—will determine what is deposited (content and format), how the digital information is managed and preserved, and how it is disseminated and accessed. Despite their different organizational models, all digital repositories will need to address the same underlying issues of infrastructure and functionality.

Reliability and Trusted Digital Repositories

RLG and the Commission on Preservation and Access published *Preserving Digital Information* in 1996. The report made a clear statement about trust in digital archives:

> For assuring the longevity of information, perhaps the most important role in the operation of a digital archives is managing the identity, integrity and quality of the archives itself as a trusted source of the cultural record. Users of archived information in electronic form and of archival services relating to that information need to have assurance that a digital archives is what it says it is and that the information stored there is safe for the long term.[4]

[4] Ibid.

Although much has been accomplished globally to move the digital archiving agenda forward in the years since the seminal report was published, there is as yet no collective agreement on a more exact definition of "trusted archive."

The archival and computer professions promulgate a host of concepts and terms that lay a foundation for establishing the defining characteristics of dependable digital archiving repositories. Commonly used terms such as *reliable*, *responsible*, *trustworthy*, and *authentic* help to define the nature of the archival enterprise and its myriad relationships with those creating, managing, and using digital objects. Computer scientists worldwide have grappled with definitions and performance measures of trusted military systems for nearly 20 years. Likewise, the airlines industry has required trustworthy, responsible, and authentic systems. In the last decade, groundbreaking work by archivists in Australia, North America, and Europe has resulted in fundamentally new approaches and tools that specify the nature and performance of accountable record keeping systems. And in the past few years, digital library experts have contributed their experience to a growing body of literature and applications pertaining to the construction and maintenance of secure systems accommodating large quantities of digital resources.

Archivists, digital library specialists, computer scientists, and others use the concept of reliability to express both expectations about and requirements for the nature of systems that contain and make available digital cultural materials. Archivists use the term *trust* to define records, record keeping systems, and archiving organizations while experts in the fields of digital libraries and computer science increasingly use it to distinguish between reliable and unreliable digital repositories. The phrase *trusted archive* appears often now in descriptions of projects and initiatives that are attempting to build real exemplars of such entities.

In an interview published last year in *RLG DigiNews*, Kevin Guthrie (president, JSTOR) opened his response to a question about JSTOR's plans for creating a trusted archive in this way:

> Well, trust in this context is both very important and rather difficult to define. It is important because the goal is to be able to establish a relationship whereby a library can rely on a third party to provide a service that has been a core function of a library; that is, archiving. That is no small responsibility and any enterprise that aims to provide such a service is going to have to earn a very high level of trust.
>
> At this early juncture, I don't think there exists a standard definition or "litmus test" of what it will take to become such a trusted archives. I do think that the mission of the enterprise is a fundamental component of the assessment.[5]

A quick survey of uses of the terms *reliable* and *trust* over the past several years can blaze a path toward clearer thinking about the nature of organizations that will undertake to offer long-term digital preservation services. At the same time, the language and concepts used over time to express expectations can be used to build toward a common definition that bridges the jargon and experiences of relevant expert communities.

A Guide to Understanding Audit in Trusted Systems was issued by the National Computer Security Center (NCSC) in 1988. This guide was a then-mature effort to identify the ways that trusted systems could be assessed and measured by third parties and focused very specifically on systems that processed classified military information. A trusted system had to be one that had built-in audit trails:

> A trusted computer system must provide authorized personnel with the ability to audit any action that can potentially cause access to, generation of, or effect the release of classified or sensitive information. The audit data will be selectively acquired based on the auditing needs of a particular installation and/or application. However, there must be sufficient granularity

[5] "Developing a Digital Preservation Strategy for JSTOR, an interview with Kevin Guthrie," *RLG DigiNews* 4, no. 4 (August 15, 2000) www.rlg.org/preserv/diginews/diginews4-4.html#feature1.

> in the audit data to support tracing the auditable events to a specific individual (or process) who has taken the actions or on whose behalf the actions were taken.[6]

And in 1992, the NCSC issued *Guidelines for Writing Trusted Facility Manuals*, in which a trusted computer system was defined as one "that employs sufficient hardware and software assurance measures to allow its use for simultaneous processing a range of sensitive or classified information." Two further definitions appearing in the document's glossary are equally illuminating:

> **Trusted Computing Base** (TCB): The totality of protection mechanisms within a computer system—including hardware, firmware, and software—the combination of which is responsible for enforcing a security policy. A TCB consists of one or more components that together enforce a unified security policy over a product or system. The ability of a TCB to enforce a security policy correctly depends solely on the mechanisms within the TCB and on the correct input by system administrative personnel of parameters (e.g., a user's clearance) related to the security policy.
>
> **Trusted Path**: A mechanism by which a person at a terminal can communicate directly with the TCB. This mechanism can only be activated by the person or the TCB and cannot be imitated by untrusted software.[7]

These NCSC documents augment a basic interpretation of *trust* with the concepts of *auditability*, *security* and *communication*. It should be noted that here *communication* refers to exchanges between a person and a machine. Later uses of the term denote different forms of exchange.

In 1996, the University of Pittsburgh's School of Information Sciences developed a set of *Functional Requirements for Evidence in Recordkeeping* in which a compliant *conscientious* record keeping organization was described as follows:

> Organizations must comply with the legal and administrative requirements for recordkeeping within the jurisdictions in which they operate, and they must demonstrate awareness of best practices for the industry or business sector to which they belong and the business functions in which they are engaged.[8]

Compliance and *auditability* are linked concepts here, establishing a direct and inarguable connection between performance and assessment. By using the term *conscientious* to describe an archiving organization, the authors affirm their belief that a repository must be meticulous in its operations if it wishes to act responsibly. Two new concepts appear in this reference: *compliance* and *conscientiousness.*

Mark Stefik, principal scientist in the information sciences and technology laboratory at the Xerox Palo Alto Research Center, has been writing about trusted systems for some years. In his 1997 *Scientific American* article, "Trusted Systems," and in an article published that same year in the *Berkeley Technology Law Review*, "Shifting the Possible: How Trusted Systems and Digital Property Rights Challenge Us to Rethink Digital Publishing," Stefik makes a series of points about ways to ensure the security and unchangeability of materials made available on the Web so as to protect both creators and distributors from unlawful copying by consumers:

> Computer scientists recognize trusted systems as those that follow rules that govern terms, conditions, and fees for using digital works.
>
> Knowing what the rules are is a central part of the design of trusted systems.

---

[6] *A Guide to Understanding Audit in Trusted Systems* (1998) www.radium.ncsc.mil/tpep/library/rainbow/NCSC-TG-001-2.html.

[7] National Computer Security Center, *Guidelines for Writing Trusted Facility Manuals, NCSC-TG-016, Version* (October 1992).

[8] *Functional Requirements for Evidence in Recordkeeping* (1996) www.sis.pitt.edu/~nhprc/prog1.html.

> Necessary security requirements vary based on the type of work being protected, but requirements for the most valuable works should include detection and prevention of tampering.
>
> One trusted system has to be able to recognize another trusted system.
>
> Copying can be permitted if it is strictly controlled in accordance with the creator's and/or distributor's and rights and interests (particularly those related to charging fees).
>
> Certification of trusted systems can ensure that they are compliant and can be relied on to follow appropriate rules and instructions.[9]

Stefik reinforces previously identified concepts of *security*, *communication*, and *compliance* but adds *certification*, *copying controls*, and *following rules* to the amalgam.

In her 1998 article, "Building Recordkeeping Systems: Archivists are Not Alone on the Wild Frontier," Margaret Hedstrom, associate professor in the School of Information, University of Michigan, surveyed the evolving field of electronic and hybrid (that is, involving paper-based and electronic source material) record keeping systems, suggesting numerous ways that the archival profession could take better advantage of work going on in the computer and digital library communities to develop trusted systems:

> Recent research and development efforts within the archival community combined with new methodologies for trusted systems provide archivists with a variety of tools to enhance the integrity, reliability, and usefulness of electronic recordkeeping systems...Recent research also illustrates that strategies and tactics for electronic recordkeeping rarely involve a simple choice between policy, standards, systems design and implementation. Rather, archivists and records managers need to pursue the right combinations of policies, standards, and system design methodologies that organizations can implement and that offer solutions which are affordable and commensurate with the risks and benefits involved.[10]

Hedstrom further posited that organizations seeking to build secure and trusted record keeping systems "are seeking trusted recordkeeping systems that follow rules for records creation, maintenance, and preservation at all times." And that:

> The expansion of electronic commerce into personal and retail consumption depends...on the ability of individuals and organizations to communicate and conduct business using trusted systems that are not predicated on prior established relationships or formal contractual agreements.

Hedstrom's points underscore the need for trusted systems that *follow rules* and that are surrounded by appropriate combinations of policies and standards so that *risk, benefit, and cost* are balanced. *Communication* (between people and businesses) and new tools developed for electronic commerce may well enhance our ability to build effective systems for long-term archiving services.

In a 1998 paper (revised in 2000) describing the Stanford Archival Vault prototype Stanford computer scientists Brian Cooper, Arturo Crespo, and Hector Garcia-Molina articulate the steps necessary to implement a reliable digital object archive.[11] Emphasizing replication strategies, the authors argue convincingly that their methodology "provides an extremely reliable storage infrastructure for preserving

---

[9] Mark Stefik, "Trusted Systems," *Scientific American* (March 1997) www.sciam.com/0397issue/0397stefik.html; Mark Stefik, "Shifting the Possible: How Trusted Systems and Digital Property Rights Challenge Us to Rethink Digital Publishing," *Berkeley Technology Law Journal* 12, no. 1 (Spring 1997) www.law.berkeley.edu/journals/btlj/articles/12_1/Stefik/html.

[10] Margaret Hedstrom, "Building Recordkeeping Systems: Archivists are Not Alone on the Wild Frontier," *Archivaria* (1998): 44-71.

[11] Brian Cooper, Arturo Crespo, and Hector Garcia-Molina, Hector, "Implementing a Reliable Digital Object Archive," in *Proceedings of the Fourth European Conference on Research and Development in Digital Libraries (ECDL)* (2000) dbpubs.stanford.edu/pub/2000-28.

digital objects, even as hardware, software and organizations evolve." The key reliability factors identified are:

- Avoidance of erasure (including deletion and overwriting by users) through write-once policies.
- Remote backup agreements that incorporate replication policies by the remote system provider.

A "reliability layer" within the distributed archival repository architecture encompasses a series of functions and mechanisms that the authors believe results in a reliable environment for preserved objects. Some of the functions identified include:

- Detection and restoration of missing/corrupted information.
- Communications among trusted components.
- User security, intellectual property management, query processing.
- Import/export facility to move objects into and out of the store.

Here is another bid for three definitional components: *communication*, *security*, and *replication. Communication* in this instance involves exchanges between parts of a system and also between federated member systems. The authors also introduce two new concepts: *backup policies* and *avoiding, detecting, and restoring lost/corrupted information*.

More recently, Cooper and Garcia-Molina expanded on their thinking about reliable digital object archives and discussed the benefits of creating replication partnerships that function as peer-to-peer trading networks.[12] In describing how these networks would interwork, the authors outlined the challenge of estimating the reliability of each partner and suggested factors that should be considered:

- Frequency of past failures (loss of data) as a predictor of future failures.
- Use of reliable hardware (disks).
- Presence of successful security measures.
- Reputation (perceived reliability).

Crespo and Garcia-Molina further refine their thinking about *reliability* and *trust* by identifying *reputation* and *performance* as important factors in determining trustworthiness.

In May 2000, the Digital Library Federation (DLF) proposed a set of *Minimum Criteria for an Archival Repository of Digital Scholarly Journals*.[13] Focusing on only one class of digital object (digital scholarly journals), their seven criteria include a mix of definitional and functional requirements. Two criteria refer to the defining characteristics of the organization itself:

> **Criterion 1**. A digital archival repository . . . will be a trusted party that conforms to minimum requirements agreed to by both scholarly publishers and libraries.
>
> **Criterion 2**. A repository will define its mission with regard to the needs of scholarly publishers and research libraries. It will also be explicit about which scholarly publications it is willing to archive and for whom they are being archived.

[12] Brian Cooper and Hector Garcia-Molina, "Creating trading networks of digital archives," in *1st ACM/IEEE Joint Conference on Digital Libraries* (2001) dbpubs.stanford.edu/pub/2001-23.

[13] Digital Library Foundation, *Minimum Criteria for an Archival Repository of Digital Scholarly Journals*, version 1.2 (May 2000) www.clir.org/diglib/preserve/criteria.htm.

The DLF placed trustworthiness at the center of its definitional requirements, although it did not specify exactly how the trust can or should be demonstrated. Further work, involving community experts, institutions, and publishers in designing trusted digital archives for journal publications, is now funded by The Andrew W. Mellon Foundation.

Trustworthiness here takes on a fuller aspect with the addition of notions about *organizational mission*, the stipulation of *agreements between creators and providers*, and the assertion that trusted repositories will *openly share information about what they are preserving and for whom*.

The US Defense Advanced Research Projects Agency (DARPA) described its Information Assurance and Survivability (IA&S) technologies while referencing a new solicitation for a Composable High Assurance Trusted Systems program in a report issued in March 2001.[14] In defining its overall aspirations, DARPA summarized its thinking about trusted systems in this way:

> Confidence in future systems must be achieved through system and network-level technologies involving approaches such as layered complementary mechanisms that will be cost-effective and scalable within three to five years. Proposed approaches must demonstrate the ability to support the advanced functionality of future trusted systems while maintaining a high level of confidence in the protection of these systems.

With this reference, the body of related terms and concepts is now enlarged with thoughts about *complementarity*, *cost-effectiveness*, *scalability*, and *confidence*.

Dr. Audun Jøsang, senior research scientist at Queensland University of Technology, has spent a good part of the past five years developing models and tools for evaluating trusted systems. One of his most innovative contributions to the field is his *A Metric for Trusted Systems* wherein he proposes a formal model for quantifying subjective beliefs about trustworthiness through the use of evaluative processes.[15] According to Jøsang, "Trust is a subjective belief [and] trust management for open computer networks . . . includes . . . the factors which influence users' trust in web sites and e-commerce." Jøsang's work involves the development of what he calls Subjective Logic and a trust inference engine based on Subjective Logic "to assist users and organisations to make trust assessments about remote parties on the Internet."[16] He suggests that security evaluation is a well understood method for determining trust in implemented system components and that a successful evaluation leads to the determination of an assurance level that reflects the trustworthiness of a system component.

Jøsang's work deepens a collective understanding of trust and at the same time provides tools to ensure that *evaluation of system components* is methodical and trustworthy in itself.

Trusted Digital Repositories: A Proposed Definition

A brief look back over a 15-year period highlights both remarkable similarities and encouraging growth in the way different sectors have matured their thinking about reliability and trust. The significant words and phrases from this limited group of experts are:

- Auditability, security, and communication.
- Compliance and conscientiousness.
- Certification, copying controls, and following rules

[14] Defense Advanced Research Projects Agency (DARPA), *BAA #01-24, Composable High Assurance Trusted Systems, CBD Reference*, (March 2001) www.darpa.mil/ito/Solicitations/CBD_01-24.html.
[15] A. Jøsang and S.J. Knapskog, *A Metric for Trusted Systems* (1998), available from citeseer.nj.nec.com/129647.html.
[16] Audun Jøsang, security.dstc.edu.au/staff/ajosang.

- Backup policies and avoiding, detecting, and restoring lost/corrupted information
- Reputation and performance.
- Agreements between creators and providers.
- Open sharing of information about what it is preserving and for whom.
- Balanced risk, benefit, and cost.
- Complementarity, cost-effectiveness, scalability, and confidence.
- Evaluation of system components.

These various definitions of responsible digital archives, record keeping systems, and computer networks can provide the substance for building blocks in the construction of a sensible and persuasive definition of a trusted digital repository. A proposed definition, incorporating many of the implicit and explicit assertions found in the literature of archival administration, computer science, and librarianship is:

> A reliable digital repository is one whose mission is to provide long-term access to managed digital resources; that accepts responsibility for the long-term maintenance of digital resources on behalf of its depositors and for the benefit of current and future users; that designs its system(s) in accordance with commonly accepted conventions and standards to ensure the ongoing management, access, and security of materials deposited within it; that establishes methodologies for system evaluation that meet community expectations of trustworthiness; that can be depended upon to carry out its long-term responsibilities to depositors and users openly and explicitly; and whose policies, practices, and performance can be audited and measured.

Attributes of a Trusted Digital Repository: A Proposed Framework

A definition, even one that can be agreed to by providers and consumers of these services, is not enough to provide guidance to those who wish either to select an archiving agency or to consider setting up and running one themselves. In order to distinguish reliable digital repositories from unreliable ones, attributes need to be identified which, when all are present, convince the community that a particular repository can be trusted with the long-term management of digital cultural materials. The following attributes represent a proposed framework for assembling the community's thinking about reliability and trusted archiving organizations:

- Administrative responsibility
- Organizational viability
- Financial sustainability
- Technological suitability
- System security
- Procedural accountability

Administrative Responsibility

A trusted digital repository will provide evidence that it has a fundamental commitment to implementing the range of community-agreed standards and best practices that affect its operations—particularly those that directly influence its viability and sustainability. Its reputation for reliability will be an indicator of its trustworthiness. Currently, the research repository community is basing systems and procedures on the

OAIS Reference Model; thus, the organization will commit to understanding the model and implementing aspects of it that reflect community-wide accepted practice. Administrative responsibility extends to meeting appropriate national and/or international standards for the physical environment (e.g., heating, ventilation, and air conditioning systems; proper shelving; fire-suppression systems; backup and recovery procedures; and security systems). The trusted repository will meet or exceed community standards for performance and will collect and share data measurements routinely with depositors. It will involve external community experts in validating and/or certifying its processes and procedures on a regular schedule. As a trusted partner in the long-term management of our digital cultural assets, a reliable repository will commit itself to transparency and accountability in all its actions.

Organizational Viability

Organizations choosing to become trusted digital repositories will establish themselves in ways that demonstrate their viability and trustworthiness. Their mission statements will reflect a commitment to the long-term retention and management of and access to digital cultural assets on behalf of depositors and users. Their legal status and standing will be appropriate to the range of responsibilities they are undertaking. Their business practices will be transparent and forthright. Staffing levels and areas of expertise will be appropriate to the work undertaken; further, staff training and professional development opportunities, including conference attendance and participation, will be given priority to ensure the currency of staff skill sets. The repository will establish management policies that reflect the commitments asserted in the mission statement. Written agreements with depositors will address all appropriate aspects of acquisition, maintenance, access, and withdrawal. Further, ongoing risk management and contingency planning will play a routine part of the organization's annual strategic planning activities. The repository will continually review its policies and procedures to ensure that appropriate growth can occur and that new processes and procedures are tested for scalability. And finally, a formal succession plan will be developed in consultation with community experts, depositors and peer organizations that identifies all relevant content and designates trusted inheritors should the repository cease to exist.

Financial Sustainability

A trusted digital repository will adhere to all good business practices and should have a solid, auditable business plan in place. Normal business and financial fitness should be reviewed at least annually to ensure the long-term sustainability of the enterprise. Standard accounting procedures should be used. Both short- and long-term financial planning cycles should be in evidence, demonstrating an ongoing commitment to seeking a balance of risk, benefit, investment, and expenditure. Operating budgets and reserves should be adequate for enterprises engaged in long-term operations serving a public good. Development opportunities for new sources of revenue should be explored routinely. All appropriate fiscal practices and conduct will be in place to ensure the sustainability of the repository.

Technological Suitability

Community experts currently advocate a range of preservation strategies. A trusted digital repository will consider all relevant options and will communicate openly about the suitability of variant strategies. It will ensure that it has in place all appropriate hardware and software to undertake the forms of acquisition, storage, and access it promises to make available. The repository will also have policies and plans for replacing technology as needed, including cycles of replacement and funding to achieve them. The repository will comply with all relevant standards and best practices, ensuring that staff have adequate expertise to understand and implement them. It will also undergo regular external audits on its system components and performance.

System Security

All systems used in the operation of trusted digital repository will be designed to assure the security of the digital assets managed there. Policies and practices will meet community requirements, particularly those pertaining to copying processes, authentication systems, firewalls, and backup systems. The repository will have written policies and plans for disaster preparedness, response, and recovery, and staff will be trained in carrying out appropriate responsibilities. Special attention will be given to processes that serve to avoid loss of data, detect changes in data, and restore lost or corrupted data. Any detected changes (including loss or corruption and restoration) will be documented and the depositor will be notified both of the changes and any resulting actions taken.

Procedural Accountability

A trusted digital repository is responsible for a range of interrelated tasks and functions (see Section 5 for details); it will therefore be accountable for all relevant policies and procedures. Repository practices will be documented and made available on request. Monitoring mechanisms that measure and ensure the continued operation of all systems and procedures will be in place. Preservation strategies undertaken (e.g., migration, emulation, etc.) will be recorded and justified in the context of community-wide best practices. Feedback mechanisms will be in place to support the resolution of problems and to negotiate the evolving requirements between archive providers and consumers.

Certification of Trusted Digital Repositories

In *Preserving Digital Information*, the CPA/RLG task force stated that "a process of certification for digital archives is needed to create an overall climate of trust about the prospects of preserving digital information."[17] The prospects of preserving digital information relate directly to the framework of trusted repository attributes outlined above: administrative responsibility, organizational viability, financial sustainability, technological suitability, system security, and procedural accountability. Unfortunately, no certification program or process covers these aspects, *in total*. Libraries and archives are left to hope that potential third-party archiving services adhere to relevant standards for data centers and computer rooms, but even those best practices lack crucial elements for trusted systems. Libraries and archives need a set of standards and/or best practices, criteria for assessment and measurement, and mechanisms to certify repositories of digital information as archives.

Types of Certification

At least two viable models for certification are in use and well known within the library and archives community: the audit model and the standards model.[18] The audit model is applicable to depositories holding government records, especially electronic records. In the US, such depositories must meet guidelines created by governmental agencies such as the Department of Defense or legislated by government.[19] The standards model operates in a variety of places throughout the library and archival community. Two examples of the standards model for certification would be guidelines for producing

[17] John Garrett and Donald Waters, *Preserving Digital Information: Report of the Task Force on Archiving of Digital Information* (Commission on Preservation and Access and RLG, 1996) www.rlg.org/ArchTF/index.html.
[18] Ibid.
[19] DOD 5015.2-STD *Electronic Records Management System Standard* jitc.fhu.disa.mil/recmgt/; National Archives and Records Administration, *NARA Regulations in the Code of Federal Regulations, Regulations in 36 CFR Chapter XII, Subchapter B - Records Management* (2001) www.nara.gov/nara/cfr/subch-b.html; National Archives and Records Administration, *Basic Laws and Authorities of the National Archives and Records Administration* (2000) www.nara.gov/nara//basiclaws.html; US House of Representatives, *Downloadable U.S. Code: Records Management By Federal Agencies,* 44 USC - Chapter 31 (January 2000) uscode.house.gov/title_44.htm.

preservation-quality microfilm and ISO interlibrary lending.[20] Institutions involved in preservation microfilming or interlibrary loan adhere to standards that appropriate agencies have certified as valid and appropriate. Peer institutions then "certify" the product or service by their acceptance and/or use of it. While both models work well, neither can completely address the range of activities, functions, and responsibilities associated with digital repositories.

In 1999, experts gathered at the Archival Workshop on Ingest, Identification, and Certification Standards (AWIICS) to begin developing standards specifically appropriate to the needs of digital repositories. Leading the discussion on certification, Bruce Ambacher, National Archives and Records Administration - Center for Electronic Records, identified four general approaches to certification: individual, archival program, process, and data.[21] These four approaches further refine the audit and standards models.

- **Individual:** Individual, professional certification or accreditation is sometimes referred to as personnel certification. In the context of traditional archival settings, certification of archivists is possible through a combination of education, work experience, and a competencies examination administered by the Academy of Certified Archivists (ACA). Nothing equivalent exists for electronic archiving or digital repository management.

- **Program:** Certification of a program or institution can be achieved through a combination of self-evaluation using standardized checklists and criteria and site inspections typical of program accreditation. Three models are the Society of American Archivists Evaluation of Archival Institutions, the HMC Approval from the Historical Manuscripts Commission (UK), and the Museum Assessment Program from the American Association of Museums. For these particular certifications, areas assessed include legal authority, governing authority, financial resources, staff, facilities, collection development, collection preservation, access, and outreach.

- **Process:** Process certification assesses methods and procedures that can be subjected to either quantitative or qualitative guidelines to guarantee adherence to all internal and external requirements. Some external standards that may be used to evaluate archival and perhaps digital repository processes include the ISO 9000 family, DoD 5015.2-STD, the Public Record Office Standard (Victoria), the Public Record Office (UK), and BSI DISC 0008.[22]

- **Data:** Data certification is concerned with two main aspects of the stored data: data persistence or reliability over time and data security. Certification for data persistence would include both internal and external quality control through processes such as ISO 9000:2000 and procedures manuals. It would also include documenting the processes used when migrating data, creating and maintaining metadata, updating data or files, and authenticating new copies. Issues related to data security have been addressed by the Public Key Certification Policy and Certification Practices Framework—because of the e-commerce boom, however, not because of the needs of digital archiving. This

---

[20] Nancy Elkington, ed., *RLG Preservation Microfilm Handbook* (Mountain View, CA: RLG, 1992); Nancy Elkington, ed., *RLG Archives Microfilming Manual* (Mountain View, CA: RLG, 1994); Interlibrary Loan Protocol Implementers Group (IPIG), *Profile for the ISO ILL Protocol: Version 2.0* (April 2001) www.nlc-bnc.ca/iso/ill/document/ipigwp/profile/ipv2_0.pdf.

[21] *Archival Workshop on Ingest, Identification, and Certification Standards (AWIICS): Draft Report* (1999) ssdoo.gsfc.nasa.gov/nost/isoas/awiics/.

[22] International Organization for Standardization (ISO), *ISO 9000/ ISO 14000*, www.iso.ch/iso/en/iso9000-14000/iso9000/iso9000index.html; Department of Defense, United States, *Design Criteria Standard For Electronic Records Management Software Applications*, jitc.fhu.disa.mil/recmgt/dod50152.doc; Public Record Office Victoria (Australia), *Standard for the Management of Electronic Records, PROS 99/00* (April 2000) www.prov.vic.gov.au/vers/standards.htm; Public Record Office (UK), *Management, Appraisal and Preservation of Electronic Records*, (1999) www.pro.gov.uk/recordsmanagement/eros/guidelines/; British Standards Institution, *Code of Practice for Legal Admissibility and Evidential Weight of Information Stored Electronically*, DISC PD 0008:1999 (1999).

framework, which was established to deal with user authentication and user communication in e-commerce transactions, handles access control issues for repositories and removes the need for additional data security certification.[23]

The participants in the AWIICS workshop agreed that, collectively, elements from each of these four certification processes could form a certification program that provides layers of trust. Such a layered approach should convey a high degree of confidence that the information an archive disseminates is the same as the information it ingested and preserved, with full documentation for all necessary modifications. A preliminary checklist for certification was created at the workshop and will serve as a tool for further work within the OAIS standardization activities.[24]

While work on certification has been delayed within the OAIS realm, both the checklist concept and the certifiable elements envisioned at the workshop provide a base for developing a certification framework.[25]

A Framework for Developing a Certification Program

Representatives from interested communities and expert stakeholders can and should develop a program for certifying trusted digital repositories. One way to frame the necessary development work is to break down the identifiable steps and illustrate each with examples from related fields and disciplines.

- Determine the need for a certifying body:

  Some certification programs are based on self-assessment while others depend on third-party examiners. Advantages and disadvantages of both should be weighed and a determination made as to which is most appropriate to this particular need.

  For most ISO standards, certification combines third-party examination and self-assessment. ISO does not check on implementations of their standards. Instead, partnerships are established with relevant communities and/or professional organizations and guides (checklists) are jointly developed to assess conformance to standards. Depending upon the standard, self-assessment may be all that is needed to judge compliance. In other cases, third-party examiners from professional organizations, private enterprise, or regulatory bodies use the checklists to assess compliance.

- Identify the attributes to be measured:

  If the proposed attributes (page 12) are used as the basis for a certification program, each component part will need to be analyzed and checklists or other measuring tools will need to be developed that allow for an objective assessment of compliance. In some cases, the translation of attributes into simple yes/no questions will be straightforward; in other cases, a range of checklists dovetailed with standard procedures and practices will need to be created.

  The preliminary checklist created at the AWIICS workshop can be used to assess a digital repository by both qualitative and quantitative measures. The checklist conflates several types of certification into a single tool, allowing assessment against a range of standards and best practices. This work, instituted through the OAIS initiative, is valuable and should be continued.

---

[23] Internet Engineering Task Force (IETF), *Internet X.509 Public Key Infrastructure Certificate Policy and Certification Practices Framework* (March 1999) www.ietf.org/rfc/rfc2527.txt.

[24] Archival Workshop on Ingest, Identification, and Certification Standards (AWIICS), *Certification (Best Practices) Checklist*, ssdoo.gsfc.nasa.gov/nost/isoas/awiics/CertifBase.ppt.

[25] The Certification group at the AWIICS workshop recommended that "accreditation of archives is important and should be pursued, but that it can only be accomplished when best practices are in place." Since the OAIS Reference Model was only in draft at the time, certification activities were delayed until the Reference Model was ready for ISO standardization; Donald Sawyer and Jerry Winkler, "Digital Archive Directions Workshop Extremely Successful," *NOST Hosts Archiving Workshop* 14, no. 4 (September 1998) nssdc.gsfc.nasa.gov/nssdc_news/sept98/01_j_garrett_0998.html.

- Specify the frequency or cycle of certification:

  Community representatives and expert stakeholders will need to agree on how long certification remains valid, and also the timeframe and associated processes for recertification.

  Community representatives and expert stakeholders interested in the long-term survival of preservation microfilm worked together to create national standards for its storage. In the US, these standards require that storage facilities are to be inspected at two-year intervals.[26] If deviations are detected or have occurred in the past, inspections are to be more frequent. Passing the inspection, which is itself also governed by a national standard, does not yield a certificate, but conveys a message of trust and responsibility through to the next inspection.

- Define the conditions for revocation of certification:

  In some certification programs, the seal of approval automatically expires after a set period if recertification process is attempted or if an attempt is unsuccessful.

  In the UK, the Historical Manuscripts Commission (HMC) is charged with tracking the existence, location, and nature of manuscripts and records for the study of British history. With that responsibility, the HMC inspects record repositories to make sure repositories are meeting the *HMC Standard for Record Repositories*.[27] Failing to conform to their HMC Standard for Record Repositories can lead to several consequences for a repository: loss of HMC certification, loss of further document deposit with the repository; and finally, failing to fix problems and regain certification can lead to removal of all documents to another approved repository.

[26] American National Standards Institute, *ANSI/PIMA IT9.11-1998 Processed Safety Photographic Films - Storage* (NY: American National Standards Institute, 1998).
[27] Historical Manuscripts Commission, *Audit Inspections and Approval of Record Repositories* (March 2000) www.hmc.gov.uk/advice/inspect.htm.

3 Responsibility and Digital Preservation

If the future of scholarship is to be secured, research repositories need to understand fully what responsibility they should assume for the preservation of digital materials. Responsibility must be understood at three basic levels. Organizations must first understand their own local requirements. Second, they need to understand which other organizations might share some of the responsibilities through geography or existing arrangements such as consortial agreements or shared user communities, disciplines, or format of materials. Third, they need to understand which responsibilities can be shared and how. Assuming that the general model for digital repositories is more or less distributed, its success relies on shared understanding across the federation or network of repositories of their respective duties and roles. Comprehensive coverage within the collections and effective interoperability across repositories will rely on such understandings.

Although a detailed discussion of the complex issues of responsibility is beyond the scope of this report, a summary of the major factors is useful.[28]

The Scope of Collections

Digital materials for libraries and archives range from simple (e.g., text-based) digital files to complex multimedia and database resources. The sheer variety of digital materials and the role that they play in the collection make development and application of collections policies very challenging. The existence or lack of a physical equivalent or counterpart influences decisions about whether and how the digital resource is preserved. For materials that have a physical counterpart, preservation decisions take into account considerations such as the condition of the original materials and the reason for digitizing (e.g., for increased access to the materials). Materials that are "born digital" can present more challenging problems because their "being digital" is not only a method of access, it represents their value as an information artifact. For many born-digital resources, effective preservation will rely as much on preservation of the object's digital characteristics or properties as on preservation of its basic intellectual content. More importantly, when a library or archive digitizes its own collections, it can control decisions about standards, formats, quality control, and documentation. The preservation of materials generated outside may not include this degree of control.

Preservation and Lifecycle Management

Preservation decisions for digital items cannot wait until continued use of the materials has proved they are worth keeping. Postponing preservation decisions can and most often will result in preservation actions that are more complex, more labor intensive, and more costly. A resource can even be held hostage by an obsolete piece of software. It is also important to accept the fact that digital information is more transitory and mutable, so there is little likelihood of its surviving through benign neglect. Preservation requires active management that begins at the creation of the material and depends on a proactive approach by digital repositories and the cooperation of the stakeholders, including data providers.[29]

[28] For a more complete discussion of the roles and responsibilities of different stakeholders in the lifecycle of digital materials, see Neil Beagrie and Daniel Greenstein, *A Strategic Policy Framework for Creating and Preserving Digital Collection*, Version 4.0, Final Draft, (1998) ahds.ac.uk/strategic.pdf.
[29] Ibid.

The Wide Range of Stakeholders

Content creators, systems developers, custodians, and future users are all potential stakeholders in the preservation of digital materials, and this complicates the determination of responsibilities—who, when, and for how long. Often, those creating digital materials or designing digital content management systems do not take great interest in their long-term preservation. For example, commercial publishers are justifiably interested in the preservation of their materials only as long as they are commercially viable, while libraries and their users are often interested in continued access to materials long after they cease to turn a profit. Similarly, for an archive, it is usually when an electronic records management system is designed (well before any records are created) that key decisions are made that affect the long-term preservation of the records themselves. In both cases, decisions about how the materials are handled when created or maintained determine how or whether the repository can preserve them.

Ownership of Material and Other Legal Issues

Responsibility for preservation has traditionally been considered alongside ownership of the materials; that is, the owner of the materials was responsible for determining their lifespan. However, ownership of digital materials is not often straightforward. While a book can be taken into the collection and set upon the shelf, digital materials are less tangible. For a growing number of digital materials considered "integral" to research collections and archives, access is provided through licensing arrangements—often through a "deal" with a regional or national consortium. Licensing arrangements can apply to either the digital content itself or to software necessary for specific functionality and access to the content. Although the organization may own the right to access material or use the software for a specified period, there is often no guarantee of rights beyond the terms of the license. While commercial publishers are beginning to provide some guarantee of continuing access, most licensing agreements are still perilously vague about how the digital repository will be maintained and how long-term access will be ensured. Reliance solely on creators or producers of digital materials for long-term preservation is potentially risky, not least because digital resources are not generally created or engineered with long-term preservation in mind.

It will be critical in the future for research repositories to work as closely as possible with content creators to ensure that long-term preservation responsibilities are clearly understood and documented in licensing agreements; this is currently being explored by The Andrew W. Mellon Foundation's e-Journal archiving program.[30] It will require increased cooperation and effective communications with publishers, software suppliers, and other producers to ensure that what is deposited is a copy of the data object in the format most suitable for preserving the materials over the long term. In this situation, understanding the important difference between long-term preservation and short term access—particularly while materials are still commercial viable—is critical. Libraries may require different license arrangements for long-term preservation than for end-user access.

Often, rights that relate to the software and systems used to create the material impinge on its preservation. Very little, if any work, has been done with software vendors to raise awareness about the longevity of their materials in the interests of future scholarship and research.

[30] A program funded by The Andrew W. Mellon Foundation designed to plan the development of e-journal repositories meeting specific requirements developed by the Digital Library Federation (DLF), Council for Library and Information Resources (CLIR), and Coalition for Networked Information (CNI). Seven major libraries have received grants, including the New York Public Library and the university libraries of Cornell, Harvard, MIT, Pennsylvania, Stanford, and Yale; see www.clir.org/diglib/preserve/ejp.htm.

Digital preservation has even wider legal implications. How preservation infringes on copyright remains unclear. For example, the content creator does not usually own the rights to the software and systems used to create the digital file. This raises legal issues when access or changes to those systems are necessary. In such cases, at best, a repository will need to arrange separate rights clearance for long-term maintenance; at worst, preservation will be compromised because rights clearances for access cannot be obtained. Some work has been done on the establishment of repositories for software to help address these concerns, however the research repository community will need to make an appeal to have this conflict taken into consideration in the creation or renewal of national deposit legislation.

Cost Implications

Although not a great deal is known about the costs of preserving complex digital objects over time, there is an accepted wisdom in the library community that digital preservation will require ongoing resource commitments—potentially more than for traditional materials, but certainly different. Traditional and digital preservation should be compared with some caution, because the complex dependencies between long-term maintenance and continuing access make comparison problematic. Indeed, for digital materials that have no analog equivalent, comparison is meaningless. Although it may be too early to compare the costs of digital and traditional preservation meaningfully, one thing is certain: preserving digital materials will require resource commitments over time. While traditional materials, for example, may have ongoing costs for stable storage environments, digital materials will also require periodic analysis and the application of new technical strategies to ensure continuing access. Digital preservation is also likely to draw on resources longer than traditional preservation does, and it may be the case that different technical strategies (e.g., different types of migration or emulation) will prescribe quite different costing timeframes and schedules.

In a recent publication entitled *Preservation Management of Digital Materials*, Jones and Beagrie suggest that digital preservation costs are based on four interrelated factors:

- The need to actively manage inevitable changes in technology at regular intervals and over a potentially infinite period.
- 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.
- The as yet unresolved means of reliably rendering certain digital publications so that they do not lose essential information after technology changes.
- 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.[31]

Digital technologies and applications shift rapidly; strategies to preserve objects resulting from new approaches must keep pace. The inherent mutability of digital preservation therefore makes it difficult, if not impossible, to establish concrete costs for all associated activities. Further, the uncertainty of the financial commitment represented by digital preservation makes assuming preservation responsibilities more complex. However, although work can be done to understand how costs will play out and where saving can be made, the preservation of digital materials cannot wait for exact information because it may never appear. What will be important is an understanding of where the main costs are likely to fall and how, within existing practices, these can be incorporated to achieve economies of scale. In addition, that ways that other stakeholders (e.g., content providers) can decrease costs (e.g., by changing the way they

[31] Neil Beagrie and Maggie Jones, *Preservation Management of Digital Materials Workbook: a pre-publication draft*, (October 2000) www.jisc.ac.uk/dner/preservation/workbook/.

supply materials for deposit—in specific formats or with better descriptions) should be explored. Repositories that currently provide guidelines for depositors include the Cornell University Library, Arts and Humanities Data Service (AHDS), and the National Library of Australia.[32] Repositories will also need to understand more about the advantages of collaborative approaches. Although, if all participants' costs are totaled, distributed archiving is unlikely to cost less than using a single, centralized repository, a shared operation will cost less per organization; that is, the costs may be easier to absorb across multiple institutions.

Research repositories need to begin work now—"this is likely to be a better strategy than only discussing and studying the problem."[33] Integrating digital preservation into the everyday management and organization of the library or archive will help ensure that the necessary skills and knowledge are embedded within the organization for the earliest and most effective savings or economies of scale.

[32] *Report of the Digital Preservation Policy Working Group on Establishing a Central Depository for Preserving Digital Image Collections*, Version 1.0 (March 2001) www.library.cornell.edu/imls/image%20deposit%20guidelines.pdf; AHDS provides guidelines for each of its service providers in Visual Arts, Performing Arts, Electronic Texts, History, and Archaeology, www.ahds.ac.uk/dephow.htm; National Library of Australia, *Safeguarding Australia's Web Resources: Guidelines for Creators and Publishers* (2000) www.nla.gov.au/guidelines/2000/webresources.html.

[33] Johan Steenbakkers, *The NEDLIB Guidelines: Setting up a Deposit System for Electronic Publications*, NEDLIB Report Series, report 5 (NEDLIB Consortium, 2000), available from www.kb.nl/nedlib.

4 Deep Infrastructure and OAIS

The Need for Deep Infrastructure

In 1995, the CPA/RLG report recognized the need for "a deep infrastructure capable of supporting a distributed system of digital archives." The report went on to suggest:

> Effective structures for digital archives in a distributed network will surely take various forms and will include corporations, federations and consortia, each of which may specialize in the archiving of digital information and range over regional and national boundaries. . . . Moreover, shared interests in, for example, intellectual discipline, in type of information, in function, such as storage or cataloging, and even interests in the output of information within national boundaries will all form a varied and rich basis for the kinds of formal and informal interactions that lead to the design of particular archival organizations.[34]

Much work remains to be done to understand models for this "deep infrastructure." Responsibility for digital archiving might be distributed in a variety of ways, involving collaboration on a variety of levels, from simple shared access to holdings to formal collaborative collections management and shared labor. However, several projects and organizations have implemented digital repositories that have proven not only feasible and practical, but able to offer important savings. The outcomes of both the Cedars Project and the NEDLIB Project have highlighted key infrastructure requirements.

- **The Cedars Project** incorporates three geographically distributed partner sites in the UK: the universities of Oxford, Cambridge, and Leeds.[35] These three sites and a handful of test sites acted as both content providers and end users of the "demonstrator archive," while a Web-based front end was developed to, in theory, allow access from any number of additional sites. In practice, access restrictions on materials used in the demonstrator limited testing to the partners and test sites, but this still allowed the development of a proof-of-concept digital repository across a total of nine different sites. A report on the Cedars Project is imminent.[36]

- **The NEDLIB Project** involved nine European national libraries and was led by the Royal Library of the Netherlands in The Hague, where the NEDLIB demonstrator system was implemented. It focused on needs of national libraries as they extend their role to take responsibility for the preservation of digital materials and to establish a Deposit System for Electronic Publications (dSEP). The NEDLIB Process Model can be distributed and "aims to develop a common architectural framework and basic tools for building such DSEP systems," incorporating the DSEP functions into existing library practices for handling digital materials.[37] Early in the project, NEDLIB adopted the OAIS model and effectively mapped it onto their existing model for a DSEP, thereby providing an implementation of OAIS that was situated within a wider digital library environment. The main concern NEDLIB raised about using OAIS in this context was its lack of explicit functions relating to strategies for continuing access and how these might be chosen

[34] John Garrett and Donald Waters, *Preserving Digital Information: Report of the Task Force on Archiving of Digital Information* (Commission on Preservation and Access and RLG, 1996) www.rlg.org/ArchTF/index.html.
[35] Cedars (CURL exemplars in digital archives) is led by the Consortium of University Research Libraries (CURL) in the UK and funded by the UK Joint Information Systems Committee through its Electronic Libraries Programme (eLib).
[36] The Cedars Project Report, April 1998–March 2001, www.leeds.ac.uk/cedars/OurPublications/cedarsrepmar01exec.html.
[37] Titia Van der Werf, *A Process Model*: *The Deposit System for Electronic Publications,* NEDLIB Report Series, report 6 (NEDLIB Consortium, 2000), available from www.kb.nl/nedlib.

> and effectively implemented in an OAIS repository. To this end, NEDLIB reworked the OAIS model itself and initiated changes that take this preservation perspective into account.[38]

While the early work of such projects suggests that a distributed system is both feasible and desirable, it also demonstrates the need for a high level of convergent development and the adoption of agreed-on standards to ensure the necessary systems supports and deep infrastructure recommended by the CPA/RLG report. A system for the certification of digital repositories would undoubtedly be useful. However, it would need to acknowledge varying degrees of "federation" within the broad term "distributed repository." That is, some collaborative repositories will be more integrated than others, sharing workflows at the lowest levels, and will rely more heavily on adopting standard practices. Currently, there are few available models for collaboration between digital repositories, and more work is needed to understand the possible levels of cooperation and the required levels of standardization.

The Open Archival Information System Reference Model

The functions of a digital repository parallel those of traditional repositories. Broadly speaking, these functions describe how materials are submitted to the repository, how they are organized and managed within the repository, and, how continuing access to them is provided. One of the greatest challenges in accepting preservation responsibility within an organization is finding a shared vocabulary for stakeholders with a variety of backgrounds to use for productive discussion of the issues. **Effective digital archiving services will rely on a shared understanding across the necessary range of stakeholders of what is to be achieved and how it will be done.**

In the past few years, a number of organizations and projects have adopted a reference model developed by the Consultative Committee for Space Data Systems called the Open Archival Information Systems (OAIS) Reference Model.[39] Despite its origins in the space data community and its initial application to satellite and GIS data, the OAIS model has attracted widespread interest in the international library and archive communities. The OAIS model can be applied to both traditional and digital repositories, but was designed primarily for a digital context. It provides a set of well-articulated concepts and a comprehensive, if daunting, vocabulary that can facilitate communications between stakeholders with a range of technical backgrounds and between disciplines.

The OAIS model is oriented to a situation where a particular type of data is closely aligned with a designated community and where the archives tends to form around the intersection of homogeneous data resources and users with very similar data needs. That this is not the situation in archives and research libraries has been cited as the major challenge to extending OAIS as a generic model. However, libraries and archives have been involved in the development of the model as it has evolved into an ISO standard and the model has a great deal to offer the library and archive community.

OAIS provides both a **functional model**—the specific tasks performed by the repository such as storage or access—and a corresponding **information model** that includes a model for the creation of metadata to support long-term maintenance and access.

As the OAIS Reference Model points out, a repository consists "of people and systems." Scalability of the model depends on the systems. Although this report does not recommend which OAIS functions might be automated, this is a key area for future work. Implementation of an OAIS-style archive on any significant scale will require automated routines for handling many of the functions.

---

[38] Titia Van der Werf, ed., *The NEDLIB Report Series* (NEDLIB Consortium, 2000), available from www.kb.nl/nedlib.

[39] Consultative Committee on Space Data Systems, *Reference Model for an Open Archival Information System (OAIS)*: Draft Recommendation for Space Data System Standards, CCSDS 650.0-R-1, Red Book, May 1999, www.ccsds.org/RP9905/RP9905.html.

Adoption of OAIS by Libraries and Archives

Many libraries and archives have begun to use the OAIS Reference Model as a model for the systems they implement or simply as a checklist for systems that are already in place or under development. Many deposit libraries, which have a legal obligation to take custody of materials published electronically, have been particularly enthusiastic about its adoption. National libraries and key research repositories with immediate preservation responsibilities have been some of the first to adopt OAIS:

- The NEDLIB Project has used the OAIS model as a basic point of reference for developing the DSEP (Deposit System for Electronic Publications).
- The British Library has based their Digital Storage Project on OAIS.
- In the US, OAIS forms the basis for development work at the Library of Congress, the National Archives, Harvard and Stanford Universities, RLG, and OCLC.

The National Library of Australia has used OAIS as a generic model to validate the functions and relationships in the Pandora Archive. OAIS development has been informed by the practical work of these projects as well as by the commitment and enthusiasm of organizations such as RLG and OCLC. The library/archive perspective has played a key role in bringing the model to the brink of adoption as an ISO standard. The NEDLIB consortium suggests an advantage of OAIS:

> From the start it was recognized that by applying the OAIS-Model, deposit libraries could benefit from the advantages of international standardization. By using a common reference model, a common terminology and a common conceptual framework, it is much easier to share ideas and exchange experiences. Not only between deposit libraries, but also across institutional boundaries, for example, between libraries and archives. . . . [Furthermore,] the longer-term it is hoped that IT-vendors and system developers will adopt the OAIS-framework as a basis for implementing deposit systems and for developing ready-to-market products. This would facilitate open systems development for the benefit of a much larger community than would be the case if archival institutions invested in tailor-made systems on an individual basis.[40]

The "open" nature of the OAIS model is one of its greatest strengths. **It is not a model for systems design but simply a reference model**. Implementations of the OAIS model using different operating systems can look quite different, but they are still based on the same concepts and standards. For many organizations that need to implement digital archiving as part of a complete digital library service, OAIS would allow for mapping of existing library service functions onto the model. It is important to realize that the functional entities described in the model do not, on implementation, have to remain separate, isolated functions. For example, a library may wish to use existing acquisitions and cataloging processes and procedures for what OAIS calls Ingest activities. Library staff may therefore consider enhancements to existing systems that allow production of the necessary metadata and procedures for materials to be, in some cases, established or deposited both within the current library management system as well as into the archival store. Access to archived materials can then also be provided through the library management system rather than through a separate dissemination system.

[40] NEDLIB Contribution to the Review of OAIS (June 2000) www.kb.nl/coop/nedlib/results/OAISreviewbyNEDLIB.html.

5 Responsibilities of a Trusted Digital Repository

The CPA/RLG report recommended "a dialogue among the appropriate organizations and individuals on the standards, criteria and mechanisms needed to certify repositories of digital information as archives."[41] Earlier sections of this report expand on how a certification program could be shaped. The OAIS model provides a useful framework from which to begin the discussions, lending depth and breadth to the proposed definition of a trusted digital repository and its associated attributes (see section 2). If a program of certification is to be established and applied in a variety of contexts and across a range of organizations, a deeper understanding of tasks and functions must be developed. This section represents a list of responsibilities initially based on the synthesis of existing work and then rigorously discussed and debated by the working group members—international experts in digital preservation. The responsibilities are intended to provide stepping stones for organizations charged with preservation of digital resources towards the establishment of trusted archiving services.

The following list of responsibilities is taken from work done by the OAIS community to define the principle obligations of an OAIS compliant repository. However it should be recognized that the application of these responsibilities to a digital repository extends beyond an implementation of the OAIS model and will therefore be of use to repositories more generally. This list is based mainly on the OAIS approach but includes one addition, acknowledging the repository's critical role in the promotion of standards in the area.

The OAIS model posits that a reliable digital repository:

- Negotiates for and accepts appropriate information from information producers and rights holders.
- Obtains sufficient control of the information provided to support long-term preservation.
- Determines, either by itself of with others, the users that make up its designated community, which should be able to understand the information provided.
- Ensures that the information to be preserved is "independently understandable" to the designated community; that is, that the community can understand the information without needing the assistance of experts.
- Follows documented policies and procedures that ensure the information is preserved against all reasonable contingencies and enables the information to be disseminated as authenticated copies of the original or as traceable to the original.
- Makes the preserved information available to the designated community.

In addition, given the importance of promoting standards for preservation, a responsible repository:

- Works closely with the repository's designated community to advocate the use of good and (where possible) standard practice in the creation of digital resources; this may include an outreach program for potential depositors.

[41] John Garrett and Donald Waters, *Preserving Digital Information: Report of the Task Force on Archiving of Digital Information* (Commission on Preservation and Access and RLG, 1996) www.rlg.org/ArchTF/index.html.

Repository Responsibilities

Negotiating for Appropriate Information from Information Producers

This responsibility refers to all transactions between the content providers (including creator, rights holders, etc.) and the repository prior to formal submission into the repository.[42] The nature of these interactions is largely determined by the control or influence the repository has over the creation of the resources. Some repositories require content creators to conform to specific restrictions on the type or format of materials, while others exert little or no influence over the creation of materials. In most cases, some activity to prepare the materials for submission (e.g., creation of metadata and documentation) will be required. In a few cases, "aggressive rescue" or salvage measures will need to be taken. Negotiations would cover:

- **Legal issues,** involving all negotiations concerning copyright and other rights (e.g., privacy, donor restrictions) and appropriate clearance for long-term maintenance and continuing access, as well as short-term or immediate access in some cases. Separate negotiations may be necessary for current access and for long-term preservation. Long-term preservation may be jeopardized if submission of materials is based solely on negotiations for current access, which can be more restricted to protect the commercial interests of the content provider.
- **Preservation metadata** based on agreed-on specifications. A repository has to ensure that agreements are in place with content providers about the bibliographic and technical metadata that accompany the submitted materials.[43]
- **Authenticity checks,** confirming that the digital materials submitted to the repository are exactly what the content provider intended.
- **Record keeping,** with adequate documentation for the transactions between the repository and content provider.

This responsibility relies on:

- Well-documented and agreed-on policies about what is selected for deposit, including, where appropriate, specific required formats.
- Effective procedures and workflows for obtaining copyright clearance for both short-term and immediate access, as necessary, and preservation.
- A comprehensive metadata specification and agreed-on standards for its implementation. This is critical for federated or networked repositories and includes standards for the provision of rights metadata from content providers and for representing technical metadata.[44]
- Procedures and systems for ensuring the authenticity of submitted materials.
- Initial assessment of the completeness of the submission.

[42] In the OAIS model, submission of materials into the repository is referred to as the Ingest function; see Appendix A.
[43] In the OAIS model, bibliographic metadata is referred to as Preservation Description Information (PDI) and technical metadata is referred to as Representation Information (RI); see Appendix A.
[44] Work already done in this area will provide a useful starting point. ONIX International, an initiative involving key players in the publishing industry, is developing a standard for the exchange of metadata for publishing, including rights metadata; see www.editeur.org/onix.html.

- Effective record keeping of all transactions, including ongoing relationships, with content providers.

Obtaining Sufficient Control of the Information

This responsibility refers to activities following on the submission of material and involves preparation of the object for storage in the repository,[45] including:

- **Analysis of the digital content:** At this point, the repository must, in consultation with the depositor/rights owner and systems managers, assess the digital object and determine which of its properties are significant for preservation. Rights clearances that have been obtained for copyrighted materials influence this analysis, as do collections management policies and documented collection strengths. This is the moment to apply policies about what formats are acceptable; any necessary migrations should have taken place.[46] Once decisions have been made about what will be preserved, this assessment should be automated as much as possible and the content analyzed systematically as material is deposited into the archive.

  Assessing an Object's Significant Properties

  Recent work has shown that before digital materials are placed into a repository, the owning institution needs to decide what level of preservation is appropriate for each digital object or class of objects. Decisions about an object's "significant properties" influence decisions about the level and method of access as well as the level of preservation metadata required for long-term maintenance. The Cedars Project and the National Library of Australia have explored the concept of significant properties, or an object's "preservable essence."[47] Particularly for repositories with a wide variety of digital materials, how decisions are made about significant properties will be important.

  For traditional materials, access and preservation are often one and the same and are generally done by the same organization. A set of papers will, in all likelihood, continue to be readable and therefore useable. However, for digital materials, simply maintaining a bytestream does not necessarily ensure the digital material will be preserved at an acceptable level to both the repository and the users. For digital materials, the level of "access" in the technical sense depends on judgments made by the collection manager and it should reflect the repository's collections management policies.

  A digital object's significant properties are not absolute, nor are they static. How a repository determines which properties are significant depends on the nature of the organization, the services it provides, and its role in preservation. The collection manager judges the appropriate levels of preservation and access to fulfill the repository's responsibilities and meet the needs of its stakeholders, including content creators and the designated community. Materials deemed to be part of the collection's core might retain all of their original functionality, while more peripheral materials might not include the full complement of these functions or properties. A library may have a choice of formats and associated functionality for a digital object (e.g., PDF, HTML, XML, or SGML for an electronic journal) and the authority to choose what to preserve. Juxtaposed to this, to ensure the authenticity of an electronic record in a legal context, an archive may face more restricted choices; for example, it may not be allowed to

[45] In the OAIS model, submission of materials is referred to as acceptance of the Submission Information Package (SIP); preparation of the objects for storage is referred to as creation of the Archival Information Package (AIP); see Appendix A.
[46] A repository may require the depositor to migrate material to an acceptable format for submission, or for accuracy and reliability the repository may choose to perform migrations itself.
[47] The Cedars Project Report, April 1998–March 2001, www.leeds.ac.uk/cedars/OurPublications/cedarsrepmar01exec.html.

make any functional changes to the electronic record. Since collections management policies may change over time, however, a repository may need to reevaluate and reassess their digital objects' significant properties.

The significant properties of a digital object (i.e., the acceptable level of functionality) dictate the underlying technical form that needs to be documented and supported to ensure preservation of those properties and the amount of metadata, including detailed technical metadata, that must be stored alongside the bytestream to ensure the object is accessible to the agreed-on level. Naturally, decisions about the significant properties have important resource implications: the more significant properties deemed to be necessary, the more associated metadata will be required. The creation and maintenance of the detailed metadata associated with the object's significant properties are critical to the repository's preservation function—the detailed descriptions and the technical information necessary for rendering the bytestream into a meaningful digital object ensure *long-term* preservation.[48] How continuing access is provided over time can and should be kept separate, conceptually, from this basic preservation function.

Significant Properties—a simple example: A repository decides that the only significant property of an electronic journal published on the Web is the text within the journal, not its layout and formatting. There is no need to store information about the HTML environment, but only to include information about retrieving or rendering an ASCII text file.

Significant Properties—a more complex example: An electronic journal that is published on the Web in HTML format includes a database that provides access to the original research data. Although end users currently access the journal in HTML, these pages are created on the fly from SGML. For archiving, the repository takes the SGML files and decides that the significant properties include the hypertext links (internal) as well as any multimedia functions (e.g., sound and video clips) and the functionality of the database, so the object is to be preserved at full functionality. Therefore, the required technical metadata includes robust technical descriptions of the objects, including the SGML DTD and other information about the systems and the software necessary to run the video and sound clips, as well as information about the database (which may or may not use standard SQL), and, finally, the arguably less complex technical metadata about retrieving the text and images.

The significant properties of digital materials need to be determined by policy makers, which a large repository cannot possibly manage object by object. Policies will need to apply to different classes of object and will need to be systematized and automated. Further, the skills necessary for the creation of detailed technical metadata to support significant properties may be beyond the human resources of most research libraries and archives. For this reason, work with software suppliers and systems designers could play a key role: it would benefit many libraries and archives to develop digital repository management systems that provide for the automatic generation of technical metadata for materials submitted to the repository.[49]

[48] In the OAIS model, the detailed pathways that render a preserved bytestream into a meaningful digital object are called Representation Networks; see Appendix A. Representation Information provides the technical means to render the digital objects and also provides detailed human-readable descriptions of the technical environment as necessary. It is the key to the preservation function with OAIS. The Cedars Project implemented the OAIS model so that Representation Information can take the form of a reference to existing information in the network for other objects already in the archive using the same technology. The Representation Information is therefore not duplicated and the archive saves resources.

[49] The Cedars Project has done some preliminary work on providing a network of Representation Information as part of the archive's function. During the Web-based ingest process, content providers simply choose the type of digital file they are submitting from a drop-down menu. If the type of file is not included in the drop-down menu, a request is forwarded to the

Likewise, repository management systems could then be designed to incorporate standard technical metadata as it is submitted alongside the digital object by the content provider, supporting important collaboration with content providers and creators.[50] In this sense, digital repositories can take advantage of the fact that technology is still evolving to influence developments to better accommodate long-term preservation.

- **Continuing Access Arrangements:** A repository needs to choose a strategy for continuing access, which will need to be reevaluated regularly as technology changes. For example, if an object relies on a complex technical environment or uses proprietary technology, an emulation of that environment might be desirable either now or in the future, which affects the level of technical metadata required. Indeed, it may be that the repository stores an emulator, in which case some standards for the development of archival-quality emulators will be necessary.[51]

- **Verification of Metadata:** Although metadata accompanies the object when it is submitted to the repository, it must be verified and, as necessary, enhanced to support the object's long-term maintenance as well as continuing access.

- **Unique and Persistent Identification of Materials:** Much work has been done on the need for unique and *persistent* naming. Nowhere is this more relevant than in long-term preservation of digital materials. A repository needs to ensure that an accepted, standard naming convention is in place that identifies its materials uniquely and persistently for use both in and outside the repository. In a distributed model, it is particularly critical that the participating organizations agree to standard naming conventions.

- **Creation of the Archival Information Package:**[52] Digital repositories can store a digital object and its associated metadata in two ways: as a single bytestream or separately. For practical reasons, repositories may prefer to store the digital object within the repository and provide only pointers or references to the associated metadata in different systems within the organization, such as bibliographic data stored in the library management system. Although such fragmentation of digital object and metadata may present problems in the future, many organizations are choosing "virtual encapsulation" to avoid duplicating metadata. However, some experts feel that, despite the increased costs of duplicating metadata, long-term preservation may be best served by storing the digital content and all of its relevant metadata as a single file within the repository.

- **Authentication and Integrity Checking:** The repository needs to ensure that mechanisms are in place for verifying the digital object, including all associated metadata. The repository should verify that the digital object can be rendered (or at least traced) from the encapsulation back into its original form as it was submitted to the repository. This should include verifying not only the integrity of the bytestream but also confirming the object's usability and functionality.

- **Archival Storage:** Whether archival storage is centralized or distributed, it relies on a robust and well-documented policy for storage and maintenance and for the expected level of service. For

---

archive administrator and the necessary Representation Information is created. In addition, there may be some scope for applying the work done on "canonicalization" of digital materials to assist in the automatic creation of Representation Information; see Clifford Lynch, "Canonicalization: A Fundamental Tool to Facilitate Preservation and Management of Digital Information," *D-Lib Magazine* 5, no. 9 (September 1999) www.dlib.org/dlib/september99/09lynch.html.

[50] NISO (National Information Standards Association, US) is working on this specifically for digital still images. Such work might prove applicable to a wider range of digital materials.

[51] The CAMiLEON project is investigating the use of emulators for continuing access to digital materials. Although as yet inconclusive, the work will eventually result in recommendations for appropriate technology for preservation-quality emulators; see the CAMiLEON Web site: www.si.umich.edu/CAMILEON/.

[52] See Appendix A.

archival storage by a third party, service-level agreements are essential. The policy must include systems for routine integrity checking of the bytestream, once it has been established within the storage facility, and for disaster preparedness, response, and recovery.

This responsibility relies on:

- Detailed analysis of an object or class of objects to assess its significant properties. Analysis should be automated as much as possible and informed by the collections management policy, rights clearances, the designated community's knowledge base, and policy restrictions on specific file formats.
- Verification and creation of bibliographic and technical metadata and documentation to support the long-term preservation of the digital object according to its significant properties and underlying technology or abstract form, with monitoring and updating of metadata as necessary to reflect changes in technology or access arrangements. This involves understanding how strategies for continuing access, such as migration and emulation, influence the creation of preservation metadata.
- A robust system of unique identification.
- A reliable method for encapsulating the digital object with its metadata in the archive.
- A reliable archival storage facility, including an ongoing program of media refreshment; a program of monitoring media; geographically distributed backup systems; routine authenticity and integrity checking of the stored object; disaster preparedness; response, and recovery policies and procedures; and security.

Determining the Repository's Designated Community

Preservation takes place for the designated community: whether to preserve an object or class of objects is initially determined by how the repository's designated community values its content. Likewise, the creation of the technical infrastructure to ensure access to the object depends entirely on the community's technical capability or knowledge base. In particular, the knowledge base determines the minimum level of associated technical metadata for long-term access. "Knowledge base" in this context does not necessarily imply any specific technical expertise on the part of the user; it is an assumption about the technical capability users will have available to them either as actual technical knowledge or through access systems.

Traditionally, knowledge of a library's designated community was gleaned through face-to-face interaction. Generally, the user community was assumed to fit within a broadly defined "research" or "academic" community. These assumptions may not have been documented explicitly as part of the library's policy; if they were, it was not to influence the preservation or long-term retention of materials. Digital repositories, however, rely on a thorough understanding of their designated communities and a federation of repositories that distributes responsibilities, identifying and understanding the designated community is critical, especially if the repositories have divided their collecting and archiving responsibilities along specific lines, such as particular data formats or subject areas. Research libraries and archives may find it difficult to identify the designated community because their users typically represent a wide variety of backgrounds and interests. A national archive or library, for example, preserves material for the whole nation—naturally, detailed knowledge of this designated community is impossible. However, with a thorough understanding of the knowledge base of their designated communities, repositories can limit the level of detail required for technical metadata and thereby contain costs. Knowing the designated community well can also help manage demand.

This responsibility relies on:

- Analysis and documentation of the repository's designated community; for federated or cooperating repositories, a shared understanding of the designated community.

Ensuring the Information to be Preserved is Independently Understandable to the Designated Community

"Independently understandable information" is information that the designated community can understand without the assistance of experts.[53] Making digital information independently understandable poses formidable challenges because it is not itself humanly readable at any level—it relies on further digital information to make it meaningful. However, it is possible to stipulate that the technical metadata required for rendering binary data into meaningful digital objects correspond to the lowest common level of technical knowledge or capability in the designated community. For example, in the current technical climate, one might assume that any user could use a Web browser and an HTML file.

Access and dissemination of objects from the repository will also need to reflect changes in both the technology and in the knowledge base of the designated community. It may be necessary to provide different migrated versions of objects as technologies change; whether this is also reflected in changes to the digital object and technical metadata will be determined by the organization's policies. Many repositories may change their continuing access methods without changing the stored object itself. In other words, "on-the-fly" migration may be provided for materials just for access, but the migrated version itself being not stored.

Clearly defining the designated community and its level of technical capability will help limit the resources necessary to support this "lowest-common-denominator" approach. For organizations such as national repositories that have a very loosely defined community with a disparate knowledge base, this could prove very labor intensive.

This responsibility relies on:

- Well-maintained and documented technical metadata that is kept aligned with the knowledge base of the designated community and with changing technologies.
- A "technology watch" to manage the risk as technology evolves and to provide continuing access and updated methods of access as necessary, such as new migrations or emulators.

Following Documented Policies and Procedures

In the past, some organizations may have relied on vague or even unwritten policy for the governance of traditional collections. However, to ensure effective and efficient mechanisms for long-term preservation of and continuing access to its digital contents, a repository requires well-documented and widely adopted policies—and well-documented procedures to ensure their effective implementation.[54] For distributed repositories, this means clearly articulated responsibilities across participating organizations and consortia.

For research repositories, a strategy or policy for preservation of digital materials may necessarily relate more widely to the organization's information strategy as a whole. More particularly, however, a policy for the preservation of digital files needs to sit comfortably within or alongside policies for nondigital content.

---

[53] The term "independently understandable" is defined by OAIS as "a characteristic of information that has sufficient documentation to allow the information to be understood and used by the designated community without having to resort to special resources not widely available, including named individuals." (Committee for Space Data Systems, *Reference Model for an Open Archival Information System (OAIS) Red Book*, February 2001)

[54] Margaret Hedstrom and Sheon Montgomery, *Digital Preservation Needs and Requirements in RLG Member Institutions* (RLG, 1998) www.rlg.org/preserv/digpres.html.

The link between policy and procedure will also be critical. If the policy of a research repository sets different levels of collection for long-term retention, each level will need a corresponding procedure. Over time, linked policies and procedures will help to reduce costs by supporting automation and scaling. Rather than consider each digital object individually at the point of deposit, procedures will automatically apply, based on the policy for a particular part of the collection.

A Simple Example of Policy and Procedure

The collections policy of the University of Appleton applies these levels of collecting to its subject areas:

- Comprehensive/Research
- Study and Teaching
- Minimal

Their policy also applies these levels of responsibility for long-term preservation:

- Archival (kept forever)
- Served (available for the foreseeable future)
- Mirrored (responsibility taken only for the short term)
- Linked (no long-term responsibility assumed)

Each of the levels has a corresponding technical procedure for digital archiving. For "Archival" materials in a "Comprehensive" collection, all significant properties would be preserved and levels of necessary metadata would be assigned automatically (ideally) at the time of submission. At the other extreme, "Linked" materials from a "Minimal" collection might have very little, if any, preservation.[55]

This responsibility relies on:

- Policies for collections development (e.g., selection and retention) that link to technical procedures about how and at what level materials are preserved and how access is provided both short and long term.
- Policies for access control to ensure all parties are protected, including authentication of users and disseminated materials.
- Policies for storage of materials, including service-level agreements with external suppliers.
- Policies that define the repository's designated community and describe its knowledge base.
- A rigorous system for updating policies and procedure in accordance with changes in technology and in the repository's designated community**.**
- Explicit links between these policies and procedures, allowing for easy application across heterogeneous collections.

Making the Preserved Information Available to the Designated Community

Providing access to materials is an integral responsibility of a digital repository, but access must be clearly defined in order for a repository to understand its implications. Immediate access to materials will require different policies, such as license arrangements, and therefore different management than access to materials over time. If materials are only accessible in a particular format to a specific group of users for a

[55] This example is based on the Berkeley Digital Library SunSITE's *Digital Library SunSITE Collection and Preservation Policy* (1996) sunsite.berkeley.edu/Admin/collection.html.

designated period, different mechanisms will need to be in place than might be appropriate later. Access arrangements will change in accordance with changes in licenses, law, and technology and with local resource constraints. Repositories also need to ensure as far as possible that decisions about access when materials are submitted do not limit what might be possible in the future.

- **Resource Discovery:** To ensure access, a repository's users need to find materials. Many libraries and archives provide this through their main catalog. In practice, many objects to be deposited in a repository will arrive with existing—often very rich—bibliographic descriptions such as MARC or Dublin Core, either accompanying the object or available in an existing system.
- **Authenticity:** The authenticity of digital materials is more complex and potentially troublesome than that of traditional library or archival materials.[56] While traditional materials can be physically verified, digital objects have less obvious evidence of authorship, provenance, or even context. For this reason, they give rise to suspicions that will be assuaged only by rigorous mechanisms throughout the repository for ensuring that the digital object is what it purports to be and that it is what was originally deposited into the repository. Authenticity checks are required at all functional levels of the digital repository: At submission, mechanisms must ensure that the object, as received, is what was intended by the content provider.[57] The stored material needs a regular system of integrity checks to ensure the bytestreams are maintained. Physical and system procedures must be maintained and the available mechanisms for access to the original bytestreams must be regularly checked; migrated versions must be verified and available emulators tested. Finally, the information provided to the user—the copy of the bytestream as well as the necessary metadata and rendering software—all requires verification.
- **Legal Issues:** Legal restrictions—licenses and legislation—govern access to materials in a repository, and over time these change. Digital repositories require an infrastructure that can support a multiplicity of access arrangements for different materials and different types of users.
- **Pricing:** Repositories that govern access with a fee structure require mechanisms for managing electronic commerce.
- **User Support:** For digital repositories, access, particularly older to materials, requires some level of user support. To a great extent, this will be determined by how the repository defines its designated community's knowledge base or technical capability.
- **Record Keeping:** As part of the repository's administration function, it may be advisable to keep track of the dissemination of objects out of the repository.

This responsibility relies on:

- A system for discovery of resources.
- Appropriate mechanisms for authentication of the digital materials.
- Access control mechanisms in accordance with licenses and laws, and an "access rights watch."
- Mechanisms for managing electronic commerce.
- User support programs.

[56] Charles T. Cullen et al., *Authenticity in a Digital Environment* (CLIR, May 2000) www.clir.org/pubs/reports/pub92/contents.html.
[57] In the OAIS model, submission is referred to as the Ingest function; see Appendix A.

Advocating Good Practice in the Creation of Digital Resources

If digital repositories advocate standard approaches to the creation of digital materials, they will be able to achieve important economies of scale and reduce costs. It would be critical to involve both content providers and software suppliers and could require an outreach program for potential depositors. It will be important for key players to facilitate this dialogue; organizations such as RLG, OCLC, or IFLA (International Federation of Library Associations and Institutions) are well placed to take the lead at the international level, where this may be more effective.

This responsibility relies on:

- Effective mechanisms for advocating good practice for content providers.

Summary

The effective establishment and maintenance of digital repository services is complex. The pace with which technology advances is at odds with the long-term view research repositories necessarily take toward long-term preservation of materials. For this reason, research repositories will need to act more quickly to preserve digital resources for posterity than they ever had to for nondigital materials. This list of responsibilities represents a follow-on from the recommendations in the CPA/RLG Task Force Report and aims to provide research organizations with a checklist of characteristics, standards and mechanisms necessary for the establishment of reliable archiving services. It is intended to facilitate the establishment of digital repositories for which there is a pressing need in the current and continuing technological environment. There are still areas that require further work and consideration in order to allow full implementation of the proposed model; equally important will be commitment and advocacy from within the research community locally, regionally, nationally, and internationally.

6 Recommendations

Recommendation 1: Guidance is needed on well-documented policies for digital collections management that logically connect selection of materials with long-term maintenance and continuing access for a designated community over time, including work to define designated communities for research repositories.

Recommendation 2: Archivists and librarians should work together to determine how to describe what is in digital repositories--what type of information needs to be made available and how, where, and to whom, whether in a MARC record, in a new kind of holdings record attached to a MARC record, as additional preservation metadata, or in some other way.

Recommendation 3: Assessing the significant properties of a digital object or of a class of digital objects will be a critical component of an effective long-term digital repository. More work is required on the possible significant properties for different classes of digital materials and how these properties in turn determine the underlying technical form or structure to be preserved and the necessary technical metadata.

Recommendation 4: The current collaboration between RLG and OCLC to facilitate development of an internationally accepted standard for digital preservation metadata should be widely supported, and both organizations working together should provide a program of implementation and support.

Recommendation 5: Technical standards are needed for a reliable digital repository, particularly from collaborative efforts such as RLG and OCLC's. The standards are needed most critically for:

- The exchange and use of existing bibliographic descriptions in preservation metadata.
- The provision of rights metadata from content provider to repository.
- The expression of technical metadata and how it can be shared between repositories.

Recommendation 6: Projects or a series of case studies are needed to develop working models for trusted repositories.

Recommendation 7: A certification framework and certification process for digital repositories are crucial and their absence has been an impediment to assigning trust. Model processes, including checklists for certification reviews, should be developed incorporating the community-approved attributes of trusted digital repositories, the work of the ISO Archiving Series, and other relevant projects.

Recommendation 8: Archivists and librarians need more thorough understanding of how cooperative digital repositories can be implemented and managed, including the use of external service suppliers. Models for the establishment of cooperative archiving services will be useful and necessary, as will be examples of service-level agreements as they apply to digital repositories (e.g., service-level agreements for external suppliers of archival storage).

Recommendation 9: Further work is needed to produce models for collaboration across research repositories taking responsibility for long-term preservation. A detailed analysis of collaborative scenarios will be necessary before a clear picture will emerge of all necessary standards and practical guidance.

Recommendation 10: Building on existing work such as Project Prism, the community requires more complete understanding of risk management in the continued evolution of technology.[58]

[58] Funded by Digital Libraries Initiative Phase 2, Project Prism at Cornell University is a four-year effort to investigate and develop policies and mechanisms for information integrity in digital libraries; see www.prism.cornell.edu.

Recommendation 11: The pressing need for unique and persistent systems of identification for digital information has supported a great deal of work in this area. However, it is not yet clear that current approaches, if any, are best suited to the purposes of long-term preservation. A concise synthesis and analysis of current work is needed, along with recommendations on the most applicable approaches.

Recommendation 12: Work is needed on obtaining copyright clearance and models for contracts or agreements between rights owners/producers and archives/libraries.

Recommendation 13: Stakeholders in the research repository community should formally endorse the Open Archival Information Systems Reference Model, actively supporting and advocating it as a standard for digital repository services. Specific guidance and support for implementation will be needed and could take the form of advisory services offered by expert organizations.

Recommendation 14: The OAIS model will be more useful to libraries and archives as many of its functions can be automated. Work will be needed to consider existing systems and provide recommendations and specifications for further development work or enhancements for automation.

Recommendation 15: Libraries and archives need more practical experience using both migration and emulation as strategies for continuing access. This can be achieved through short-term pilot projects and through raised awareness of existing work, particularly outside the library and archives communities. Guidance should include both technical and legal (copyright) implications for migration and emulation. Further work is necessary on how specific strategies for continuing access may affect preservation metadata created when materials are submitted to the collection; for example, what information is required for assuring emulation is a viable future access strategy?

Recommendation 16: A study is needed to test whether creating rich digital masters does support their long-term preservation and usability.

Selected Resources

Projects

CAMILEON (Creative Archiving at Michigan and Leeds: Emulating the Old on the New): www.si.umich.edu/CAMILEON/

Cedars (CURL Exemplars in Digital Archives) Project: www.leeds.ac.uk/cedars

NEDLIB (Networked European Deposit Library): www.kb.nl/coop/nedlib/

PANDORA (Preserving and Accessing Networked Documentary Resources of Australia) Project: www.nla.gov.au/pandora

Preserving Access to Digital Information (PADI): www.nla.gov.au/padi

Publications

Beagrie, Neil and Daniel Greenstein, *A Strategic Policy Framework for Creating and Preservation Digital Collections*, E:Lib Supporting Study P3 (London: Library and Information Technology Centre, 1998) ahds.ac.uk/strategic.htm.

Consultative Committee for Space Data Systems (CCSDS), *Reference Model for an Open Archival Information System (OAIS)* (July 2001) www.ccsds.org/documents/pdf/CCSDS-650.0-R-2.pdf.

Garrett, John and Donald Waters, *Preserving Digital Information: A Report of the Task Force on Archiving of Digital Information*. (Washington, DC: Commission on Preservation and Access, and Mountain View, CA: RLG , 1996) www.rlg.org/ArchTF/index.html.

Gatenby, Pam "Digital Archiving: Developing Policy and Best Practice Guidelines at the National Library of Australia." Paper presented at An Interactive Workshop sponsored by ICSTI and ICSU Press, January 2000: www.icsti.org/icsti/2000workshop/gatenby.html.

Hedstrom, Margaret and Sheon Montgomery, *Digital Preservation Needs and Requirements in RLG Member Institutions* (Mountain View, CA: RLG, 1998) www.rlg.org/preserv/digpres.html.

Jones, Maggie and Neil Beagrie, *Preservation Management of Digital Materials Workbook* (London: Re:source, 2000) www.jisc.ac.uk/dner/preservation/workbook/.

OCLC/RLG Working Group on Preservation Metadata, *Preservation Metadata for Digital Objects: A Review of the State of the Art* (January 2001) www.oclc.org/digitalpreservation/presmeta_wp.pdf.

Papers presented at Preservation 2000: An International Conference on the Preservation and Long Term Accessibility of Digital Materials (2000) www.rlg.org/events/pres-2000/prespapers.html.

Public Record Office, *Management, Appraisal and Preservation of Electronic Records* (Kew: Public Record Office, 1999) www.pro.gov.uk/recordsmanagement/eros/guidelines/.

RLG DigiNews, a bimonthly newsletter: www.rlg.org/preserv/diginews.

Rothenberg, Jeff, *Avoiding Technological Quicksand: Finding a Viable Technical Foundation for Digital Preservation* (Washington, DC: Council on Library and Information Resources, 1999) www.clir.org/pubs/reports/rothenberg/pub77.pdf.

Appendix A: OAIS Technical Overviews

The OAIS Information Model and Digital Preservation Metadata

The foundation of an OAIS digital repository is the **Information Package**, which includes both a digital object and the necessary associated metadata. As objects are submitted to the repository, they arrive as a **Submission Information Package (SIP),** which contains the digital object and any other information the content provider deemed relevant or available. On submission into the repository, the SIP is enhanced as necessary and then encapsulated as an **Archival Information Package (AIP)**, including the data object and all its associated metadata. The AIP is the cornerstone of the digital repository. When a user requests access to an object, a **Dissemination Information Package (DIP)** is provided, which typically contains a copy of the digital object as well as the necessary metadata and support systems to retrieve and use the digital object.

The original digital object is stored as a bytestream in the AIP, along with the metadata necessary for making that bytestream into a meaningful and useable digital resource. For digital materials in the future, what is known about them will come from the information or metadata stored with them—the binary data is meaningless without some description of what it is and how it works. OAIS provides for two main types of metadata: **Content Information** and **Preservation Description Information (PDI)**. The Content Information contains both the digital object (as a bytestream) and all the necessary technical metadata (called **Representation Information** or **RI**) to support its transformation into a meaningful digital object. The PDI is all of the other descriptive information that, although not critical for actual conversion of the bytestream, is deemed necessary for long-term preservation. The creation and maintenance of the preservation metadata (both the technical RI and the PDI) will most likely represent the bulk of the initial costs and complexities of digital preservation.

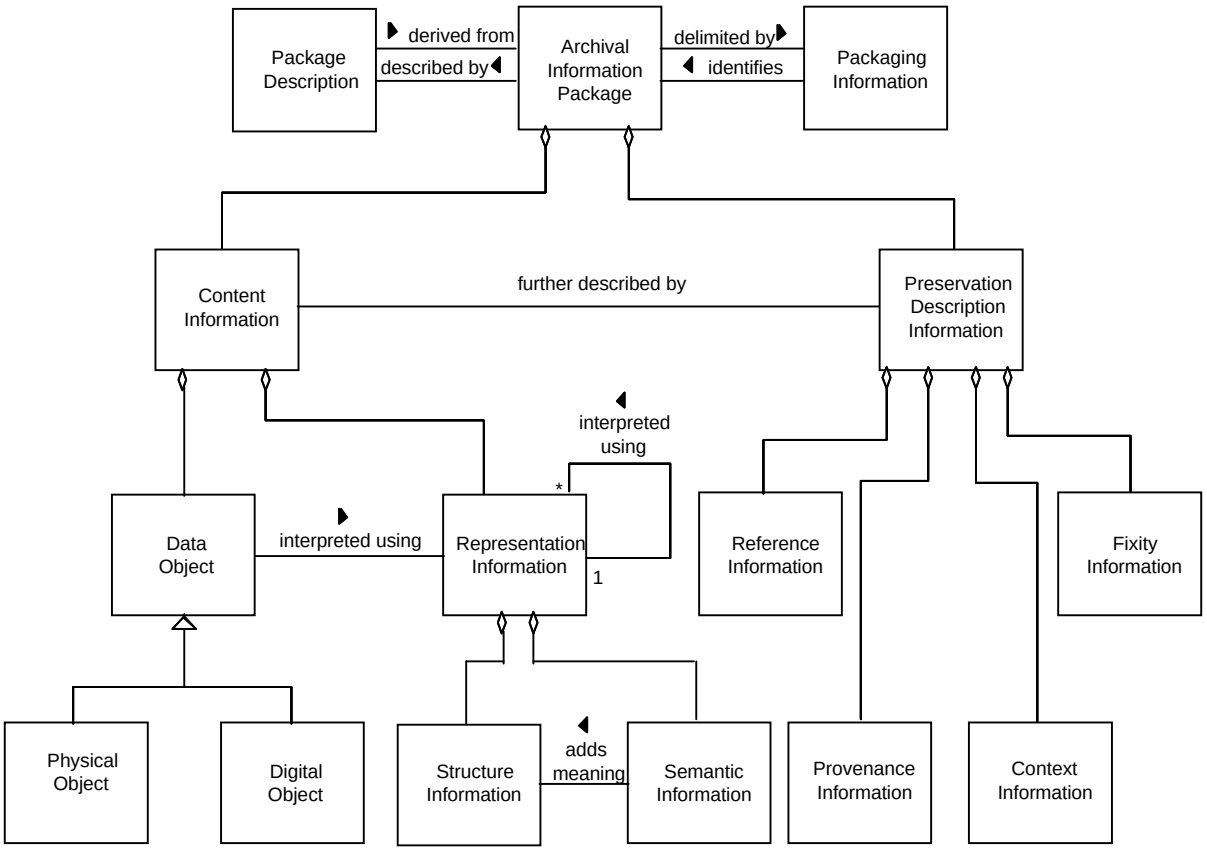

Figure 1. The Archival Information Package (AIP), detailed view[59]

Preservation Description Information (PDI)

Preservation Description Information is the metadata that includes traditional bibliographic information as well as more detailed information to ensure the material is effectively preserved. This metadata includes details of ownership, copyright, and other intellectual property rights, such as licensing arrangements and access restrictions. Preservation Description Information is broken down into Provenance, Reference, Fixity, and Context Information.

Representation Information (RI)

Representation Information is the technical metadata for mapping the bytestream into specific data types and formats. Without adequate Representation Information, the bytestream is not retrievable as a meaningful digital object: the Representation Information provides meaning to the bits. In practice, Representation Information involves many different descriptions for a variety of relevant technologies. For example, even a simple Web page that contains graphics requires descriptions of the Web environment (browser, etc.), the text (ASCII standard), and the image files. This is a recursive system—all the Representation Information requires additional Representation Information in order to be understood. Representation Information is therefore likely to involve references to other Representation Information elsewhere in the repository and will take the form of a **Representation Network**. In theory, if the digital object is to remain accessible for the long term, this recursion will stop only when a physical form is

[59] Consultative Committee on Space Data Systems, *Reference Model for an Open Archival Information System (OAIS)*: Draft Recommendation for Space Data System Standards, CCSDS 650.0-R-1, Red Book, May 1999, www.ccsds.org/RP9905/RP9905.html.

encountered, such as a system specification or technical manual. However, in practice, the "end node" of the Representation Network will be the software or hardware that is part of the knowledge base of the designated community, which allows end users to understand the digital object without recourse to the repository or the data creator. It is the level of technology or technical capability that can be assumed to be supported outside the repository itself—a kind of lowest common denominator. These end nodes of the Representation Information will need to be closely monitored. As any technology threatens to become obsolete, Representation Information will need to be generated and stored in the repository to support that technology.

International Collaboration and Digital Preservation Metadata

A number of initiatives have developed preliminary specifications for preservation metadata (both RI and PDI). Although beyond the scope of this report, the work of Harvard University, the National Library of Australia, the NEDLIB and Cedars projects, and others have played a key role.[60] RLG and OCLC are now taking forward the work done by these and other organizations to build consensus and develop a standard framework for metadata for the long-term retention of digital materials.[61]

[60] Harvard University Library, *Digital Repository Services (DRS)User Manual for Data Loading,* available from hul.harvard.edu/ois/systems/drs/doc.html; Harvard University Library, Library Digital Initiative, *Image Reformatting,* hul.harvard.edu/ldi/html/reformatting_image.html; Harvard University Library, *Library Preservation at Harvard: Image Digitization*, preserve.harvard.edu/resources/digitization/image.html; National Library of Australia, *Preservation Metadata for Digital Collections: Exposure Draft* (1999) www.nla.gov.au/preserve/pmeta.html; Catherine Lupovici and Julien Masanes, *Metadata for the Long Term Preservation of Electronic Publications*, NEDLIB Report Series, report 2 (NEDLIB Consortium, 2000), available from www.kb.nl/nedlib. Cedars Project Team, *Metadata for Digital Preservation: The Cedars Outline Specification* (June 2000) www.leeds.ac.uk/cedars/metadata.html.

[61] The OCLC/RLG Working Group on Preservation Metadata, *Preservation Metadata for Digital Objects: A Review of the State of the Art. A White Paper.* DRAFT (January 2001) www.oclc.org/digitalpreservation/presmeta_wp.pdf.

The OAIS Functional Model

Along with its model for necessary metadata, OAIS includes a comprehensive logical model for the functions of a repository. Any scheme for describing the attributes of a reliable digital repository will require that a variety of communities share an understanding of the repository's basic functions. The OAIS functions include:

- Submission or "pre-Ingest" activities
- Ingest
- Archival Storage
- Data Management
- Preservation Planning
- Archive Administration
- Access / Dissemination[62]

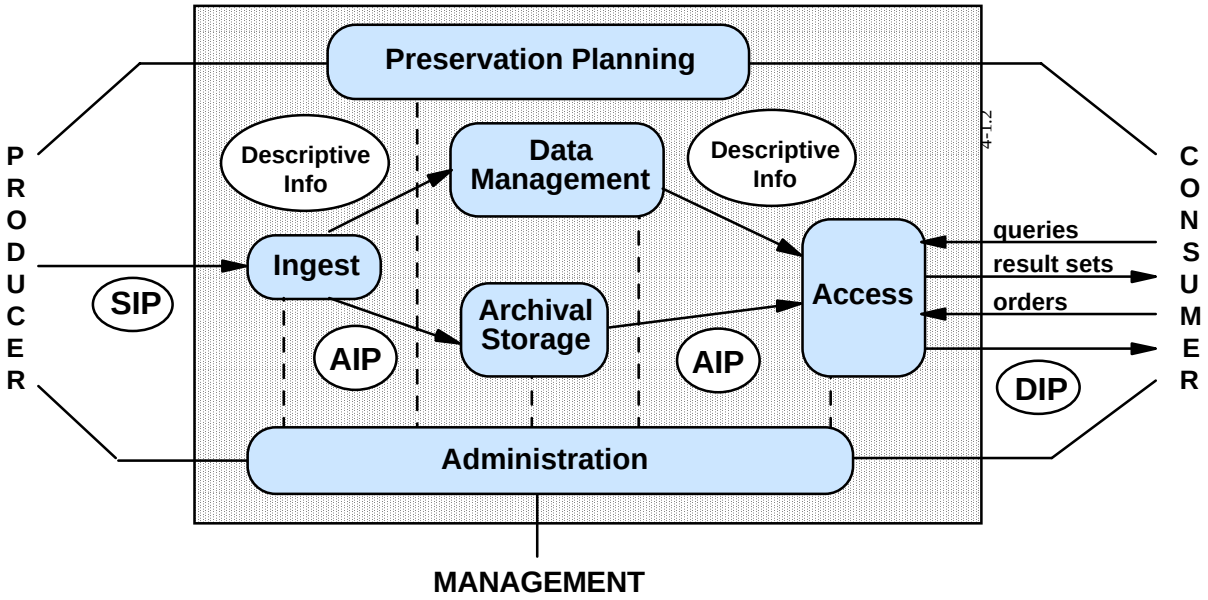

Figure 2. Overview of the Open Archival Information Systems Reference Model

The following descriptions of these functions are based both on the reference model itself and on implementations of the model in a library or archive environment, mainly by the Cedars and NEDLIB projects[63].

[62] The OAIS model contains a complete glossary of its terms.
[63] On Cedars, see The Cedars Project Report, April 1998–March 2001, www.leeds.ac.uk/cedars/OurPublications/cedarsrepmar01exec.html; Johan Steenbakkers, *The NEDLIB Guidelines: Setting up a Deposit System for Electronic Publications,* NEDLIB Report Series, report 5 (NEDLIB Consortium, 2000); available from www.kb.nl/nedlib.

Submission and "Pre-Ingest" Activities

Before a repository can accept responsibility as a reliable archiving service, management tools must be in place, covered by a well-documented and agreed-on collections policy document. For most libraries and archives, this will be in the form of a collections management/development policy. Digital materials, in this sense, are not always different from traditional materials and many of the same criteria apply to both. However, for digital materials these elements are the most critical:

- Evaluation criteria for assessing potential submissions; that is, selection criteria for digital preservation.
- Collections development strategies and technical strategies for continuing access.
- Collections development procedures, including review procedures pertaining to retaining and deaccessioning materials.

Where appropriate, the repository needs to ensure availability of copyright and other intellectual property rights or privacy/confidentiality information, including licenses, schedules for deposit (where regular updates will be forthcoming), and appropriate documentation, and may even include details of preferred formats and media.

As part of the "pre-Ingest" activities, the repository also needs to:

- Check any existing deposit schedules to ensure everything expected has been received.
- Assign the digital object's unique identifier(s), if not already available, and provide labels for the physical artifact.
- Check for viruses and validate the integrity of the digital object and its physical carrier.
- Assess in detail the significant properties of the digital object, such as its look and feel, or functionality.
- Validate or improve the documentation.
- Where appropriate, reformat the digital object according to repository policies.
- Ensure that all necessary metadata for long-term maintenance and continuing access accompanies the object.

Ingest

Once the digital resource has been properly prepared, the Ingest function allows for materials submitted as a Submission Information Package (SIP) to be prepared as Archival Information Packages (AIPs) for storage. At this stage, the Content Information and related PDI are established in the repository. Creation of the AIP is the foundation of the long-term preservation function.

Ingest to the repository on a practical level involves:

- Assignment and/or validation of unique identifier. This identifier must be unique not only within the repository but also within the repository's wider community or the federation of which it is a part.
- Selection and validation of the agreed-on underlying technology or underlying abstract form based on the object's significant properties.[64]

[64] For a discussion of an object's significant properties, see Section 5.

- Transformation of the object as it was submitted, along with its associated metadata, into a bytestream that can be stored on suitable hardware in the repository.
- Establishment of necessary Representation Information.
- Verification of all Preservation Description Information.

In a distributed environment, the Ingest function can be performed across participating repositories. In practice, this requires the adoption of strict policies for the assignment of unique identifiers or use of existing identifiers and a well-documented and fully implemented specification for preservation metadata—both PDI and RI. The outcome of ongoing work with unique identifiers and emerging standards in this area will be important.

Archival Storage

The Archival Storage function in an OAIS repository is the logical component that contains the necessary services for the effective storage and retrieval of AIPs. Such functions include:

- Moving AIPs from Ingest into permanent storage.
- Managing the storage hierarchy.
- Refreshing the storage media.
- Providing all necessary information to allow objects to be disseminated from the repository.

Like the Ingest function, Archival Storage can be centralized or distributed. Whichever model is adopted, it will rely on a robust and well-documented policy that describes how the material is stored and maintained and what level of service is expected. Archival Storage can be done either in-house or by a third party; service-level agreements are essential with third parties.

Archival Storage must also include systems for routine integrity checking. The stored bytestream should be checked regularly to assess whether it is truly identical to the original bytestream. Authentication and integrity checking should be a regular activity of the repository's administration.

Another critical component of any storage facility is a robust system for disaster recovery, with these main components:

- An ongoing program of media refreshment, transferring bytestreams onto newer, fresher media. This includes a program of monitoring repository media for possible degradation and subsequent integrity checking on refreshed bytestreams.
- Geographically distributed backup systems—ideally, more than one.

Data Management

Data Management covers all aspects of an OAIS repository and is essential for both long-term preservation and day-to-day administration and use. It represents good record keeping at every stage described by the OAIS functions. The activities in the Data Management component are determined by the policies developed and maintained by the repository's Management and Administration. Some of the components that the OAIS model includes in Data Management are:

- Pricing information (if applicable) and access controls.
- Customer profiles.
- Tracking of user requests.

- Security information, including any usernames, passwords, digital certificates, etc.—anything used to authenticate users of the repository.
- Statistical information to improve operation.
- Accounting information.

The Cedars Project added these elements to Data Management:

- Records from Pre-Ingest negotiations, such as immediate or short-term access arrangements.
- Policies for and monitoring of the allocation of unique identifiers.
- Maintenance of records of holdings for use with finding aids.[65]

Preservation Planning

Preservation Planning is the function responsible for monitoring the OAIS environment and providing recommendations to the repository (through the Archive Administration function) to ensure that materials are accessible to the designated community over the long term. This function detects critical information about shifts in the knowledge base of the designated community, allowing the repository to accommodate shifts in technology. Preservation Planning includes:

- Monitoring the designated community.
- Monitoring technology.
- Monitoring the significant properties of the repository's contents, as necessary.
- Developing preservation strategies and standards for continuing access.
- Developing packaging designs and migration or routine transfer plans.

The addition of this function to OAIS in March 2001—thanks particularly to the efforts of the NEDLIB Project and the National Library of Australia—reflects the influence of libraries and archives on the reference model. Implementation of the model by libraries and archives has made clear the need for a function that explicitly allows for the application of processes such as migration or emulation for continuing access.

Archive Administration

Archive Administration is the OAIS function that contains all services needed for day-to-day maintenance of the repository. This includes functions such as management of the systems configurations (statistics, etc.) and the more strategic issues of repository policy development and maintenance. Archive Administration includes:

- Negotiating submissions agreements with content producers and providers.
- Reviewing procedures.
- Maintaining systems configurations for hardware and software.
- Developing and maintaining repository policies and standards, including policies and standards recommended and enhanced by Preservation Planning.
- Providing user support.

[65] The Cedars Project Report, April 1998–March 2001, www.leeds.ac.uk/cedars/OurPublications/cedarsrepmar01exec.html.

- Interacting with management outside of the repository.

To this list, the Cedars Project added:

- Reviewing and maintaining Representation Information and Networks, based on recommendations from Preservation Planning.
- Negotiating user access agreements with service providers or others.
- Communicating with other repositories.[66]

Archive Administration may also include monitoring of legal status and relevant changes in national and possibly international law.

Access/Dissemination

In the OAIS model, access to archived materials is provided through the Dissemination Information Package (DIP): a copy of the digital object along with the necessary metadata, and software as necessary. In addition to preparation of the DIP, Dissemination requires mechanisms for both verifying the integrity of the information in the DIP and for ensuring that users have permission for access to the material.

The DIP differs from the AIP in that it contains only a copy of the digital object (or a new object generated from the AIP such as in a migration) and only the metadata necessary for access to the appropriate level. It probably does not include the full complement of rich, descriptive PDI or RI. The important separation of AIP and DIP allows for materials to be disseminated in different ways and by different methods of continuing access. For example, a database might be available both as a migrated version accessible in the current version of Microsoft Access or with an emulator of its original database package. This separation nicely illustrates the difference between long-term maintenance and continuing access.

Although it would be possible to simply copy the AIP and distribute it to a user, this is probably not practical. Most users would have no need for all the detailed metadata that supports the bytestream long term. The DIP should contain only the metadata necessary for the required level of access. For example, a PDF file may be disseminated as a bytestream with a copy of Adobe Acrobat Reader™, while within the repository the technical metadata includes much richer information about the PDF format.

Technically, the level of access provided for digital materials depends on the judgment of the archivist and/or the collection manager, based on the objects' significant properties. Strategies such as emulation or migration of some kind are the obvious choices for providing access over time. For many materials, access will change over time and will vary according to the designated community. For example, changes in licensing arrangements or in copyright may require updates to the way access is provided, but not necessarily to the way the bytestream is preserved. Likewise, particular designated communities may have different levels of access to particular materials at different times (e.g., researchers vs. undergraduate students or rights holders vs. users).

Digital materials cannot be considered preserved without meaningful access. However, publishers and other rights holders are often cautious about the preservation of their materials if they think unlimited or unrestricted public access is a necessary precondition. The distinction between access for current users and long-term maintenance and access needs to be explored more deeply and better understood to ensure interests in the one do not jeopardize the other.

[66] Ibid.

Summary

A review of the Open Archival Information Systems model provides insight into the functions necessary for the establishment of a long-term digital repository. Its adoption by libraries and repositories is increasing, due mainly to continued involvement of the library and archives communities in the development of the model itself. Although work is still to be done on many of the standards necessary for effective distributed implementation of the model, OAIS provides a critical framework for establishing or enhancing digital archiving services.

Appendix B: Definitions of Terms

Archival Information Package (AIP): An Information Package, consisting of the Content Information and the associated Preservation Description Information, which is preserved within an OAIS.

Archival Storage: The OAIS entity that contains the services and functions for the storage and retrieval of Archival Information Packages.

Content Information: The set of information that is the primary target for preservation. It is composed of a Data Object and its Representation Information. For example, Content Information could be a single table of numbers representing, and understandable as, temperatures, but excluding the documentation that would explain its history and origin, how it relates to other observations, etc.

Context Information: The information that documents the relationships of the Content Information to its environment, including why the Content Information was created and how it relates to other Content Information.

Data migration: One current strategy for providing continuing access to archived digital materials over time. It is perceived to be the most reliable strategy for providing continuing access to many types of digital materials because it has been used for years for routine migration of homogeneous digital materials. However, the library community lacks documented practical experience that shows it is a reliable approach for heterogeneous digital collections such as multimedia CD-ROMs or electronic journals, so it has yet to see widespread adoption for these materials. For the purposes of this report, the definition of "data migration" is based on that provided in the CPA/RLG report: "a set of organized tasks designed to achieve the periodic transfer of digital materials from one hardware/software configuration to another, or from one generation of computer technology to a subsequent generation."[67]

Although in some cases data migration is straightforward, it can be a complex process and may or may not be reversible, which will have important implications for its effectiveness as a continuing access strategy. For example, in the early 1990's many organizations with files stored in EBCDIC code opted for immediate migration to ASCII despite the fact that the ASCII character set didn't allow for all character information from EBCDIC files to be migrated without loss of information. This was not a reversible migration and therefore some information was lost. When UNICODE was subsequently introduced its character set did allow for all EBCDIC characters to be represented – therefore a more effective migration was possible from original EBCDIC files directly to UNICODE. For organizations where the original EBCDIC files were discarded, there was irretrievable loss of information. This illustrates nicely why reversible migrations are preferred for the purposes of preservation.[68]

Currently, there is a lack of practical experience to provide guidance on migrating complex digital materials through different data formats. The CAMiLEON project and Cornell University have some initial results comparing different approaches to data migration.[69]

---

[67] John Garrett and Donald Waters, *Preserving Digital Information: Report of the Task Force on Archiving of Digital Information* (Commission on Preservation and Access and RLG, 1996) www.rlg.org/ArchTF/index.html.

[68] Example adapted from David Holdsworth and Derek M. Sergeant, *A Blueprint for Representation Information in the OAIS Model* (2000) gps0.leeds.ac.uk/~ecldh/cedars/nasa2000/nasa2000.html or esdis-it.gsfc.nasa.gov/MSST/conf2000/index.html.

[69] CAMiLEON, part of the JISC/NSF International Digital Libraries Program (DLI2), is a partnership between the University of Michigan and the University of Leeds to explore use of emulation as a strategy for digital preservation, see www.si.umich.edu/CAMILEON/; Gregory W Lawrence et al., *Risk Management of Digital Information: A File Format Investigation* (CLIR, June 2000) www.clir.org/pubs/reports/pub93/contents.html; Paul Wheatley, *Migration – a CAMiLEON discussion paper*, www.leeds.ac.uk/camileon.

Data Object: A digital object (can also be a physical object).

Designated community: An identified group of potential users of the archive's contents who should be able to understand a particular set of information. The designated community may be composed of multiple user communities.

Digital preservation strategy: A digital preservation strategy is a particular technical approach for providing continued access to archived digital materials. At this time, three main strategies are used to keep materials within the repository "fresh" and ensure that they are accessible using current technology: data migration, persistent object transformation, and technology emulation.

Dissemination Information Package (DIP): The Information Package, derived from one or more AIPs, received by the user in response to a request to the repository.

Fixity Information: The information that documents the authentication mechanisms and provides authentication keys to ensure that the Content Information object has not been altered in an undocumented manner. For example, a CRC code for a file or a checksum.

Independently understandable: A characteristic of information that has sufficient documentation to allow the information to be understood by the designated community without having to resort to special resources not widely available, including named individuals.

Information Package: Content Information and associated Preservation Description Information that is needed to aid in the preservation of the Content Information. The Information Package has associated Packaging Information used to delimit and identify the Content Information and Preservation Description Information.

Ingest: The OAIS entity that contains the services and functions that accept Submission Information Packages from producers, prepare Archival Information Packages for storage, and ensure that Archival Information Packages and their supporting Descriptive Information become established within the OAIS-compliant repository.

Knowledge base: A set of information, incorporated by a person or system, that allows that person or system to understand received information.

Long term: A period long enough to raise concern about the affect of changing technologies, including support for new media and data formats, and of a changing user community.

Long-term preservation: The act of maintaining correct and independently understandable information over the long term.

Metadata for digital preservation: The effective management and use of digital resources in a repository will rely on a robust system of resource description—for resource discovery, access, and preservation. Metadata research continues to generate interest worldwide; to date, most activity has focused on metadata for resource discovery (e.g., MARC, Dublin Core, CIDOC). However, there is increasing awareness that reliable digital repositories will depend on the creation and storage of information required to support a chosen preservation strategy, such as migration, emulation, or technology preservation. This information will need to describe the data in detail including, broadly speaking, both descriptive and structural metadata. Although these have been defined in different ways, within the context of the OAIS model, preservation metadata takes two forms:

- Preservation Descriptive Information, which includes general resource description as well as rights management information and descriptions of actions taken for the purposes of preservation.

- Representation Information, which maps the stored data into more meaningful concepts; that is, systems information that renders bits and bytes into a meaningful digital object. For example, the ASCII definition, which maps data (bits) into readable symbols.

Open Archival Information System (OAIS) Reference Model: Developed by the Consultative Committee on Space Data, a conceptual framework and reference tool for defining a digital repository. It provides a model of the environment, functions, and data types for implementing a digital repository. The OAIS is undergoing approval as an ISO standard and its publication as an international standard is expected later this year.

Persistent object transformation: Although not widely used in the library and archive community, a strategy for providing continuing access that has been widely publicized internationally. Persistent object transformation may appear to be migration by another name; in fact, it takes a longer term approach, focusing not on the object's technical environment or on moving it into current technology, but attempting to define the essential attributes and methods of the object. These attributes and methods are then made explicit within the objects themselves through tagging and/or encapsulation that are independent of the object's current or original technical infrastructure.[70]

Preservation Description Information (PDI): The information that is necessary for adequate preservation of the Content Information; it can be categorized as Provenance, Reference, Fixity, and Context Information.

Producers: The people or systems that provide information to the archive.

Provenance Information: The information that documents the history of the Content Information. This information tells the origin or source of the Content Information, any changes that may have taken place since it was originated, and who has had custody of it since it was originated. Examples of Provenance Information are the principal investigator who recorded the data, and the information concerning its storage, handling, and migration.

Reference Information: The information that identifies and, if necessary, describes one of more mechanisms used to provide assigned identifiers for the Content Information. It also provides identifiers that allow outside systems to refer unambiguously to a particular Content Information; for example, an ISBN.

Rendering software: Software that displays Representation Information of an Information Object in forms understandable to humans.

Repository: An organization that intends to maintain information for access and use.

Representation Information: The information that maps a data object into more meaningful concepts. For example, the ASCII definition that describes how a sequence of bits (i.e., data object) is mapped into a symbol.

Representation Network: The full set of Representation Information that describes the meaning of a digital object. This can apply to a single data object or to an entire archive.

Significant properties: The technical and other characteristics of the digital object that the depositor and/or repository agree to be most important for preservation over time. For digital materials, simply maintaining a bytestream does not ensure the digital object will be preserved at a level acceptable to the repository and its users. A digital object's significant properties are not assumed to be absolute; repositories

[70] Reagan Moore et al., "Collection-Based Persistent Digital Archives – Part 1," *D-Lib Magazine* 6, no. 3 (March 2000) www.dlib.org/dlib/march00/moore/03moore-pt1.html, and Reagan Moore et al., "Collection-Based Persistent Digital Archives – Part 2," *D-Lib Magazine* 6, no. 4 (April 2000) www.dlib.org/dlib/april00/moore/04moore-pt2.html.

will make judgments that fulfill their preservation responsibilities and meet the needs of their user communities and the wishes of the depositor. Significant properties may apply to a single digital object or to an entire class of digital objects within a repository. This term was first coined, defined, and employed by the Cedars Project.

Submission Agreement: An agreement reached between an OAIS-compliant repository and the producer that specifies a data model for a data submission. This data model identifies format/contents and the logical constructs used by the producer and how they are represented on each media delivery or in a telecommunication session.

Submission Information Package (SIP): An Information Package that is delivered by the producer to the repository for use in the construction of one of more AIPs.

Systems manager: Broadly, the technical specialist involved in the management and preservation of digital collections.

Technology emulation: A strategy for continuing access to digital materials that mimics or re-creates the digital object's original technical environment using current technology. Access to the object relies on a copy of the original bytestream (as deposited) and an emulation of its original operating environment. Emulation can take place at either the hardware or software level. Emulation may be particularly useful for preserving the "look and feel" of the object; however, like migration, there is little if any practical experience in applying emulation in a production environment. Preliminary work by IBM and others suggests that emulation may be the most practical (and in some cases the only) strategy for some more complex or esoteric digital materials, particularly for preserving the original functionality.[71] Recent findings of the CAMiLEON project challenge the perceived wisdom that migration and emulation are two completely different approaches: preliminary work with obsolete computer systems suggests that, as we build a more complete understanding of the various methods for migrating data, it will become clear that emulation and migration represent related approaches on a graduated scale.[72] It may be that skepticism about emulation will be revealed as a lack of understanding of the complexities of data migration.

[71] Raymond A Lorie, *The Long Term Preservation of Digital Information* (IBM Almaden Research Center, October 2000) www.almaden.ibm.com/u/gladney/Lorie.pdf. For more on different approaches to using emulation for digital preservation, see David Holdsworth and Paul Wheatley, *Emulation, Preservation and Abstraction,* www.leeds.ac.uk/camileon and Jeff Rothenberg, *Avoiding Technological Quicksand: Finding a Viable Technical Foundation for Digital Preservation* (CLIR, January 1999) www.clir.org/pubs/reports/rothenberg/contents.html.

[72] Paul Wheatley, *Migration – a CAMiLEON discussion paper,* www.leeds.ac.uk/camileon.

Appendix C: Roster of the RLG/OCLC Working Group

Neil Beagrie, Joint Information Systems Committee

Marianne Doerr, Bayerische Staatsbibliothek

Margaret Hedstrom, University of Michigan

Anne Kenney, Cornell University

Catherine Lupovici, Bibliothèque nationale de France

Kelly Russell, Cedars Project (CURL exemplars in digital archives)

Colin Webb, National Library of Australia

RLG Liaison: Robin Dale, Member Programs & Initiatives

OCLC Liaison: Meg Bellinger, Preservation Resources