An Audit Checklist for the Certification of Trusted Digital Repositories

DRAFT FOR PUBLIC COMMENT

RLG
Mountain View, CA
August 2005

National Archives and Records Administration
8601 Adelphi Road, College Park, Maryland, 20740-6001 USA

RLG
2029 Stierlin Court, Suite 100, Mountain View, California, 94043-4684 USA

Table of Contents

RLG-NARA TASK FORCE ON DIGITAL REPOSITORY CERTIFICATION II

INTRODUCTION 1

RLG-NARA DIGITAL REPOSITORY CERTIFICATION TASK FORCE 2
INTENDED AUDIENCE: 2
TERMINOLOGY 3
DRAFT FOR PUBLIC COMMENT 4

I. THE AUDIT & CERTIFICATION PROCESS 5

RELEVANT STANDARDS & BEST PRACTICES 5
THE AUDIT INSTRUMENT 5
THE CERTIFICATION PROCESS 6

II. AUDIT & CERTIFICATION CRITERIA 8

A. ORGANIZATION 8
A1. Governance & organizational viability 9
A2. Organizational structure & staffing 9
A3. Procedural accountability & policy framework 10
A4. Financial sustainability 11
A5. Contracts, licenses, & liabilities 12
B. REPOSITORY FUNCTIONS, PROCESSES, & PROCEDURES 14
B1. Ingest/acquisition of content 14
B2. Archival storage: management of archived information 19
B3. Preservation planning, migration, & other strategies 22
B4. Data management 27
B5. Access management 28
C. THE DESIGNATED COMMUNITY & THE USABILITY OF INFORMATION 32
C1. Documentation 32
C2. Descriptive metadata appropriate to the Designated Community 33
C3. Use & usability 34
C4. Verifying understandability 37
D. TECHNOLOGIES & TECHNICAL INFRASTRUCTURE 39
D1. System infrastructure 39
D2. Appropriate technologies 42
D3. Security 42

III. AUDIT INSTRUMENT FOR THE CERTIFICATION OF TRUSTED DIGITAL REPOSITORIES 45

GLOSSARY 59

REFERENCES 61

APPENDIX 1: A DISCUSSION OF UNDERSTANDABILITY & USE 62

RLG-NARA Task Force on Digital Repository Certification

Bruce Ambacher, ***Co-Chair***
National Archives and Records Administration

Kevin Ashley
University of London Computing Centre

John Berry
Internet Archive

Connie Brooks
Stanford University

Dale Flecker
Harvard University

David Giaretta
Rutherford Appleton Laboratory, Council for the Central Laboratory of the Research Councils, UK

Babak Hamidzadeh
Library of Congress

Keith Johnson
Stanford University

Maggie Jones
Digital Preservation Coalition, UK

Nancy McGovern
Cornell University

Don Sawyer
National Aeronautics and Space Administration

Johan Steenbakkers
Koninklijke Bibliotheek

Taylor Surface
OCLC

RLG staff liaison:
Robin Dale, ***Task Force Co-Chair***
RLG

> *For assuring the longevity of information, perhaps the most important role in the operation of a digital archive is managing the identity, integrity and quality of the archives itself as a trusted source of the cultural record.*
>
> Preserving Digital Information, 1996

Introduction

Almost a decade ago, the Task Force on Archiving of Digital Information (1996) declared "a critical component of digital archiving infrastructure is the existence of a sufficient number of trusted organizations capable of storing, migrating, and providing access to digital collections." The task force saw that "trusted" or trustworthy organizations could not simply identify themselves. To the contrary, the task force declared, "a process of certification for digital archives is needed to create an overall climate of trust about the prospects of preserving digital information." The task force stopped short of articulating the details of such a certification process. Certainly one obstacle was the fact that few organizations actually had digital repositories or archives at that time.

Work in articulating responsible digital archiving infrastructure was furthered by the development of the Open Archival Information System (OAIS) Reference Model. Designed to create a consensus on "what is required for an archive to provide permanent or indefinite long-term preservation of digital information," the OAIS addressed fundamental questions regarding the long-term preservation of digital materials that cut across domain-specific implementations. The reference model (ISO 14721) provides a common conceptual framework describing the environment, functional components, and information objects within a system responsible for the long-term preservation of digital materials (CCSDS, 2002). Long before it became an approved standard in 2002, many in the cultural heritage community had adopted OAIS as a model to better understand what would be needed from digital preservation systems. Institutions began to declare themselves "OAIS-compliant" to underscore the trustworthiness of their digital repositories, but there was no established definition of "OAIS-compliance," let alone a mechanism for measuring it.

In 2002, RLG and OCLC jointly published *Trusted Digital Repositories: Attributes and Responsibilities* (TDR), which further articulated a framework of attributes and responsibilities for trusted, reliable, sustainable digital repositories capable of handling the range of materials held by large and small cultural heritage and research institutions. The framework was broad enough to accommodate different situations, technical architectures, and institutional responsibilities while providing a basis for the expectations of a trusted repository. The document has proven to be useful for institutions grappling with the long-term preservation of cultural heritage resources and has been used in combination with the OAIS as a digital preservation planning tool.[1] As a framework, this document concentrated on high-level organizational and technical attributes and discussed potential models for digital repository certification. It refrained from being prescriptive about the specific nature of rapidly emerging digital repositories and archives and instead reiterated the call for certification of digital repositories, recommending the development of certification program and articulation of auditable criteria.

RLG-NARA Digital Repository Certification Task Force

In 2003, RLG and the National Archives and Records Administration created a joint task force to specifically address digital repository certification. The goal of this task force has been to develop criteria to identify digital repositories capable of reliably storing, migrating, and providing access to digital collections. The challenge has been to produce certification criteria and delineate a process for certification applicable to a range of digital repositories and archives, from academic institutional preservation repositories to large data archives and from national libraries to third-party digital archiving services.

Digital preservation infrastructure continues to grow through institutional funding and national initiatives like the US National Digital Information Infrastructure and Digital Preservation Program or the European Union Research Policy & Funding Framework Programmes. Research and development projects continue to address the remaining digital preservation challenges. While these challenges have not yet been resolved, the proliferation of experience, research, and infrastructure throughout the cultural heritage community has made trustworthy digital repositories conceptually realistic. Over the last two years, projects, programs, and collaborative work have begun to cultivate a shared view among stakeholders on well-defined infrastructure and processes for achieving certain digital preservation objectives.

The groundwork has been laid for the establishment of the long-awaited certification process for digital repositories. Over the last two years, the RLG-NARA task force has worked to define and articulate metrics or indicators of trustworthiness and reliability for digital repositories. The process has been iterative and the following criteria and documentation is the fifth generation of the expert group’s work. It represents best, current practice and thought about the organizational and technical infrastructure required to be considered trustworthy and capable of certification. It documents criteria that trustworthy repositories will be able to meet, providing explanations and examples. Section III of the document is the audit tool.

Intended audience

This document is produced primarily produced primarily for those who work in or are responsible for digital repositories seeking to be certified against its requirements and for those who will carry out the audit and certification process. To a great extent, these groups may be

[1] The Cornell University Library Digital Preservation Management considers *Trusted Digital Repositories: Attributes and Responsibilities* (TDR) and the *OAIS Reference Model* to be two foundational documents for institutions approaching digital preservation. The Cornell workshop advocates a “merged model” of digital preservation that combines the TDR and OAIS because, according to the workshop, the TDR on its own lacks an implementation model while the OAIS on its own lacks an organizational context. Together, Cornell states, they leverage community-based efforts and enable collaborative initiatives.

expected to overlap. Many of the requirements can only be tested by those with working knowledge of digital repository operations, so an audit is likely to be carried out, at least in part, by peers.

The requirements touch on every level of a repository's functions, so the document will be relevant to staff with many different roles within a repository and the organization of which it is part. The organization’s senior management and policy makers will need to be aware of at least the requirements of section A of the Audit and Certification Criteria. Systems and network staff, who may be responsible for many parts of infrastructure other than those specific to the repository, will have an interest in section D, which is also of relevance to those responsible for matters such as building security and fire protection. Those who deal with external users, whether producers of material or consumers of it, will find much that is relevant to their work in sections B and C.

The document is expected to be of interest to a wider community than this, however. Organizations planning repositories, and repositories that do not expect to seek certification but are, for example, themselves part of a chain of preservation, or are unsure of its relevance to them, are still likely to find much that is of interest here. The analysis of a repository’s functions, the itemized requirements, and the explanations of how they can be tested, can all help a repository plan and review its working practices. This may reassure the repository that it is operating in accordance with recognized best practice, it may help staff and users understand the repository’s actions, and it can help an organization focus limited resources where they will best ensure that digital resources will survive.

Both producers of digital material that will be preserved for long periods and users of this material will find much useful information here to help them understand what to expect from the repositories they deal with. It may help some producers streamline their interactions with the repositories that take long-term responsibility for their materials.

Even in the absence of any formal certification process, this document will help organizations considering outsourcing some or all aspects of digital preservation by showing how they can ensure that the organizations they contract with are carrying out the task of digital preservation in a way that deserves trust.

Terminology

Digital preservation interests a range of different communities, each with a distinct vocabulary and local definitions for key terms. A glossary is included to convey exact meaning of many of the terms in this document, but it is important to draw attention to the usage of several key ones.

In general, key terms in this document have been adopted from the OAIS Reference Model. One of the great strengths of the OAIS Reference Model has been to provide a common terminology made up of terms “not already overloaded with meaning so as to reduce conveying unintended meanings” (OAIS, 2002). Because the OAIS has become a foundational document for digital preservation, the common terms are well understood and are therefore used within this document. Definitions for some OAIS terms appear in the glossary along with other terms.

The OAIS Reference Model uses “digital archive” to mean the organization responsible for digital preservation. In this document, the term “repository” or phrase “digital repository” is used to convey the same concept in all instances except when quoting from the OAIS. It is important to understand that in all instances in this document, “repository” and “digital repository” are used to convey digital repositories and archives that have long-term preservation responsibilities and functionality.

Finally, this document names criteria which combined, evaluate the trustworthiness of digital repositories and archives. While the correct phase to describe such entities is “trustworthy digital repositories,” the community has long used “trusted digital repositories” to convey that same value assessment. While grammatically incorrect, it is never-the-less the phrase most familiar and engaged within the community. Therefore, this document does refer to trustworthiness and *trusted* digital repositories.

Draft for public comment

This is a draft for public comment. Comments and constructive criticism about the audit instrument as well as the documentation are welcomed and encouraged. Since this work may affect all digital repositories, comments from all stakeholders will be critical to the final audit instrument and certification process. All comments will be considered by the reconvened task force in early 2006.

Please send comments to Robin Dale at RLG (Robin.Dale@rlg.org) by the close of business 15 January 2006.

I. The Audit & Certification Process

Certification for digital repositories will involve far more than the documentation of criteria. To be useful, a full certification process must provide tools to allow for planning and self-examination, as well as an external, objective audit. It must recognize standards and best practices relevant to the community of the repository, as well as those of the information management industry as a whole. In other words, audit and certification of trusted digital repositories cannot exist in a vacuum. This effort was conceived because no single organization or standards body had created the necessary tools and evaluative process. Organizations and service providers have been claiming trustworthiness without a way to prove it. For institutions seeking third-party archiving solutions, as well as government funding agencies providing mass amounts of funding for digital content creation, this is untenable. A standard, objective method for audit and certification of digital repositories is a mandatory tool in the digital preservation arsenal. This work of the RLG-NARA Task Force on Digital Repository Certification is the first phase of fulfilling that requirement.

Relevant standards & best practices

This work started with research for existing, relevant documentation and standards. The results, as well as ongoing initiative tracking, highlighted numerous documents and standards from which pieces were applicable or related to this work. None, however, could be used in lieu of developing this specialized audit and certification process. A few examples:

- The ISO 9000 family of standards addresses components of organization and system management that, while valuable, were not specifically developed to gauge the trustworthiness of organizations operating digital repositories.
- Similarly, ISO 17799 was developed specifically to address data security and information management systems. Like ISO 9000, it has some very valuable components to it but it was not designed to address the trustworthiness of digital repositories. Its requirements for information security seek data security compliance to a very granular level, but do not address organizational, procedural, and preservation planning components necessary for the long-term management of digital resources.

It is important to acknowledge that while not completely duplicating the digital repository audit requirements, there is real value in knowing whether an institution is certified to related standards. In fact, it is all but ensured that if an organization is ISO 17799 certified, it will completely meet all of the criteria found in Section D of the Audit and Certification Criteria of this digital repository audit instrument. Certainly, an institution that has undertaken any kind of certification process—even if none of the evaluated components overlap with a digital repository audit—will be better prepared for digital repository certification.

The audit instrument

This new audit tool is the work of many experts representing the widest international range of communities in research, governments, and cultural heritage organizations. All were chosen because of their experience building and managing digital repositories. To be of true value, the tools for auditing repositories needed to be developed by practitioners. The group gathered for this task represents over 180 years of collective experience in information technology and systems and more than 135 years of collective experience in the preservation of digital information (data archives, electronic records, digital repositories, etc.).

The tool went through several different frameworks before taking its current form. It is divided into four sections:

A. Organization
B. Repository functions, processes, and procedures
C. Designated community and intended uses of the information
D. Technology and technical infrastructure

These are explained in Section II, Audit & Certification Criteria.

Section II provides the textual documentation of the metrics/criteria, explanations where necessary, and examples of how an institution may be able to prove it meets the metric/criteria. For institutions using the audit instrument as a planning tool or for self-evaluation, the documentation should provide sufficient examples though in almost all instances, the example information is illustrative rather than prescriptive.

Section II groups the metrics by the sections named above. Individual metrics are further numbered within each section. This numbering sequence allows easy cross referencing across and between metrics, as well as with the actual audit instrument (Section III).

The tool is designed to allow an auditor to record the level of readiness or fulfillment of each metric. For each metric, the auditor can rate an institution's progress toward satisfying the criteria: planning, documentation, implementation, and evaluation. It is assumed that a repository will go through each of these steps to achieve proficiency and meet the requirements of the metric. In some cases, a repository may have plans and documentation that meet requirements, but may not have implemented or evaluated—or verified—its plans. The audit tool allows for complete evaluation and documentation of the cycle of development without imposing a pass/fail rating.

It is hoped that as a draft for public comment this document will encourage institutions and individuals to read, understand, and potentially use this draft to evaluate their local repositories. Certainly any certification process will always begin with an institution performing a self-assessment, so this documentation provides an early opportunity to begin the process. Comments and constructive criticism are welcome to help the task force ensure that this document and audit instrument play a valuable role in the future certification of digital repositories.

The certification process

The effort to take this work from proposal to implementation is already underway. Under the auspices of the Center for Research Libraries and through a grant from The Andrew W. Mellon Foundation, this audit instrument and the work of the task force are being leveraged to develop the final piece, the certification process. In tandem with the public comment phase, the CRL Audit and Certification of Digital Archives Project (www.crl.edu/content.asp?l1=13&l2=58&l3=142) will use the audit instrument to test audit three archives. Test audit results will be private because the project is designed only to contribute to the further development of the audit and certification process.

The mechanics of the certification process have yet to be determined though will also be addressed as a part of the CRL project. The goal is to develop a single process that can apply to all types and development levels of digital repositories. It is likely that the eventual certification will involve levels of certification which signify an organization's readiness and development level. Clearly the goal would be to have all repositories meet the highest level of standards and be certified as fully trustworthy, but it is assumed that some organizations may step through the levels as they advance in their development phases. Levels of certification would allow for objective, truthful assessments of organizational infrastructure, technical infrastructure, and preservation readiness. It would allow repositories to participate in a growing culture of assessment rather than deter them for not having reached the highest level of trustworthiness. In the end, objective audit and certification ratings are beneficial for repositories, producers of information, funders, government agencies, and especially the users of the material.

II. Audit & Certification Criteria

A. Organization

Though adequate technical architecture, processes, and capabilities underpin a trusted digital repository, the technical aspects are only one piece of the overarching infrastructure supporting the digital repository functions. Organizational attributes of digital repositories are equally critical.

Organizational attributes are characteristics of the repository organization that affect performance, accountability, and sustainability. *Trusted Digital Repositories: Attributes and Responsibilities* (2002) grouped these types of attributes into four of its "Attributes of a Trusted Digital Repository": administrative responsibility, organizational viability, financial sustainability, and procedural accountability. In their training workshop *Digital Preservation Management*, Cornell University Library refers to these characteristics as "organizational infrastructure." According to Cornell, "an organization's infrastructure is best embodied in its policies and procedures" and documentation of organizational infrastructure is embodied in three distinct levels: policy framework, policies and procedures, and plans and strategies (Cornell, 2005). Organizational attributes are indicators of a digital repository's comprehensive planning, readiness, ability to address its responsibilities, and trustworthiness.

Organizational infrastructure includes but is not restricted to these elements:

- governance
- organizational structure
- mandate or purpose
- scope
- roles and responsibilities
- policy framework
- funding system
- financial issues, including assets
- contracts, licenses, and liabilities
- transparency

Metrics/criteria addressing these elements are grouped in these five sections:

A1 Governance and organizational viability
A2 Organizational structure and staffing
A3 Procedural accountability and policy framework
A4 Financial sustainability
A5 Contracts, licenses, and liabilities

A1. Governance & organizational viability

A repository must demonstrate an explicit, tangible, and long-term commitment to compliance with prevailing standards, policies, and practices.

A1.1 Repository has a mission statement that reflects a commitment to the long-term retention of, management of, and access to digital information on behalf of depositors.

The mission statement of the repository must be accessible to depositors and other stakeholders and contains an explicit long-term commitment.

A1.2 Repository has a formal succession plan, contingency plans, and/or escrow arrangements in place in case the repository ceases to operate or substantially changes its scope.

Part of the repository's perpetual-care promise is a commitment to identifying appropriate successors or arrangements should the need arise. Consideration needs to be given to this responsibility while the repository is viable, not when a crisis occurs, to avoid irreparable loss. A formal succession plan should include the identification of trusted inheritors (if applicable), return of digital objects to depositors with adequate prior notification, etc.

A2. Organizational structure & staffing

A repository must have designated staff with requisite skills and training and must provide ongoing development.

A2.1 Repository staff have skills and expertise appropriate to their duties.

The repository must demonstrate that the staff and consultants have the range of requisite skills—e.g., archival training, technical skills, and legal expertise.

A2.2 Repository has the appropriate number of staff to support all functions and services designated in agreements with depositors.

Staffing for the repository must be adequate for the scope and mission of the archiving program. Understaffing indicates that the repository cannot fulfill its agreements.

A2.3 Repository commits to professional development to keep staff expertise and skills current.

Technology will continue to change, so the repository must have a lifelong learning approach to developing and retaining staff. As the requirements and expectations pertaining to each functional area evolve, the repository must demonstrate that staff are prepared to face new challenges.

A3. Procedural accountability & policy framework

A repository must provide clear and explicit documentation of its requirements, decisions, development, and actions to ensure long-term access to digital content in its care. This documentation assures Consumers, management, Producers, and certifiers that the repository is meeting its requirements and fully performing its role as a trusted digital repository. Certification, the clearest indicator of a repository's sound and standards-based practice, is facilitated by procedural accountability that results in comprehensive and current policies, procedures, and practice.

A3.1 Repository has a mechanism in place for reviewing, updating, and developing comprehensive policies and procedures as repositories grow and as the community practice evolves.

The policies and procedures of the repository must remain current and must evolve to reflect changes in requirements and practice. The repository must demonstrate that a policy and procedure audit and maintenance is in place and regularly applied. Versions of these documents must be well managed by the repository (e.g., outdated versions are clearly identified or maintained offline) and qualified staff and peers must be involved in reviewing, updating, and extending these documents.

A3.2 Repository has monitoring and feedback mechanisms in place to ensure continued operation, resolve problems, and address evolving requirements of providers and Consumers.

The repository must demonstrate reliability in all its operations and support to its range of users. Reliability and sustainability are essential to establishing trust in the repository.

A3.3 Repository is committed to formal, periodic review and assessment to ensure continued development.

Long-term preservation is a shared and complex responsibility. A trusted digital repository contributes to and benefits from the breadth and depth of community-based standards and practice. Regular review is a requisite for ongoing and healthy development of the repository.

A3.4 Repository has a documented history of the changes to its operations, procedures, software, and hardware, traceable to its preservation strategies where appropriate.

The repository must demonstrate the full range of its activities and developments over time. Documenting decisions about the organizational and technological infrastructure of the repository is a core responsibility.

A3.5 Repository commits to transparency and accountability in all actions supporting the operation and management of the repository.

Transparency is the best assurance that the repository operates in accordance with accepted standards and practice. Accountability cannot be achieved without transparency. The two together are the basis for trust in the repository. Both are achieved through active, ongoing documentation.

A3.6 Repository commits to define, collect, track, and provide, on demand, its information integrity measurements.

The repository must develop or adapt appropriate measures for ensuring the integrity of its holdings. The chain of custody for all of its digital content from the point of deposit forward must be explicit, complete, correct, and current. The repository must demonstrate that the content it has matches the content it received. Losses associated with migration and other preservation actions should also be documented. (See D1.5 and D1.6.)

A3.7 Repository commits to a regular schedule of certification and to notifying certifying bodies of operational changes that will change or nullify its certification status.

A repository cannot self-certify. Therefore, certification is the best indicator that the repository meets its requirements, fulfills its role, and adheres to appropriate standards. The repository must demonstrate that it integrates certification preparation and response into its operations and planning.

A4. Financial sustainability

A trusted digital repository should be able to prove its financial sustainability over time. Overall, trusted repositories will adhere to all good business practices and should have a sustainable business plan in place. Normal business and financial fitness should be reviewed at least annually. Standard accounting procedures should be used. Both short- and long-term financial planning cycles should demonstrate an ongoing commitment to a balance of risk, benefit, investment, and expenditure. Operating budgets and reserves should be adequate.

A4.1 Repository has short- and long-term business planning processes in place to support sustainability.

The repository must demonstrate that it has formal, cyclical, proactive business planning processes in place as enumerated in the following metrics. Similar to metric A1.2 (succession/contingency/escrow planning), the repository must establish these processes when it is viable to avoid business crises.

A4.2 Repository has in place at least annual processes to review and adjust business plans as necessary.

The repository must demonstrate its commitment to proactive business planning by performing cyclical planning processes at least yearly.

A4.3 Repository business planning and practices are transparent, compliant with relevant accounting standards and practices, and auditable.

The repository must demonstrate that it adjusts its business practices as necessary over time to keep them transparent, compliant, and auditable.

A4.4 Repository has ongoing commitment to risk, benefit, investment, and expenditure analysis and reporting (including assets, licenses, and liabilities).

The repository must commit to at least these categories of analysis and reporting, and maintains an appropriate balance between them.

A4.5 Repository recognizes the eventual strong possibility of a gap between repository-generated funding and the funding necessary to meet the repository's commitments to its depositors. It commits to bridging these gaps by securing funding and resource commitments specifically for that purpose; these commitments can come either from the repository itself or parent organizations, as applicable.

Even with effective business planning procedures in place, any repository with long-term commitments will likely face some kind of resource gap in the future. The repository must provide essentially an insurance buffer as a first—and hopefully effective—line of defense, obviating the need to invoke a succession plan except in extreme situations (such as the repository ceasing operations permanently).

A5. Contracts, licenses, & liabilities

A5.1 If repository manages, preserves, and/or provides access to digital materials on behalf of another organization, it has and maintains appropriate contracts or deposit agreements.

Repositories, especially those with "third-party" deposit arrangements, should guarantee that appropriate contracts, licenses, or deposit agreements express rights, responsibilities, and expectations of each party. Contracts and formal deposit agreements should be countersigned and current.

When the relationship between depositor and repository is less formal (i.e., a faculty member depositing work in an academic institution's preservation repository), documentation articulating the repository's capabilities and commitments should be provided to each depositor.

A5.2 Repository's contracts or deposit agreements specify and/or transfer appropriate preservation rights, as necessary.

Because the right to change or alter digital information is often restricted by law to the creator, it is important that digital repositories address the anticipated need to be able to work with and potentially modify digital objects to keep them accessible over time. Repositories should have written policies and agreements with depositors that specify and/or transfer certain rights to the repository enabling appropriate and necessary preservation actions to take place to the digital objects within the repository.

Because legal negotiations can take time, potentially preventing or slowing the ingest of digital objects at risk, it is acceptable for a digital repository to take in or accept digital objects even with only minimal preservation rights in an open-ended form and then deal with expanding to detailed rights later.

A repository's rights must at least limit the repository's liability or legal exposure that threatens the repository itself. A repository cannot be said to have sufficient control of the information if the repository itself is legally at risk.

A5.3 Repository tracks and manages copyrights and restrictions on use as required by contract or license.

The repository should have a mechanism for tracking licenses and contracts to which it is obliged. Whatever the format of the tracking system, it must be sufficient for the institution to track, act on, and verify rights and restrictions related to the use of the digital objects within the repository.

A5.4 If repository ingests digital content with unclear ownership/rights, it has policies addressing liability and challenges to those rights.

The repository's policies and mechanisms must be vetted by appropriate institutional authorities and/or legal experts to ensure that responses to challenges adhere to relevant laws and requirements, and that the potential liability for the repository is minimized.

B. Repository Functions, Processes, & Procedures

This section addresses the repository functions, processes, and procedures needed to ingest, manage, and provide access to digital objects for the long term. It specifically does not cover the technical or system infrastructure requirements. (See Section D for those criteria.) Requirements for these functions are categorized into five sections based on archive functionality, allowing grouping under the well-known OAIS functional entities.

Section B1 articulates the requirements associated with the ingest or acquisition of digital content. Ingest is the crucial interaction between repository and depositor. Successful ingest also marks the ability of the repository to gain sufficient control over the content. It involves procedural and system-related tasks for the repository.

Section B2 establishes a minimal set of conditions for long-term preservation of AIPs. The system infrastructure (discussed in D1) must provide suitable services to allow higher-level repository functions operating on AIPs to perform their tasks in a reliable manner. But if the higher-level functions do not use these services, or do not use them properly, then preservation is not assured.

Section B3 addresses the current, sound, and documented preservation strategies a repository must have in place and demonstrably implemented to assure that the digital content will remain accessible over the long-term.

Section B4 addresses the requirements for minimal level metadata that allow digital objects to be located, as well as managed within the system.

Section B5 establishes requirements for a repository’s access functionality. It addresses the ability to produce and disseminate accurate and authentic versions of the digital objects within the repository.

B1. Ingest/acquisition of content

“Ingest” is a generic term to describe the processes that take place before the final, preserved form of an object is present in the repository. Repositories are likely to differ the most in this area, depending on the type of material they collect and their relationships with its Producers. For any repository, it can be stated with some confidence that ingest finishes when an AIP and its associated metadata are secure in the repository, including the creation of any security copies. It is more difficult to make a general statement about when ingest begins. Some repositories will have content submitted to them by Producers, perhaps unexpectedly. Others will actively go out and seek content and request it from Producers. Some Producer-repository relationships will be more collaborative, making it less clear-cut who initiates a particular transaction.

Relationships between Producers and repositories that affect ingest can differ greatly in their formality and the extent to which obligations are placed on different parties. National archives and copyright libraries may be able to compel their Producers (government agencies and publishers) to provide content, but may have little or no control over its form. Other repositories may not be able to compel Producers to offer content, but might be able to select the form of acceptable content, whether that applies to file formats or minimal metadata standards, for instance. Some repositories (Web archives, for example) may have little or no relationship with the Producers of the content they preserve.

Given these differences, some of the requirements here are very general, and require judgments about what is appropriate for a repository given its stated mission and the needs of its Designated Community. But the result that all repositories are trying to achieve is the same: to preserve content that is understandable and usable in the long term.

The digital objects a repository accepts for preservation should reflect both its mission statement and its Designated Communities' spheres of interests. Users should clearly understand the relationship between a repository, its mission statement, and its collections. Likewise, the documentation associated with the primary digital objects of a collection should be just as logical. The information to be transferred with specific digital objects (primary and others) will be enunciated in the specific transfer agreement for those objects and should be the information and items necessary for Consumers to use the objects without resort to the Producers, to other experts, and hopefully to subject matter experts in the repository itself.

In a general, generic statement it is impossible to specifically enumerate the documentation required for each digital object being preserved by a certified repository. Complete documentation may include metadata, codes, sample forms, record layouts, explanations of the universe, minimum and maximum values, and related studies and results. The documentation is collected both to ensure completeness of the collection and to help the Consumer determine the accuracy or correctness of the data itself. That determination is normally made by the Producer and the Consumer, rather than the repository.

Fundamentally, the repository is tasked to preserve information, which means digital objects together with their Representation Information. This is the primary information to be preserved and is called the Content Information in OAIS terminology. (This is also applicable to the Preservation Description Information—Provenance, Context, Fixity, and Reference.) A fundamental decision, to be taken by the repository together with the Producer, is the definition of what constitutes the information to be preserved, or Content Information. The OAIS recommendation is to start by deciding what is the Primary Digital Object and then to address the extent of the Representation Information that needs to accompany this digital object. The extent of this Representation Information is not predefined and may vary widely from one submission to another even within a given repository.

Useful examples that demonstrate the types and extent of documentation that should be collected for various types of data objects and archival information collections can be seen in Annex A of Consultative Committee for Space Data Systems, Reference Model for an Open Archival Information System (OAIS) (ISO 14721); and in the U.S. National Archives and Records Administration Electronic and Special Media Records Services Division, Accessioning Procedures Handbook (College Park, MD, loose leaf June 2000). Cooperative efforts include the Data Documentation Initiative (www.icpsr.umich.edu/DDI/org/index.html), formally established in 2003, which is promoting an XML Document Type Definition that has been widely adopted in some disciplines, and the Council of European Social Science Data Archives (www.nsd.uib.no/cessda/), which promotes the preservation and exchange of data and technology and the establishment of new organizations to do the same through the use of metadata standards, common thesauri and standardized rights management, as well as standardized cataloguing of data object entries.

B1.1 Repository identifies properties it will preserve for each class of digital object.

This process begins in general with the repository's mission statement and is further specified in pre-accessioning agreements with Producers or depositors (e.g., Producer-archive agreements) and made very specific in deposit or transfer agreements for specific digital objects and their related documentation. For example, one repository may only commit to preserving the textual content of a document and not its exact appearance on a screen; another may wish to preserve the exact appearance and layout of textual documents.

B1.2 Repository has specified all appropriate aspects of acquisition, maintenance, access, and withdrawal issues in written agreements with depositors.

The deposit agreement specifies all aspects of these issues that are necessary for the repository to carry out its function. There may be a single agreement covering all deposits, or specific agreements for each deposit, or a standard agreement supplemented by special conditions for some deposits. These special conditions may add to the standard agreement or override some aspects of the standard agreement. Agreements may need to cover restrictions on access and will need to cover all property rights in the digital objects. Agreements may place responsibilities on depositors, such as ensuring that SIPs conform to some pre-agreed standards, and may allow repositories to refuse SIPs that do not meet these standards. Other repositories may take responsibility for fixing errors in SIPs. The division of responsibilities must always be clear.

B1.3 Repository has an identifiable, written definition for each SIP or class of information ingested by the repository.

The written inventory portion of a deposit agreement or transfer agreement should specify exactly what digital object(s) are transferred, what documentation is associated with the object(s), and any restrictions on access.

B1.4 Repository has a process to ensure that the information is acquired from the expected source.

The repository's written standard operating procedures (SOP) and the actual practices followed must ensure the digital object(s) are obtained from the expected source and are the expected object(s). Confirmation occurs through digital processing and data verification and validation and through exchange of appropriate instrument of ownership (deed of gift).

B1.5 Repository obtains sufficient physical control over the digital objects to preserve them.

The repository can obtain physical control of the digital objects through several of these activities:

- **Analysis of the digital content:** Depending upon a repository's mission and goals, this may involve the repository, in consultation with the depositor/rights owner and systems managers, assessing the digital object and determining which of its properties are significant for preservation. For other repositories and digital archives, analysis of digital content may be accomplished through automated tools that compare the digital objects against expected and/or acceptable formats or other mechanisms that analyze the content systematically as material is deposited into the repository. To ensure long-term preservation, digital repositories need to decide what level of preservation is appropriate for each digital object or class of objects. 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 bitstream to ensure the object is accessible to the agreed-on level.

- **Verification, analysis, and creation of metadata**: Any metadata that accompanies the object when it is submitted to the repository must be verified and, as necessary, enhanced to support the object's long-term maintenance as well as continuing access. 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 interpreting the bitstream as a meaningful digital object ensure current usability by the contemporaneous Designated Community and form the basis of long-term preservation. How continuing access is provided over time can and should be kept separate, conceptually, from this basic preservation function.

- **Authentication and integrity checking:** The repository needs to ensure that mechanisms are in place for verifying the digital object, including all associated metadata. This should include verifying not only the integrity of the bitstream but also confirming the object's usability and functionality. Integrity of bitstreams should be verified as well as usability for whole classes of digital objects, for example, the preservability of the format should be examined.

- **Creation of the Archival Information Package:** Digital repositories can store a digital object and its associated metadata in two ways: as a single bitstream 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 other systems, such as bibliographic data stored in the library management system. Such "virtual encapsulation" avoids duplicating metadata, but separating a digital object and its metadata may present problems in the future. Some experts feel that long-term preservation may be best served by storing the digital content and as much as possible of its relevant metadata as a single file.

> The Metadata Encoding and Transmission Standard (METS) and the emerging XFDU standard (the extensible data packaging format) are exemplars of potential AIP encapsulation structures. The METS standard (2005) was created by the cultural heritage community for encoding descriptive, administrative, and structural metadata. Depending on its use, a METS document could take the role of Submission Information Package (SIP), Archival Information Package (AIP), or Dissemination Information Package (DIP). The XFDU standard (2005), currently under construction within the Consultative Committee on Space Data Systems (CCSDS), is similar to METS and can serve the same "packaging" functions as METS. Other communities may use different, community-generated packaging or encapsulation structures. The only requirement for packaging structures is that they are well documented and, if not openly accessible, the documentation can be produced on demand for auditors.

B1.6 Repository's ingest process verifies each SIP for completeness and correctness.

Information collected during the ingest process must be compared with information provided by the Producer to verify the *completeness* of the data transfer and ingest process. Checking for completeness can mean simply checking that a file has not been truncated during transfer. It can also mean checking that a group of files contains all the expected members—all of the images from a satellite on a particular day, or all the documents pertaining to a meeting, to give two examples.

Information collected during the ingest process must be compared with information provided by the Producer to verify the *correctness* of the data transfer and ingest process. The extent to which a repository can determine correctness will depend on what it knows about the SIP and what tools are available for verifying correctness. It can mean simply checking that file formats are what they claim to be (TIFF files are valid TIFF format, for instance), or can imply checking the content. This might involve human checking in some cases, such as confirming that the description of a picture matches the image.

B1.7 Repository provides Producer/depositor with appropriate responses at predefined points during the ingest processes.

Based on the initial processing plan and agreement between the repository and the Producer/depositor, the repository must provide the Producer/depositor with progress reports at specific, predetermined points throughout the ingest process. Responses can include initial ingest receipts, or receipts that confirm that archiving is complete. Sample reports could include copies of the ingest completeness and correctness reports and error reports and any final transfer of custody document.

B1.8 Repository can demonstrate that all SIPs are either accepted as whole or part of an eventual AIP, or otherwise disposed of in a recorded fashion.

The timescale of this process will vary between repositories from seconds to many months, but SIPs must not remain in a limbo-like state forever. The accessioning procedures and the internal processing and audit logs should maintain records of all internal transformations of SIPs and thus demonstrate that they either become AIPs (or part of AIPs) or are disposed of. Appropriate descriptive information should also document the provenance of all digital objects.

B1.9 Repository can demonstrate when preservation responsibility is formally accepted for the contents of the SIP.

A key component of a repository's responsibility to “gain sufficient control” of digital objects is the point when the transformation is made from ingested SIP to AIP. At this point, the repository formally accepts preservation responsibility of digital objects from the depositor. Repositories generally will mark this acceptance with some form of notification to the depositor. (This may depend on repository responsibilities as designated in the depositor agreement.) A repository may mark the transfer by sending a formal document, often a final signed copy of the transfer agreement, back to the depositor signifying the completion of the SIP-AIP ingest transformation process. Other approaches are equally acceptable. Brief daily updates may be generated by a repository that only provides annual formal transfer reports.

B2. Archival storage: management of archived information

Digital repositories must take actions to preserve the ingested information and the things they disseminate to end users must be strongly linked to the original objects that were deposited.

To paraphrase the OAIS, the requirements of this section are meant to ensure information (digital objects and all appropriate metadata) received and verified from each Producer, is put into the archival form (AIP) and is stored in Archival Storage for long-term preservation. More specifically, the repository must actually complete the ingest process, creating some appropriate form—identifiable as archival storage—in which to store the information.

B2.1 Repository has an identifiable, written definition for each AIP or class of information preserved by the repository.

It is merely necessary that definitions exist for each AIP, or class of AIP if there are many instances of the same type. Repositories that store a wide variety of object types may need a specific definition for each AIP they hold, but it is expected that most repositories will establish class descriptions that apply to many AIPs. It must be possible to determine which definition applies to which AIP.

This metric is primarily concerned with issues of format and representation. Note that the next metric places more stringent conditions on the content of the definitions to ensure that they are fit for the intended purpose. Separating the two metrics is important, particularly if a repository does not satisfy one of them. It is important to know whether the problem is that some or all AIPs are not defined, or that the definitions exist but are not adequate.

B2.2 Repository has a definition of each AIP (or class) that is adequate to fit long-term preservation needs.

In many cases the mere existence of the definitions required by the previous metric will mean that this metric is also satisfied, but it may also be necessary for the definitions to say something about the semantics or intended use of the AIPs if this could affect long-term preservation decisions. To take a simple example, two repositories may both only preserve digital still images, and each uses multi-image TIFF files as their preservation format. Repository 1 consists entirely of real-world photographic images intended for viewing by people, and it has a single definition covering all of its AIPs. (The definition may refer to a local or external definition of the TIFF format.) Repository 2 contains some images, such as medical x-rays, that are intended for computer analysis rather than viewing by the human eye, and other images that are like Repository 1. Repository 2 should perhaps define two classes of AIPs, even though it only uses one storage format for both. A future preservation action may depend on what the intended use of the image is—an action that changes the bit-depth of the image in a way that is not perceivable to the human eye may be satisfactory for real-world photographs but not for medical images, for example.

B2.3 Repository has a definition of how AIPs are derived from SIPs.

The repository must be able to show how the preserved object is derived from the object initially submitted for preservation. In some cases the AIP and SIP will be identical apart from packaging and location, and the repository need only state this to meet the metric. More commonly, complex transformations may be applied to objects during the ingest process and a precise description of these actions may be necessary to ensure that the preserved object represents the information in the submitted object. Some repositories may need to produce these on a case-by-case basis, in which case diaries or logs of actions taken to produce each AIP will be needed. Other repositories that can run a more production-line approach may have a description for how each class of incoming object is transformed to produce the AIP. It must be clear which definition applies to which AIP. If, to take a simple example, two separate processes each produce a TIFF file, it must be clear which process was applied to produce a particular TIFF file.

B2.4. Repository has and uses a naming convention that can be shown to generate visible, unique identifiers for all AIPs.

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. Equally important is a system of reliable linking/resolution services in order to find the uniquely named object, no matter its physical location. This is so that actions relating to AIPs can be traced over time, over system changes, and over storage changes. Ideally the unique ID lives as long as the AIP; if it does not, there must be traceability. The ID system must be seen to fit the repository's current and foreseeable future requirements for things like numbers of objects. It must be possible to demonstrate that the identifiers are unique.

B2.5 If unique identifiers are associated with SIPS before ingest, they are preserved in a way that maintains a persistent association with the resultant AIP.

SIPs will not always contain unique identifiers when they are received by the repository. But where they do, and particularly where those identifiers were widely known before the objects were ingested, it is important that they are either retained as is, or that some mechanism allows the original identifier to be transformed into one used by the repository.

For example, consider an archival repository whose SIPs consist of file collections from electronic document management systems (EDMS). Each incoming SIP will contain a unique identifier for each file within the EDMS, which may just be the pathname to the file. The repository cannot use these as they stand, since two different collections may contain files with the same pathname. It may generate unique identifiers by qualifying the original identifier in some way (e.g., prefixing the pathname with a unique ID assigned to the SIP of which it was a part.) Or it may simply generate new unique numeric identifiers for every file in each SIP. If it qualifies the original identifier, it must explain the scheme it uses. If generates entirely new identifiers, it will probably need to maintain a mapping between original IDs and generated IDs, perhaps using object-level metadata.

B2.6. Repository verifies each AIP for completeness and correctness when generated.

If the repository has a standard process to verify SIPs for either or both completeness and correctness and a demonstrably correct process for transforming SIPs into AIPs, then it simply needs to demonstrate that the initial checks were carried out successfully and that the transformation process was carried out without indicating errors.

Repositories that must create unique processes for many of their AIPs will also need to generate unique methods for validating the completeness and correctness of AIPs. This may include performing tests of some sort on the content of the AIP that can be compared with tests on the SIP. Such tests might be simple (counting the number of records in a file, or performing some simple statistical measure such as calculating the brightness histogram of an original and preserved image), but they might be complex or contain some subjective elements.

B2.7. Repository provides an independent mechanism for audit of the integrity of the repository collection/content.

In general it is likely that a repository that meets all the previous metrics will satisfy this one without needing to demonstrate anything more. As a separate metric it demonstrates the importance of being able to audit the integrity of the collection as a whole.

For example, if a repository claims to have all e-mail sent or received by The Yoyodyne Corporation between 1985 and 2005, it has been required to show that:

- The content it holds came from Yoyodyne’s e-mail servers.
- It is all correctly transformed into a preservation format.

- Each monthly (say) SIP of e-mail has been correctly preserved, including original unique identifiers such as Message-IDs.

However it may still have no way of showing whether this really represents all of Yoyodyne's e-mail: if there is a three-day period with no messages in the repository, is this because Yoyodyne was shut down for those three days, or was the e-mail lost before the SIP was constructed ? This case could be resolved by the repository amending its description of the collection, but others may not be so straightforward.

A familiar mechanism from the world of traditional materials in libraries and archives is an accessions or acquisitions register that is independent of other catalog metadata. A repository should be able to show, for each item in its accessions register, which AIP(s) contain content from that item. Alternatively it may need to show that there is no AIP for an item, either because ingest is still in progress, or because the item was rejected for some reason. Conversely, an arbitrary AIP should be able to be related to an entry in the acquisitions register.

B3. Preservation planning, migration, & other strategies

A repository must have current, sound, and documented preservation strategies in place and demonstrably implemented. It is not enough simply to preserve information. A repository must do so in accordance with predefined, documented policies and procedures. Without documents a repository cannot pass an audit, even if its work is otherwise exemplary.

Documents need not be particularly complex. They also do not need to prescribe in detail how a repository will deal with the unknown. For instance, a repository cannot be required to document how it will preserve a file format that has not yet been invented. But it may be expected to describe what it will do when first presented with an object in a format that it has not encountered before. Organizational policy may be to reject it or to investigate the feasibility of dealing with it, or the decision may depend on other factors, such as who the object is from or what information it contains.

A trusted digital repository cannot simply say what it will do; it must demonstrate—be transparent about—its policies, practices, and procedures. This documentation should be explicit, comprehensive, current, complete, and available.

The repository must be able to demonstrate:

- Relevant decisions about acceptable formats.
  Examples: standalone or portions of policies that restrict, define, or stipulate formats that may be accepted by the repository.

- Comprehensive automated and/or manual workflow for bringing in appropriate digital objects.
  Examples: protocols for transfer, including roles and responsibilities of the Producer and the repository; explicit evidence of conversions that occur in AIPs that are generated from SIPs; quality assurance mechanisms and measures for assuring the completeness

and correctness of resulting AIPs.

- Anticipated and/or applied preservation actions pertaining to individual and classes of AIPs.
  Examples: preservation plans—planned, tested, and/or applied; preservation action logs; policies that address preservation strategies.

- Archival storage policies, procedures, and practices that ensure effective capture, ongoing and reliable archival storage, and responsiveness to inevitable technological change.
  Examples: storage management investment and planning documents, comprehensive security plans to enable the workflow, measures and monitoring protocols for stored AIPs.

- Independent means to verify expected repository content based on a secure trace of digital objects received.
  Examples: an auditable acquisitions register, an inventory that cannot be altered.

This is a key set of activities for collecting those things that make the information available and usable for future generations. The preservation strategy lays out a plan for carrying this out within an evolving environment (social/technological, etc). The strategy must provide for:

- A process for monitoring change that might affect preservation.
- An understanding/expertise for interpreting the impact/implications of these changes.
- A planned response to these changes.
- An implementation of this response.

A strategy must also state the conditions under which the deletion of AIPs is allowed. For example, a repository holds a proprietary format, the software for which is expected to become unsupportable and eventually unusable. Potential strategies are:

- Transform data upon ingest of the format.
- Keep original format and wait for others to produce a solution to the support of the software.
- Produce a supportable emulation environment to enable the proprietary software to continue to run.

Strategies may be needed for each class (e.g., format) of digital data held by the repository.

A strategy would also be expected to have special checks on AIPs over and above those performed as part of the normal robust infrastructure. These would include packaging the various components of the AIP—Content Information, Representation Information, Preservation Description Information (PDI), Packaging Information, and Package Description—and fixity checks on access to or movement of data, e.g., checksums, digests, error correction encodings, etc., including random sampling of holdings to monitor possible degradation of media. Updates are allowed to AIPs, e.g., to incorporate additional PDI. This must produce a new edition/version of an AIP.

Other transformations may be applied to SIPs and AIPs to generate further AIP versions. For example the repository may wish to keep a more easily preservable format for a particular type of data—making life easier for the repository and more suitable for the Community. It is important that contemporaneous records (e.g., logs of processes, history, etc.) be kept of these transformations as well as at least the receipt of SIPs and creation of AIPs. It is difficult to specify the level of detail of this recording. This logging may be of sufficient detail to allow one to regenerate one version from the next or vice versa in a reversible way. In this case the repository would be able to generate versions of AIPs as required.

Alternatively, if the logging is not sufficiently detailed for this then each AIP version would have to be kept or the deletion of intermediate versions recorded. The original AIP should never be deleted unless allowed as part of an approved strategy.

B3.1 Repository has documented preservation strategies.

The repository must show these are indeed preservation strategies.

B3.2 Repository implements/responds to strategies for AIP storage and migration.

At least two aspects of the strategy must be acted upon: that which pertains to how AIPs are currently stored (including physical requirements, media requirements, location of copies, formats and metadata) and that which may require AIP migration of any form.

If a repository has not existed long enough to have needed to carry out any sort of AIP migration, it must demonstrate that its policy has not required migration yet.

B3.3 Repository uses appropriate international Representation Information (including format) registries.

The Global Digital Format Registry (GDFR), the UK National Archives' file format registry PRONOM, and the UK Digital Curation Centre's Representation Information Registry are three emerging examples of potential international standards a repository might adopt. Whenever possible, the repository should use these types of standards to identify the Representation Information components of Content Information and PDI. This will reduce the long-term maintenance costs to the repository and improve quality control.

B3.4 Repository records/registers Representation Information (including formats) ingested.

When international standards for the associated Representation Information are not available, the repository needs to capture such information and register it so that it is readily findable and reusable. Some of it may be incorporated into software. It is critical to the ability to turn bits into useable information. The Representation Information must be permanently associated with the Content Information.

B3.5 Repository preserves the Content Information of AIPs.

The repository must be able to demonstrate that the AIPs faithfully reflect what was captured during ingest and that any subsequent or future planned transformations will continue to preserve that aspect of the repository's holdings.

B3.6 Repository acquires Preservation Description Information for its associated Content Information.

Preservation Description Information (PDI) is needed not only by the repository to help ensure the Content Information is not corrupted (Fixity) and is findable (Reference Information), but to help ensure the Content Information is adequately understandable by providing a historical perspective (Provenance Information) and by providing relationships to other information (Context Information). The extent of such information needs is best addressed by members of the Designated Community. The PDI must be permanently associated with Content Information.

B3.7 Repository actively monitors AIP integrity.

In OAIS terminology this means that the repository must have Fixity Information for AIPs and must make some use of it. At present, most repositories deal with this at the level of individual information objects by using a checksum of some form, such as MD5. In this case the repository must be able to demonstrate that the Fixity Information (checksums, and the information that ties them to AIPs) are stored separately or protected separately from the AIPs themselves, so that someone who manages to maliciously alter an AIP would be unlikely to be able to alter the Fixity Information as well. A combination of logs that show this check being applied and an explanation of the system security that keeps the two classes of information separate will meet this requirement.

AIP integrity also needs to be monitored at a higher level, ensuring that all AIPs that should exist actually do exist, and that the repository does not possess AIPs it is not meant to. Checksum information alone will not be able to demonstrate this.

B3.8 Repository has contemporaneous records of actions taken associated with ingest and archival storage processes and those administration processes that are relevant to the preservation.

These records must be created on or about the time of the actions they refer to. The records may be automated or may be written by individuals, depending on the nature of the actions described. Where community or international standards are used, such as PREMIS (2005), the repository must demonstrate that all relevant actions are carried through.

B3.9 Repository has mechanisms in place for monitoring and notification when Representation Information (including formats) approaches obsolescence or is no longer viable.

For most repositories the concern will be with the Representation Information (including formats) used to preserve information, which may include information on how to deal with a file format or software that can be used to render or process it. Sometimes the format needs to change because the repository can no longer deal with it. Sometimes the format is retained and the information about what software is needed to process it needs to change.

In all cases the repository must show that it has some active mechanism to warn it of impending obsolescence. The obsolescence is determined largely in terms of the knowledge base of the Designated Community. This metric ensures that the preserved information remains understandable and usable by the Designated Community. This may be dependent on an external registry, in which case the repository must demonstrate how it uses the information from that registry.

B3.10 Repository has mechanisms to change its preservation plans as a result of its monitoring activities.

Information from monitoring sometimes requires a repository to change how it deals with the material it holds in unexpected ways. The repository must demonstrate that it understands this, for example, by providing a description of how it reacts to the results of monitoring. Plans as simple as migrating from format X to format Y when the registries show that format X is no longer supported are not sufficiently flexible: other events may have made format Y a bad choice. The repository must be prepared for that eventuality. Another possible response is for the repository to create additional Representation Information and/or PDI.

B3.11 Repository can provide evidence of the success of its preservation planning.

The repository should be able to demonstrate the continued preservation, including understandability to the appropriate Designated Community, of its holdings over a number of years, given the age of the repository and its holdings.

This could be evaluated at a number of degrees of severity and depends on the specificity of the Designated Community. If the Designated Community is fairly broad then an auditor could represent the test subject in the evaluation. More specific Designated Communities could require significant efforts to verify if the auditor is not representative of the Designated Community or its knowledge base. It may be that, at an assessment, judgment must be exercised as to whether adequate efforts have been made, but such a course must be justified in detail. The same tests should apply as for C4.2, which requires that the information is understandable to the Designated Community. (See Section C, Designated Community and the Usability of Information and Appendix 1 for more about preservation and understandability.)

B4. Data management

A critical component of any repository is its data management functionality. Regardless of technical composition and regardless of whether it is considered a “light” or “dark” repository, the system still needs to be able to store and use descriptive information (metadata) for access and retrieval. Descriptive information in this sense includes many more things than the narrative description that might be familiar to the user of a traditional library or archive catalog. It also includes technical information necessary to preserve and manage the object.

In simple terms, this means that people have to be able to find what they are interested in the repository. Having found it, they need to know enough about it to be able to get a useable copy of it. That may mean they need to know how big it is, or what software they will need to interpret it. At minimum it means that a search needs to return enough information to allow them to order copies of the things that the search has found—usually some unique identifier for each object of interest, such as a catalog number or an archival reference, for instance. This requirement establishes the principle that it is not enough simply to preserve. If people cannot find what they want then the repository is not serving the needs of its users.

It is the repository’s job to ensure that each and every stored object has descriptive information associated with it. How the repository does this is not specified by this document, but the repository itself must be able to make clear how this happens. It may place the requirement entirely on the Producers of information, by having agreements with them that say that material offered to the repository must contain a minimum amount of metadata that enables the descriptive information to be stored. The repository may take on the task of producing the information itself. Or it may have a hybrid scheme that involves the repository filling in the gaps in what Producers provide—using their metadata when it is sufficient, and adding metadata when it is not. Whichever it does, it must set out in advance what the minimum metadata requirements are to enable material to be discovered and identified again.

The minimum metadata requirements for data management purposes may be very basic. In most cases the minimum requirement for discovery may be nothing more than an identifier by which the Designated Community would know and request a deposited object. For most repositories, the focus of this section is on creating and maintaining metadata that enables material to be located. (Defining “minimum metadata” that is relevant and understandable by the Designated Community is addressed in Section C2.)

In some cases the metadata that enables the material to be retrieved may warrant closer inspection. This can be important if the repository’s holdings vary greatly in size and the larger objects are not suitable for downloading over a network connection, for instance. Information about size would enable a user to choose a more optimum delivery method, such as a tape to be delivered by mail. In other cases a repository’s holdings may require special software to be available to the user to allow an object to be interpreted. Users must be able to determine this in advance, rather than possibly paying to acquire material only to discover that they do not have the tools to use it. (See section C3, Use and Usability for related use requirements.) A repository may choose to meet this requirement in more general information it makes available to its users, rather than placing specific information in the descriptive information for each AIP. For instance, a repository all of whose holdings consist of PDF files can:

- State in the information for each AIP that it is a PDF file.
- Have general information on how to use the repository that states that you will need a PDF reader to use its holdings.
- Say that its Designated Community is people with access to a PDF reader.

B4.1 Repository captures or creates minimum descriptive metadata and ensures that it is associated with the AIP.

The repository has said what metadata is needed; now it needs to say how it gets it. Does it require the Producers to do it (and so, for instance, refuse a deposit that doe not contain it) or does it know or agree that it must supply some metadata itself during the ingest process? Either is acceptable, but the responsibility must clearly fall somewhere.

Association is important. This is not a one-to-one correspondence, and is not necessarily stored with the AIP. Hierarchical schemes of description allow some descriptive elements to be associated with many items. The association should be unbreakable in the sense that it must never be lost sight of even if other associations are created.

B4.2 Repository can demonstrate that referential integrity is created between all AIPs and associated descriptive information.

Every AIP must have some descriptive information and all descriptive information must point to at least one AIP, such that the integrity can be validated. This should be an easy requirement to satisfy and is a prerequisite for the next one.

B4.3 Repository can demonstrate that referential integrity is maintained between all AIPs and associated descriptive information.

Particular attention must be paid to operations that affect AIPs and their identifiers over time and how integrity is maintained during these operations. There may be times, depending on system design, where this cannot be demonstrated because some system component is out of action, but it must be possible to know when this occurs and the demonstration must happen.

B5. Access management

These requirements establish that access is implemented according to the repository’s stated policies. They fall into three groups. B5.1 and B5.2 are primarily concerned with security of access—who can access what. B5.3 to B5.5, taken together, ensure that the access function is implemented correctly. Access should always deliver what is required, or else make clear that it is not possible for whatever reason, and it should do so in a timely fashion. Timeliness may be measured in seconds or weeks, since access may be an online function or a postal function or may be mediated through some other mechanism or a combination of them.

The final requirement, B5.6, stands alone. It makes a specific, additional requirement over and above the need to simply provide access to a repository's holdings. For the repository to be trusted, it must be able to provide a copy of material that can be traced back to originals.

B5.1 Repository access management system fully implements access policy.

All the policies should be seen to be implemented. This may be partly by computers and partly by humans, as will be the case with some forms of access validation—checking passports, for instance, before issuing a userid and password may be an appropriate part of access management for some institutions.

B5.2 Repository logs all access management failures, and staff review inappropriate "access denial" incidents.

A repository should have some mechanism to filter out anomalous or unusual denials and use them to either identify security threats or failures in the access management system (such as valid users being denied access). This does not mean looking at every denied access.

B5.3 Repository can demonstrate that the process that generates the DIP is completed in relation to the request.

If a user expects a set, the user should get the whole set. If the user expects a file, the user should get the whole file. If the user's request cannot be satisfied, the user should get told this. (For instance, resource shortages may mean a valid request cannot be satisfied.)

Acceptable scenarios include:

- The user receives the complete DIP asked for and it is clear to the user that this has happened.
- The user is told that the request cannot be satisfied.
- Part of the request cannot be satisfied, the user receives a DIP containing the elements that can be provided, and the system makes clear that the request is only partially satisfied.

Unacceptable scenarios include:

- The request can only be partially satisfied and a partial DIP is generated, but the response delivered to the user does not indicate that it is partial.
- The request is delayed indefinitely because something it requires. such as access to a particular AIP, is not available, but no notification is given to the user nor is there any indication as to when the conflict will be resolved, if ever.
- The user is told the request cannot be satisfied, implying nothing can be delivered, but actually receives a DIP, and is now unsure of its validity or completeness.

B5.4 Repository can demonstrate that the process that generates the DIP is correct in relation to the request.

The right material should be delivered and appropriate transformations should be applied (if necessary) to generate the DIP. A simple example is that if the repository stores TIFF images but delivers JPEGS, the conversion should be shown to be correct to whatever standards seem appropriate. If the repository offers delivery as JPEG or PNG, then the user should receive the format requested. Many repositories may apply more complex transformations to generate DIPs from AIPs.

B5.5 Repository demonstrates that all access requests result in a response of acceptance or rejection.

Eventually a request must succeed or fail, and there must be boundaries on how long it takes for the user to know this. The repository must record some information about access requests, even if it does not retain the information for long.

B5.6 Repository enables the dissemination of authentic copies of the original or objects traceable to originals.

Part of trusted archival management deals with the authenticity of the objects that are disseminated. A repository must enable end users either to be confident that what they have is an authentic copy of the original object, or is traceable in some auditable way to the original object. This distinction is made because objects are not always disseminated in the same way, or in the same groupings, as they are deposited. A database may have subsets of its rows, columns, and tables disseminated so that the phrase "authentic copy" has little meaning. Ingest and preservation actions may change the formats of files, or may group and split the original objects deposited. The requirement seeks to ensure that these actions do not lose information that would support an auditable trail between the original deposited object and the eventual disseminated object.

A repository should be able to demonstrate the processes to construct the DIP from the relevant AIP(s). This is a key part of establishing that DIPs reflect the content of AIPs, and hence of original material, in a trustworthy and consistent fashion. DIPs may simply be a copy of AIPs, or may result from a simple format transformation of an AIP. But in other cases they may be derived in complex ways from a large set of DIPs. A user may request a DIP consisting of the title pages from all e-books published in a given period, for instance, which will require these to be extracted from many different AIPs. If the repository allows complex DIPs of this nature to be requested, it will need to put more effort into demonstrating how this metric is met compared with a repository that only allows requests for DIPs that correspond to an entire AIP.

A repository is not required to show that every DIP it provides can be verified as an authentic copy of the original; it must show that it can do this when it is required.

The distinction between authentic copies of the original and objects traceable to originals is made here because some types of object are rarely disseminated as a whole. When they are not, copies cannot be spoken of as simply being authentic copies, but rather need to be able to show a chain of trust from the original object(s) to the disseminated fragment.

The distinction can also be important when transformation processes are applied. For instance, consider a repository that stores digital audio from radio broadcasts, but enables the dissemination of derived text that is constructed by automated voice recognition from the digital audio stream. This is likely to be an imperfect process but may still be worthwhile for many of the purposes of its users. But no one would attest that these texts were authentic copies of the original audio. Producing an authentic copy means either handing out the original audio stream or getting a human to verify and correct the transcript against the stored audio.

The level of authentication is to be determined by the Designated Community, and the requirement is just to enable high levels of authentication, not to impose it on all copies.

C. The Designated Community & the Usability of Information

The OAIS Reference Model specifies that a Designated Community is “an identified group of potential Consumers who should be able to understand a particular set of information. The Designated Community may be composed of multiple user communities.” The Designated Community may be associated with a class of objects within one or more Repositories, rather than a single Designated Community for each Repository. It’s important to note that a Designated Community will likely include both information Consumers and information Providers and the repository’s assumptions will take their interests and needs into account.

In order to adequately provide digital preservation services, the repository must state its assumptions about the intended use of the *information objects* (i.e., Content Information and PDI) it will hold and preserve. The assumptions provide the foundation of the scope of services required to satisfy the information needs of the users of the collections in the repository. Without this foundation the repository cannot state the boundaries of its expected capabilities. Fundamentally, the Content Information and PDI are to be independently understandable to an appropriate Designated Community. The purpose is to avoid needing experts who are intimately familiar with the Content Information or PDI on hand to assist users (Consumers) because this is expensive and limits how long the information can be retained without loss.

The Designated Community can be defined and communicated through a variety of mechanisms. Perhaps the most straightforward way is through marketing materials used for courting depositors and users. However, the assumptions about the nature of the Designated Community can be intimated in Producer agreements, user agreements, Web sites, and policy documents.

A step toward public presentation of this information will likely be found in system requirements documents, use cases, and management policies and procedures. However, without public availability of this information trust in the repository is jeopardized.

C1. Documentation

C1.1 Repository has a definition of its Designated Community/ies—who it is, what its knowledge base is, what levels of service it expects, etc.

Examples of Designated Community definitions include:

- General English-reading public educated to high school and above, with access to a Web Browser (HTML 4.0 capable).
- For GIS data: GIS researchers—undergraduates and above—having an understanding of the concepts of Geographic data and having access to current (2005, USA) GIS tools/computer software, e.g., ArcInfo (2005).
- Astronomer (undergraduate and above) with access to FITS software such as FITSIO, familiar with astronomical spectrographic instruments.
- Student of Middle English with an understanding of TEI encoding and access to an XML rendering environment.
    - Variant 1: Cannot understand TEI

    - Variant 2: Cannot understand TEI and no access to XML rendering environment
    - Variant 3: No understanding of Middle English but does understand TEI and XML
- Two groups: the publishers of scholarly journals and their readers, each of whom have different rights to access material and different services offered to them.

The Designated Community definition is arrived at through the planning processes used in creating the repository and defining its services. The definition will be drawn from various sources ranging from market research, to service level agreements for Producers, to the mission or scope of the institution within which the repository is embedded.

C1.2 Repository makes the definition of its Designated Communities available.

A public statement of the intended users, the Designated Community, of the repository creates a touch-point for Consumers to evaluate the service expectations in a broad sense. This information may be communicated through a variety of means including display on a public Web site, as part of the charter of the repository, in dataset descriptions, in printed publicity materials, through informational material for potential depositors, through usage agreements with Consumers, etc.

C1.3 Repository defines, communicates, and commits to a definition of "understandability" with its Designated Community.

For a given submission of information, the repository must make clear the operational definition of understandability that is associated with the corresponding Designated Community. The Designated Community may vary from one submission to another, as may the definition of understandability that establishes the repository's responsibility in this area. This may range from no responsibility for the case of bit preservation only, to the maintenance of a particular level of use for the case where understanding by the members of the Designated Community is determined outside the repository, to the case where the repository has a responsibility to ensure a given level of Designated Community human understanding and must ensure the required Representation Information to meet this understanding level has been obtained.

It must also include a definition of the Designated Community's application tools that are to use the information (possibly after transformation by repository services). For example, if the Designated Community is defined as readers of English with access to widely available document rendering tools, and if this definition is clearly associated with a given set of Content Information and PDI, then the requirement is met.

C2. Descriptive metadata appropriate to the Designated Community

A repository's minimum descriptive metadata requirements must match the needs of the repository's Designated Community. This does not mean being capable of responding to every request for additional catalog information that comes from its users. Rather, it must make an assessment, based on utility and cost, of what a representative member of its Designated Community would want. If the repository serves multiple communities, each of which are interested in different segments of its holdings, then it is acceptable that the minimum requirements may vary from AIP to AIP. It is natural to expect that if a repository holds both digital films and digital music that the minimum descriptive elements required for film and music would differ.

For instance, a repository may hold and preserve a model demonstrating a point made in a book. End users would be expected to "discover" the work through the text of the book, and the only discovery requirement for the repository would be to respond to a request for the published identifier of the model. A slightly more expansive requirement might be if the repository holds digital books, it may decide that a title is the minimum information required. This is permissible even if it is capable of holding much richer sets of metadata per book. It is also permissible even if some of its users would like to be able to search on more exotic criteria, such as the original publication price, the number of pages, or the fonts used in the text. If the repository can make a good case that the minimum requirements it has established are sufficient for the majority of expected uses, it satisfies this requirement.

Focusing on the Designated Community ensures that the repository is not faced with the impossible task of satisfying all possible users. A scientific data repository may have established that its Designated Community consists of English-speaking physicists educated to at least doctorate level. It is not then a failing if a schoolchild has difficulty in locating material, nor if the same problems face a professor of biology or a Russian-speaking postdoctoral physicist. This requirement can also be applied to repositories that expect their holdings to be searched by other computer programs rather than directly by humans. The repository can establish what types of software can interrogate its holdings, perhaps by specifying an information exchange protocol that they must support.

C2.1 Repository articulates minimum metadata requirements to enable the Designated Community to discover and identify material of interest.

There is a distinction here between retrieval metadata and other metadata that describes what has been found. For example, in a library we might say that a book's title is mandatory, but its publisher is not, because people generally search on the title.

This is not about being able to satisfy every possible request, but about dealing with the types of request that a typical user from the Designated Community would want. The minimum requirements must be articulated. Note that the minimum may be nothing more than an identifier by which the Designated Community would know and request a deposited object.

C3. Use & usability

It is important to realize that repositories are not required to have rigid service levels to which they must always conform. Guarantees of this sort are appropriate to some repositories. They may take a form like, "We will always ship your order within 48 hours, or we will inform you by e-mail if that is not possible." If such guarantees are provided it is important that the repository can demonstrate that they are adhered to. But a small repository may not be able to provide them. It would be acceptable then for it to say that order processing will depend on staff availability and demand, but that you can always check the status of the order via a Web page, e-mail, or a telephone call, for instance.

Meeting these requirements is not simply about providing access to everyone. A repository's Consumers (including the Designated Community/ies) may be a small set of people, and confidentiality requirements or Producer agreements may mean that different members of it may only be entitled to access highly restricted subsets of the repository's holdings. It is important for the repository to demonstrate that it applies these restrictions properly. One of the dimensions of trust is the trust that information Producers have in the repository. Where information Producers place requirements on the repository to permit only specific forms of access or use to specific, identified communities, they must have confidence that the repository will implement these restrictions in a suitably secure manner.

Not all repositories will have restrictions on access. At the other extreme, some repositories may hold information where the presumption is that access is not permitted to anyone (even the Producers) without a court order. This would be appropriate for highly confidential information that is being preserved for future access, perhaps many years hence. In some cases repositories may not even permit the public to know that they hold particular items.

All such ways of working are acceptable provided that the repository makes its policies clear, and can be seen to adhere to them.

C3.1 Repository documents and communicates to its Designated Community what access and delivery options are available.

Repository policies should document the various aspects of access to and delivery of the preserved information. There is a general expectation that the policies, or at least the consequences of them, are made known to the Designated Community. The users should know what they can ask for, when, how, and whether it will cost (among other things).

The repository must make clear what users can and cannot do, and that the things they can do meet their reasonable needs. Can users search a catalog via the Web? Can they visit the repository to speak to someone to help them find information? Can they download copies instantly or must they be ordered for postal delivery? Can they request subsets of AIPs, or multiple AIPs in a single request? Can they choose different file formats for delivery? What types of searches can they perform?

Repository policies should clearly define access and delivery mechanisms available to its Designated Communities. Repositories do not have to support any particular type of request; they just need to state which types of request they can handle (online, batch, on-site, incidental, programmed or repeated requests—either to be notified when new material of a given type appears, or automatically receive copies of certain types of material). Similarly, a repository does not have to support any particular kind of delivery mechanism, but it does need to describe and communicate what types of delivery it can perform. Are the types of delivery defined and announced (digital files, sent for example as an e-mail attachment, by Web, by FTP, or sent by mail on disk, tape, in print)? Are there limits on the types or the size of the result sets?

In the case where there are charges associated with using the digital objects within the repository, repository policies should clearly define what charges it applies to services for its Designated Communities. Not all repositories will charge; some will only charge for certain services; some may have annual subscription fees with unlimited usage and others may charge per item or even per search. In some repositories, charges will be calculated automatically by an online ordering system whereas in others the charge for delivering an item may not be known until the item is produced. The latter may be the case where substantial manual work is required to produce a DIP and the work is charged by the hour, for instance. Any and all of these policies are acceptable so long as the charging mechanism and the services to which it applies are made known to those who might have to pay the charges. The repository should be able to show that the charging mechanisms are applied consistently.

Note that repositories might have to deal with a single, homogeneous, or with multiple or disparate communities. Different policies might be needed for different communities as well as for different collection types.

C3.2 Repository has implemented a policy for recording all access actions (includes requests, orders etc.) that meet the requirements of the repository and information Producers/depositors.

A repository need only record the actions that meet the requirements of the repository and its information Producers/depositors. This may mean that little or no information is recorded about access. That is acceptable if the repository can demonstrate that it does not need to do more. Some repositories may want information about *what* is being accessed, but not about *who* is doing it. Others may need much more detailed information about access. A policy should be established and implemented that relates to demonstrable needs. Are these figures being monitored? Are statistics produced and made available?

C3.3 Repository ensures that agreements applicable to access conditions are adhered to.

The repository must be able to show what Producer/depositor agreements (if any) apply to which AIPs and must validate user identities in order to ensure that the agreements are satisfied. Although it is easy to focus on denying access when considering conditions of this kind (that is, preventing unauthorized people from seeing material), it is just as important to show that access is granted when the conditions say it should be (that is, people who should be permitted to do things are actually permitted to do them).

Access conditions are often just about who is allowed to see things, but they can be more complex. They may involve limits on quantities—all members of community A are permitted to access 10 items a year without charge, for instance. Or they may involve limits on usage or type of access—some items may be viewed but not saved for later reuse, or items may only be used for private research but not commercial gain, for instance.

Various scenarios may help illustrate what is required:

- If a repository's material is all open access, the repository can simply demonstrate that access is truly available to everyone.
- If all material in the repository is available to a single, closed community, then the repository must demonstrate that it validates that users are members of this community, perhaps by requesting some proof of identity before registering them, or just by restricting access by network addresses if the community can identified in that manner. It should also demonstrate that all members of the community can indeed gain access if they wish.
- If different access conditions apply to different AIPs, then the repository must demonstrate how these are realized.
- If access conditions require users to make some declaration before receiving DIPs, then the repository must be able to provide evidence that the declarations have been made. These might be signed forms, or evidence that a statement has been viewed online and a button clicked to signify agreement. The declarations might involve nondisclosure or agreement to no commercial use, for instance.

C3.4 Repository has documented and implemented access policies (authorization rules, authentication requirements) consistent with deposit agreements for stored objects.

User credentials are only likely to be relevant for repositories that serve specific communities or that have access restrictions on some of their holdings. A user credential may be as simple as the IP address from which a request originates, or may be a username and password, or may be some more complex and secure mechanism. Thus, this may be a null requirement for some repositories and require very formal validation for others. The key thing is that the access and delivery policies are reflected in practice and that the level of validation is appropriate to the risks of getting validation wrong. Some of the requirements may emerge from agreements with Producers/depositors and some from legal requirements.

Repository staff will also have occasional need to access stored objects, whether to complete ingest functions, perform maintenance functions such as verification and migration, or to produce DIPs. The repository must have policies and mechanisms to protect stored objects against deliberate or accidental damage by staff (see A5.1).

C4. Verifying understandability

The primary purpose of most repositories is to preserve information so that it can be usable over time. The potential users of the digital resources, the Designated Community, will have specific expectations of usability and understandability.

It is the repository's responsibility to make sure it has mechanisms in place to ensure it obtains, manages, and makes available the Content Information and the PDI in forms that allow the digital objects to be understandable and usable.

C4.1 Repository has a documented process to test understandability to the Designated Community, as previously defined, of the information content associated with the Content Information and PDI, and this includes defining needed steps should the agreed level of understandability not be met.

It may be that the Content Information or PDI is held by the repository in a form that is not directly usable by current Designated Community application tools. In such a case the repository needs to have a defined process for the transformation to a usable form, or how additional Representation Information is made available (see B3.9).

Repositories that share the burden of ensuring that adequate metadata (documentation) are captured or generated to meet a required degree of Designated Community human understanding may implement any number of procedures to address this requirement. Such repositories typically have a narrowly defined Designated Community, such as a particular science discipline. Examples of approaches to meeting this requirement include the retention of individuals with the discipline expertise, or the periodic assembly of outside community members, to evaluate and identify additional metadata needed for human understanding.

C4.2 Repository has verified that Content Information and PDI are understandable to Designated Community.

The repository needs to verify that Content Information and PDI held as AIPs can be made available in a form that is utilizable by typical Designated Community tools. One way to do this by selecting subsets of the AIPs, extracting the Content Information, running it through any documented transformation processes, and then using it with typical Designated Community tools. The result should match with that expected, and should be recorded for comparisons over time to ensure long-term preservation.

Repositories that also participate in ensuring that the Content Information and PDI are sufficiently humanly-understandable to the Designated Community need to document the execution of their process for checking and ensuring this level of understanding. One way to do this is for a Designated Community proxy-reviewer to sign off on the understandability requirement. Another is to document the assembly of a review team and the results of that review, including any efforts made to bring the metadata (documentation) to an adequate level of understandability.

D. Technologies & Technical Infrastructure

This section is about the system, technologies, and technical infrastructure required to ensure AIPs can be preserved for the long term, once they are ingested. It does not prescribe specific hardware and software, but describes best practices for data management and security. Criteria here are similar to the good computing practices required in international management standards like ISO 17799. Repositories or organizations that have undergone ISO 17799 certification are very likely to meet all these criteria.

This section is broken into three layers. This first layer stresses general system infrastructure requirements. The second layer addresses appropriate technologies, building on the system infrastructure requirements, with additional criteria specifying the use technologies and strategies appropriate to the repository's Designated Community. The final layer addresses security. "System" in this section, as in many others, refers to more than IT systems, such as servers, firewalls, or routers. Fire protection systems and flood detection are significant, as are systems that involve actions by people.

D1. System infrastructure

Without a secure and trusted infrastructure, the functions carried out on AIPs cannot be trusted—they are built on a house of cards. Actions specified here are general enough to apply to systems other than repositories and archives.

D1.1 Repository functions on well-supported operating systems and other core infrastructural software.

The metric specifies well-supported as opposed to manufacturer-supported or other similar phrases. The level of support for these elements of the infrastructure must be appropriate to their uses; the repository must understand where the risks lie. The degree of support required relates to the criticality of the subsystem involved. A repository may deliberately have an old system using out-of-date software to support some aspects of its ingest function. If this system fails it may take some time to replace it, if it can be replaced at all. As long as its failure does not affect mission-critical functions, this is acceptable. Systems used for internal development may not be protected or supported to the same level as those for end-user service.

D1.2 Repository ensures that all platforms have a backup function sufficient for the repository's services and for the data held, e.g., metadata associated with access controls, repository main content, etc.

The repository needs to be able to justify its backup systems. Some will need much more elaborate backup plans than others.

D1.3 Repository stipulates the number and location of copies of all digital objects.

The repository must identify the number of copies of all stored digital objects, and the location of each object and their copies. This applies to what are intended to be identical copies, not versions of objects or copies.

The location must be described such that the object can be exactly located without ambiguity. It can be an absolute physical location or a logical location within a storage media or a storage subsystem. One way to test this would be to look at a particular object and ask how many copies there are, what they are stored on, and where they are.

A repository can have different policies for different classes of objects, depending on factors such as the Producer, the information type, or its value. Some repositories may have only one copy of everything, stored in one place though this is definitely not recommended. There may be additional identification requirements if the data integrity mechanisms use alternative copies to replace failed copies.

D1.4 Repository has mechanisms in place to insure any/multiple copies of digital objects are synchronized.

If multiple copies exist, there has to be some way to ensure that changes to an object are propagated to all copies of the object. There must be an element of timeliness to this. It must be possible to know when the synchronization has completed, and ideally to have some estimate beforehand as to how long it will take to complete. Depending whether it is automated or requires manual action (such as the retrieval of copies from off-site storage) the time involved may be seconds or weeks. The duration itself is immaterial—what is important is that there is understanding of how long it will take.

There must also be something that addresses what happens while the synchronization is in progress. This has an impact on disaster recovery: what happens if a disaster happens while an update is in progress? If one copy of an object is altered and a disaster occurs whilst other copies are being updated, it is essential to be able to ensure later that the update is successfully propagated.

D1.5 Repository has effective mechanisms to detect data corruption or loss.

The repository must detect data loss accurately to ensure that any losses fall within the tolerances established by policy (see A.3.6). Data losses must be detected and detectable regardless of the source of the loss. This applies to all forms and scope of data corruption, including detecting missing objects and false objects, corruption within an object, and copying errors during data migration or synchronization of copies. Ideally, the repository will demonstrate that it has all the AIPs it is supposed to have and no others, and that they and their metadata are uncorrupted.

The approach must be documented and justified. Common hazards, such as hardware failure, human error, and malicious action, should be mitigated. Repositories that use well-recognized mechanisms such as MD5 signatures need only recognize their effectiveness and role within the overall approach. But to the extent the repository relies on homegrown schemes they must provide convincing justification that data loss and corruption are detected within the tolerances established by policy.

Data losses must be detected promptly enough that routine systemic sources of failure, such as hardware failures, are unlikely to accumulate and cause data loss beyond the tolerances established by policy. For example, consider a repository that maintains a collection on identical primary and backup copies with no other data redundancy mechanism. If the media of the two copies have a measured failure rate of 1% per year and failures are independent, then there is a 0.01% chance that both copies will fail in the same year. If the policy were the repository could not lose more than 0.001% of the collection per year, then the repository would need to confirm media integrity at least every 72 days to achieve an average time-to-recover of 36 days or about one tenth of a year. This simplified example illustrates the kind of issues a repository should consider, but the objective is a comprehensive treatment of the sources of data loss and their real world complexity.

D1.6 Repository reports to its administration all incidents of data corruption or loss, and steps taken to repair/replace corrupt or lost data.

The repository must record, report, and repair as possible all violations of data integrity. A record of incidents, recovery actions, and their results should be available. The repository should document procedures to take when loss or corruption is detected, including standards for measuring the success of recoveries.

D1.7 Repository has defined processes for storage media migration.

The repository should have should be triggers for initiating action and understanding of how long it will take for storage media migration, or “refreshing”—copying between media without reformatting the bitstream. Will it finish before the media is dead, for instance? Copying large quantities of data can take a long time and can have impact on other system performance.

It is important that the process incorporates a check that the copying has happened correctly. (See B3.2.)

D1.8 Repository has a documented change management process that identifies changes to critical processes.

Examples of this would include changes in processes in Data Management, Access, Archival Storage, Ingest, Security. The really important thing is to be able to know what changes were made and when they were made. Traceability makes it possible to understand what was affected by particular instances of systems.

D1.9 Repository has a process for testing the effect of critical changes to the system.

This could cover many different things, from whole-system testing to unit testing. It could be very expensive and complex safety-type tests are not required. But there should be some recognition of the fact that a completely open regime where no changes are ever evaluated or tested is a recipe for problems. There are other ways of dealing with this problem. One is to ensure that many operations are reversible, so that if a component is later discovered to have failed, the change can be undone.

D1.10 Repository has a process to stay current with the latest operating system security fixes

The repository must show evidence of how it is operated; automated updates and manual review by system staff are all acceptable. Database applications, Web servers, etc., are all significant along with operating systems.

D2. Appropriate technologies

A repository should use strategies and standards relevant to its Designated Communities and its digital technologies.

D2.1 Repository has hardware technologies appropriate to the services it provides to its Designated Communities and has procedures in place to monitor and receive notifications when hardware technology changes are needed.

The repository needs to be aware of the types of access services its Designated Community expects, including where applicable the types of media to be delivered, and needs to make sure its hardware capabilities can support these services. For example, it may need to improve its networking bandwidth over time to meet growing access data volumes and expectations.

D2.2 Repository has software technologies appropriate to the services it provides to its Designated Communities and has procedures in place to monitor and receive notifications when software technology changes are needed.

The repository needs to be aware of the types of access services the Designated Community expects, and to make sure its software capabilities can support these services. For example, it may need to add format translations to meet the needs of currently widely used application tools, or it may need to add a data subsetting service for very large data objects.

D2.3 Repository has procedures in place for monitoring or receiving notifications about changes in the needs of its Designated Communities (e.g., surveys, formal reviews, workshops, and individual interactions).

The repository may use various mechanisms to maintain awareness of its Designated Community needs. These mechanisms need to be incorporated into its operating procedures.

D3. Security

“System” here refers to more than IT systems, such as servers, firewalls, or routers. Fire protection and flood detection systems are as significant, as are systems that involve actions by people. The first two of these requirements are general and the third addresses internal security, while the remainder address disaster recovery.

D3.1 Repository maintains a systematic analysis of its environment: data, systems, personnel, physical plant, security needs, etc.

This can mean everything from temperature and humidity requirements to building security to staff vetting.

D3.2 Repository has implemented mechanisms (processes) to adequately address each of the defined security needs.

The repository must show how it has dealt with its security requirements. If it knows some types of material are likely to be subject to high levels of attack, it will need to provide additional levels of protection, for instance.

D3.3 Repository staff have delineated roles, responsibilities, and authorizations.

Whose job involves repository functions? What are the responsibilities of staff? Is the management's picture of this the same as the staff's? Authorizations are about who can do what—who can add users, who has access to change metadata, who can get at audit logs.

D3.4 Repository has written disaster preparedness and recovery plan(s), including at least one off-site copy of all deposited data.

The repository must have a written plan with some approval process for what happens in specific types of disaster (fire, flood, system compromise, etc.) and for who has responsibility for actions. Multiple off-sites copies are expected of most repositories, but others may be able to justify not providing these. The level of detail in a disaster plan, and the specific risks it addresses, needs to be appropriate to the repository's location and service expectations. Fire is an almost universal concern, but earthquakes probably do not require specific planning all locations. The disaster plan must, however, deal with unspecified situations that can be foreseen to have specific consequences, such as lack of access to a building.

D3.5 Repository tests disaster plans regularly.

There needs to be evidence that the plan is tested. Among the things to be tested are that the relevant personnel are aware of their roles in any disaster. Staff should be able to answer questions like these in compliance with the disaster plan: "What would you do if the fire alarm went off?" "Do you take decisions during a disaster or follow someone else's instructions? What if that someone else is absent on that day?"

Experience has shown that untested plans do not work. Tests can be walkthroughs of people's roles or actual system shutdowns. Walkthroughs are more likely to be feasible in many repositories, as the loss of service involved in a more rigorous test may not be acceptable.

Tests do not need to show perfect results, but there should be a way to learn from mistakes. Each test of a plan will usually identify gaps that require corrective action. Also, plans need to evolve over time as the system infrastructure and the external environment change. The repository should have a documented rationale for how frequently the disaster plan is tested.

D3.6 Repository has defined processes for service continuity and disaster recovery.

The needs here, and hence the processes, will depend to a great extent on the Designated Community, their requirements, and the funding available. Service continuity can be anything from no outage to many weeks outage for some types of disaster. The best response to some extreme events is a mechanism to hand over responsibility for the repository's holdings to some other organization(s), a process that may take many months or even years. The repository is simply required to show that some process or agency will ensure that the AIPs are preserved and that access to them is restored eventually. But no mechanism is perfect and a repository can acknowledge that it cannot afford to defend its holdings against some risks.

III. Audit Instrument for the Certification of Trusted Digital Repositories

Institution/Organization: ____

Form Completed By: ____

Date: ____

Stage of Development: ____

Description of Digital Repository:

[This page intentionally left blank.]

	Planned?	Documented?	Implemented?	Evaluated?	Notes
A. The Organization					
A1. Governance & organizational viability					
A1.1. Repository has a mission statement that reflects a commitment to the long-term retention of, management of, and access to digital information on behalf of depositors.					
A1.2. Repository has a formal succession plan, contingency plans, and/or escrow arrangements in place in case repository ceases to operate or substantially changes its scope (i.e., return with adequate prior notification of digital objects to depositors and/or trusted inheritors identified).					
A2. Organizational structure & staffing					
A2.1. Repository staff have skills and expertise appropriate to their duties.					
A2.2. Repository has appropriate number of staff to support all functions and services designated in agreements with depositors.					
A2.3. Repository commits to professional development to keep staff expertise and skills current.					
A3. Procedural accountability & policy framework					
A3.1. Repository has a mechanism in place for reviewing, updating, and developing comprehensive policies and procedures as repositories grow and as the community practice evolves.					
A3.2. Repository has monitoring and feedback mechanisms in place to ensure continued operation, support problem resolution, and address evolving requirements of providers and consumers.					
A3.3. Repository is committed to formal, periodic review and assessment to ensure continued development.					
A3.4. Repository has a documented history of the changes to its operations, procedures, software, hardware, traceable to its preservation strategies where appropriate.					
A3.5. Repository commits to transparency and accountability in all actions supporting the operation and management of the repository.					

	Planned?	**Documented?**	**Implemented?**	**Evaluated?**	**Notes**
A3.6. Repository commits to define, collect, track, and provide on demand, its information integrity measurements.					
A3.7. Repository commits to a regular schedule of certification and to notifying certifying bodies of operational changes that will change or nullify its certification status.					
A4. Financial sustainability					
A4.1. Repository has a short- and long-term business planning process in place to support sustainability.					
A4.2. Repository has in place at least annual processes to review and adjust business plans as necessary.					
A4.3. Repository business planning and practices are transparent, compliant with relevant accounting standards and practices, and auditable.					
A4.4. Repository has ongoing commitment to risk, benefit, investment, and expenditure analysis and reporting (including assets, licenses, and liabilities).					
A4.5. Repository recognizes the eventual strong possibility of a gap between repository-generated funding and the funding necessary to meet the repository's commitments to its depositors. It commits to bridging these gaps by securing funding and resource commitments specifically for that purpose; these commitments can come either from the repository itself or parent organizations, as applicable.					

	Planned?	Documented?	Implemented?	Evaluated?	Notes
A5. Contracts, Licenses and Liabilities					
A5.1 If repository manages, preserves, and/or provides access to digital materials on behalf of another organization, it has and maintains appropriate contracts or deposit agreements.					
A5.2 Repository contracts or deposit agreements must specify and/or transfer appropriate preservation rights, as necessary.					
A5.3 Repository tracks and manages copyrights and restrictions on use as required by contract or license					
A5.4 If repository ingests digital content with unclear ownership/rights, it has policies addressing liability and challenges to those rights.					

	Planned?	Documented?	Implemented?	Evaluated?	Notes
B. Repository Functions, Processes & Procedures					
B1. Ingest/acquisition of content					
B1.1. Repository identifies properties it will preserve for each class of digital object.					
B1.2. Repository has specified all appropriate aspects of acquisition, maintenance, access, and withdrawal issues in written agreements with depositors.					
B1.3. Repository has an identifiable, written definition for each SIP or class of information ingested by the repository.					
B1.4. Repository has a process to ensure that the information is acquired from the expected source.					
B1.5. Repository obtains sufficient physical control over the digital objects to preserve them.					
B1.6. Repository's ingest process verifies each SIP for completeness and correctness.					
B1.7. Repository provides producer/depositor with appropriate responses at predefined points during the ingest processes.					
B1.8. Repository can demonstrate that all SIPs are either accepted as whole or part of an eventual AIP, or otherwise disposed of in a recorded fashion.					
B1.9. Repository can demonstrate when preservation responsibility is formally accepted for the contents of the AIP.					

	Planned?	Documented?	Implemented?	Evaluated?	Notes
B2. Archival storage: management of archived information					
B2.1. Repository has an identifiable, written definition for each AIP or class of information preserved by the repository.					
B2.2. Repository has a definition of each AIP (or class) that is adequate to fit long-term preservation needs.					
B2.3. Repository has a definition of how AIPs are derived from SIPs.					
B2.4. Repository has and uses a naming convention that can be shown to generate visible,unique IDs for all AIPs.					
B2.5. If unique identifiers are associated with SIPS before ingest, they are preserved in a way that maintains a persistent association with the resultant AIP.					
B2.6. Repository verifies each AIP for completeness and correctness when generated.					
B2.7. Repository provides an independent mechanism for audit of the integrity of the repository collection/content.					

	Planned?	Documented?	Implemented?	Evaluated?	Notes
B3. Preservation planning, migration, & other strategies					
B3.1. Repository has documented preservation strategies.					
B3.2. Repository implements/responds to strategies for AIP storage and migration.					
B3.3 Repository uses appropriate international representation information [including format] registries					
B3.4. Repository records/registers representation information [including formats] ingested					
B3.5. Repository preserves the content information of AIPs.					
B3.6 Repository acquires Preservation Description Information for its associated content information.					
B3.7. Repository actively monitors AIP integrity.					
B3.8. Repository has contemporaneous records of actions taken associated with ingest and archival storage processes and those administration processes which are relevant to the preservation.					
B3.9. Repository has mechanisms in place for monitoring and notification when format (or other representation information) obsolescence is near/or are no longer viable.					
B3.10. Repository has mechanisms to change its preservation plans as a result of its monitoring activities.					
B3.11. Repository can provide evidence of the success of its preservation planning					

	Planned?	Documented?	Implemented?	Evaluated?	Notes
B4 Data Management					
B4.1. Repository captures or creates this minimum descriptive metadata and ensures it is associated with the AIP					
B4.2. Repository can demonstrate that referential integrity is created between all AIPs and associated descriptive information.					
B4.3. Repository can demonstrate that referential integrity is maintained between all AIPs and associated descriptive information.					
B.5 Access Management					
B5.1. Repository access management system fully implements access policy					
B52. Repository logs all access management failures, and staff review inappropriate "access denial" incidents.					
B5.3. Repository can demonstrate that the process that generates the DIP is complete in relation to the request.					
B5.4. Repository can demonstrate that the process that generates the DIP is correct in relation to the request.					
B5.5. Repository must demonstrate that all access requests result in a response of acceptance or rejection.					
B5.6. Repository enables the dissemination of authentic copies of the original or objects traceable to originals					

	Planned?	Documented?	Implemented?	Evaluated?	Notes
C. Designated Community and the Usability of Information					
C1. Documentation					
C1.1. Repository has a documented definition of its designated community/ies--who it consists of, its knowledge base, what levels of service it expects, etc.					
C1.2. Repository makes the definition of its Designated Communities available.					
C1.3. Repository defines, communicates, and commits to a definition of "understandability" with its Designated Community.					
C2. Descriptive Metadata Appropriate to Designated Community					
C2.1. Repository articulates minimum metadata requirements to enable the Designated Community to discover and identify material of interest.					
C3. Use and Usability					
C3.1. Repository documents and communicates to its designated community what access and delivery options are available.					
C3.2. Repository has implemented a policy for recording all access actions (includes requests, orders etc.) that meet the requirements of the repository and information producers/depositors.					
C3.3. Repository ensures that agreements applicable to access conditions are adhered to.					
C3.4. Repository has documented and implemented access policies (authorization rules, authentication requirements) consistent with deposit agreements for stored objects.					

	Planned?	Documented?	Implemented?	Evaluated?	Notes
C4. Verifying Understandability					
C4.1. Repository has a documented process to test ‘understandability to the Designated Community’, as previously defined, of the information content associated with the Content Information and PDI, and this includes defining the appropriate steps necessary should the agreed level of ‘understandability’ not be met.					
C4.2. Repository has verified that Content Information and PDI are understandable to Designated Community.					

	Planned?	Documented?	Implemented?	Evaluated?	Notes
D. Technologies & Technical Infrastructure					
D1. System infrastructure					
D1.1. Repository functions on well-supported operating systems and other core infrastructural software.					
D1.2. Repository ensures that all platforms have a backup function, sufficient for the repository's services and for the data held (e.g., metadata associated with access controls, repository main content, etc.)					
D1.3. Repository stipulates the number and location of copies of all digital objects.					
D1.4. Repository has mechanisms in place to insure any/multiple copies of digital objects are synchronized.					
D1.5. Repository has effective mechanisms to detect data corruption or loss.					
D1.6. Repository reports to its administration all incidents of data corruption or loss, and steps taken to repair/replace corrupt or lost data.					
D1.7. Repository has defined processes for storage media migration.					
D1.8. Repository has a documented change management process that identifies changes to critical processes.					
D1.9. Repository has a process for testing the effect of critical changes to the system.					
D1.10. Repository has a process to stay current with the latest operating system security fixes.					

	Planned?	Documented?	Implemented?	Evaluated?	Notes
D2. Appropriate technologies					
D2.1. Repository has hardware technologies appropriate to the services it provides to its designated communities and has procedures in place to monitor and receive notifications when hardware technology changes are needed.					
D2.2. Repository has software technologies appropriate to the services it provides to its designated communities and has procedures in place to monitor and receive notifications when software technology changes are needed.					
D2.3. Repository has procedures in place for monitoring or receiving notifications about changes in the needs of its Designated Communities (e.g., surveys, formal reviews, workshops and individual interactions).					
D3. Security					
D3.1. Repository maintains a systematic analysis of its environment: data, systems, personnel, physical plant, security needs, etc.					
D3.2. Repository has implemented mechanisms (processes) to adequately address each of the defined security needs.					
D3.3. Repository staff have delineated roles, responsibilities, and authorizations.					
D3.4. Repository has written disaster preparedness and recovery plan(s) (including at least one off-site copy of all deposited data).					
D3.5. Repository tests disaster plans regularly.					
D3.6. Repository has defined processes for service continuity and disaster recovery.					

[This page intentionally left blank.]

Glossary

Many of these terms are taken from the glossary of OAIS (2002).

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

Backup: The periodic capture of information to guard against system or component failure or against accidental or deliberate corruption of the system or system metadata. It is separate from the actions that most repositories will take of holding multiple copies of AIPs. Backups should ensure that lost or corrupted metadata can be restored, or that a failed system can be rebuilt and reintegrated into the repository with minimum loss of information. Backups are not expected to prevent all information loss. They are intended to restore a system or a component to a known state in a manner consistent with other system components, where this is applicable.

Content Information: The set of information that is the original target of preservation. It is composed of the digital object and its Representation Information.

Copies: Different logical or physical instances of the same object. Usually this will mean bit-wise identical copies stored on different file systems, on different media and/or in different locations. Most, but not all, repositories will have more than one copy of each AIP to guard against media failure or system failure. Some may choose to protect against certain software failures by using two different mechanisms to store the same object, such as having both a TAR and a ZIP file containing the same collection of files. In this case the bitstreams are different because the encapsulation format is different, but there is no question that they represent the same digital object. “Copies” also be taken to refer to different forms of the same entity that a repository may choose to hold for operational reasons. One trivial example might be the storage of TIFF and JPEG versions of an image to speed the production of DIPs in JPEG format. Here one form is clearly derived from the other, but it is important that changes in one form are propagated to the other in a predictable fashion.

Descriptive Information: The set of information, consisting primarily of Package Descriptions, that is provided to Data Management to support the finding, ordering, and retrieval of OAIS information holdings by Consumers.

Designated Community: An identified group of potential Consumers who should be able to understand a particular set of information. The Designated Community may be composed of multiple user communities.

Digital repository / Digital archive: These two terms are often used interchangeably. OAIS uses *archive* when referring to an organization that intends to preserve information for access and use by a Designated Community. *Trusted Digital Repositories: Attributes and Responsibilities* prefers the term *digital repository*. Digital archives and digital repositories should not be confused with either *digital libraries*, which collect and provide access to digital information, but may not commit to its long-term preservation, or *data archives*, which do include long-term preservation but limit their collections to statistical datasets.

Disaster: Any event that threatens or interrupts the operation of the repository and that, without corrective action, threatens the long-term preservation of its holdings. Disasters can include things that threaten the physical environment such as fire, flood, and explosion. They can also include the loss of facilities such as protracted network outages, or the inability to gain access to a building for prolonged periods due to severe weather or other contingencies.

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

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 an official ISO standard (14721).

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.

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

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

Versions of an object: A phrase that will not apply to all repositories. It is referenced here to avoid possible confusion, as Section D makes no requirements as to how versions of an object are handled. It refers to the fact that some objects can be considered to be later or alternative forms of other objects, such as the director’s cut of a film compared with the original cinema version, or different editions of the same book, or draft and final versions of a given document. A repository will usually choose to identify, through descriptive metadata, this type of relationship but it does not impinge on the preservation requirements of each object.

References

Consultative Committee for Space Data Systems. *Reference Model for an Open Archival Information System*. 2002. (ISO Standard 14721). www.ccsds.org/documents/650x0b1.pdf

Consultative Committee for Space Data Systems. *Producer-Archive Interface Methodology Abstract Standard*. 2003. www.ccsds.org/documents/651.0-R-1.pdf

Cornell University Libraries. *Digital Preservation Management: Implementing Short-term Strategies for Long-term Problems.* 2004. www.library.cornell.edu/iris/tutorial/dpm/index.html

Data Dictionary for Preservation Metadata: Final Report of the PREMIS Working Group. Dublin, Ohio and Mountain View, CA: OCLC and RLG, May 2005. www.oclc.org/research/projects/pmwg/premis-final.pdf

ISO 9000:2000 Quality management systems—Fundamentals and vocabulary. Geneva, Switzerland: International Organization for Standardization.

ISO/IEC 17799:2005 Information technology—Security techniques—Code of practice for information security management. Geneva, Switzerland: International Organization for Standardization.

Metadata Encoding and Transmission Standard (METS) version 1.4. Washington, DC: Digital Library Federation, 2005. www.loc.gov/standards/mets

Minnesota Historical Society, State Archives Department. *Trustworthy Information Systems Handbook.* 2002. www.mnhs.org/preserve/records/tis/tis.html

National Institute of Standards and Technology. *Security Self-Assessment Guide for Information Technology Systems (NIST Special Publication 800-26)*. Washington, DC: NIST, 2001. csrc.nist.gov/publications/nistpubs/800-26/sp800-26.pdf

National Institute of Standards and Technology. *Revised NIST SP 800-26 System Questionnaire with NIST SP 800-53 References and Associated Security Control Mappings*. Washington, DC: NIST, April 2005. csrc.nist.gov/publications/nistpubs/800-26/Mapping-of-800-53v1.doc

Task Force on Archiving of Digital Information. *Preserving Digital Information*. Washington, DC, and Mountain View, CA: Commission on Preservation and Access and the Research Libraries Group, 1996. www.rlg.org/legacy/ftpd/pub/archtf/final-report.pdf

Trusted Digital Repositories: Attributes and Responsibilities. Mountain View, CA: RLG, May 2002. www.rlg.org/en/pdfs/repositories.pdf

Appendix 1: A Discussion of Understandability & Use

OAIS states that the "Repository must ensure that the information to be preserved is independently understandable to the Designated Community." In other words, the community should be able to understand/use the information without needing the assistance of the experts who produced the information.

The definition of "understandable" in the context of a digital repository must be clearly defined, communicated, and committed to between repository and Producer. It may lie somewhere between "reproduce the bitstream as deposited," in the case where the bitstream by itself is always usable by the Designated Community, and "ensure the Information Content is rendered or performed, intelligible, and usable to the Designated Community given its current knowledge base, tools, and practices."

As a part of the process of submitting material to a digital repository, the repository must address the issue of the submission's information content and the extent to which this content is understandable to its Designated Community. The repository's responsibility should be defined in its charter, and it may be further elucidated in the Submission Agreement negotiated between the repository and the Producer. The extent of this responsibility can vary widely. If the repository is only tasked to preserve bits, not information content, for a submission, then this responsibility is not relevant.

More complex cases requiring information preservation can be viewed from two extremes. When a repository has minimal responsibility, the repository may be assured by the Producer that the information submitted is understandable to the Designated Community. The repository must have a clear definition of the Designated Community that includes the extent to which the repository needs to ensure the information content can be used by the Community's application tools. For example, if the Designated Community is defined as readers of English with access to widely available document rendering tools, the repository must ensure that the submitted information meets these criteria at the time of submission and that the corresponding information it delivers continues to do so (see Section B3 on preservation activities).

When a repository takes maximum responsibility, the repository cannot rely solely on the Producer's planned submission and must take additional steps to ensure that the information it receives for preservation can be understood by the Designated Community and is sufficiently usable. Again, the definition of the Designated Community should include the extent to which the repository needs to ensure the information content can be used by the community's application tools. The steps taken may include consulting with outside sources to evaluate the degree to which the information is understandable, and efforts by the repository to gather the additional metadata needed. This enables the repository to perform information preservation as well as bit preservation, and to do so long after the original Producers of the information are no longer available. The two major categories of information that must be understandable to the Designated Community are the Content Information and PDI. This discussion addresses only the Content Information, but it applies to the PDI too as it will have its own Representation Information.

Once the Primary Digital Object, its Representation Information (i.e., Content Information), and a definition of "understandable" have been determined, it is possible to ask whether what the repository disseminates is understandable to the Designated Community. In other words, it must be possible to apply the Representation Information to the Primary Digital Object and have the result be understandable to typical members of the Designated Community. This application process could take place within the repository, with only the result presented to the Designated Community in some new representation, or it could be left for the Designated Community to accomplish. If the process takes place within the repository, the repository must maintain its ability to perform it. If it is left to the Designated Community, the repository must also maintain the Representation Information so that it is "understandable" to the Designated Community, and it will need to periodically verify that most of the Designated Community can still perform the process, or the Designated Community must formally commit to this responsibility.

For example, the repository may maintain software that uses, or even partially or fully embodies, the Representation Information to render the Primary Digital Object in an informative visual or auditory manner for human consumption. Alternatively, the repository's software may present all the information through an interface acceptable to Designated Community applications. Or the repository may provide the Primary Digital Object and Representation Information, including their relationships, directly to the Designated Community with the understanding that the Designated Community understands how to apply the Representation Information to the Primary Digital Object to obtain understandable information. Scientific datasets often fall into this last category.

In short, the repository must maintain whatever is agreed to constitute the Content Information and its understandability requirements for the Designated Community. For certification, it is important for the repository to make clear its criteria for determining the Content Information and for determining its Designated Community's understandability requirements so that a third party can evaluate them in specific cases and with respect to the repository's charter.

Some examples can clarify the relationships among Representation Information, a Designated Community's needs, and the repository service that makes the information available to the Designated Community:

1. Digital Object Type: Word version 3 binary file from a government agency.

- Representation Information: Identifier of the format being "Word v3" and being proprietary.
- Content Information: Information from a government agency in a Word document.
- Designated Community: General public with access to widely available document rendering tools.
- Definition of "understandable": The Content Information is in a format currently renderable with widely available document rendering tools.
- Repository Access Service: Provides a binary file in a format currently renderable with widely available document rendering tools along with a unique identifier of the format type and the PDI. Upon request, may send the original binary file with its unique Representation Information identifier, assuming these are different. Note that for this proprietary format, the full Representation Information may only be available in the form of "embedded within the rendering software."

2. Digital Object Type: Binary file produced by the PDF application.

- Representation Information: Identifier of PDF-A format, described in a registry.
- Content Information: Document describing a medical procedure.
- Designated Community: English readers having a knowledge base typical of second-year medical students.
- Definition of “understandable”: Visually rendered exactly like visual rendering of original submission.
- Repository Access Service: Binary file, PDI, and PDF-A rendering application is made available.

3. Digital Object Type: Binary file containing observations from an instrument on a satellite.

- Representation Information: Binary file format definition and the definition of the meaning of the fields in the format (including detailed sensor characteristics of the satellite instrument), all given in an EAST (a formal syntax language) description with associated Data Dictionary.
- Content Information: Data from an instrument on a satellite.
- Designated Community: English readers having a third-year graduate school education in the associated scientific discipline.
- Definition of “understandable”: Original binary file is accompanied with sufficient Representation Information to allow a member of the Designated Community to understand how to access all the fields in the binary file, to understand what each field means, and to understand the relationships among the fields, and, using the PDI, to understand the context in which the field values were obtained.
- Repository Access Service: Provides the binary file, the Data Dictionary, EAST description, PDI information, and an identifier that allows a person to find the standards document that is the definition of EAST description language.

4. Digital Object Type: Software source code to perform simple function ‘A’.

- Representation Information: Identification of the language the code is written in, and a pointer to a definition of that language. If available, a natural language description of what function ‘A’ does. Also, a description of the inputs and the expected outputs, all understandable to the Designated Community.
- Content Information: Understandable and useable software source code.
- Designated Community: Software developers who may have an interest in code for functions like ‘A’.
- Definition of “understandable”: Fully documented source code is delivered with references (pointers) to primary technology dependencies such as language definition, system call, operating systems dependencies, build system, software environment requirements, relevant data standards, etc. All text is delivered in a currently usable character set. Information is sufficient to allow a member of the Designated Community to either compile and use the code correctly or to successfully transform the function to another language.
- Repository Access Service: Provides the software source code, Representation Information, and PDI upon request.

5. Digital Object Type: Software executable code.

- Representation Information: Identification of the platform environment in which the software can run, possibly including pointers to full descriptions of that environment and perhaps additionally to an emulation of that environment. Hopefully available, a natural language description of what function the code performs. Also, a description of the inputs and expected outputs, all understandable to the Designated Community.
- Content Information: Useable software executable.
- Designated Community: Software developers who may have an interest in code performing such functions.
- Definition of “understandable”: Binary object, executable in the environment specified in the Representation Information, with Representation Information and PDI that may be read and understood by members of the Designated Community.
- Repository Access Service: Provides the software executable code, Representation Information, and PDI upon request.

6. Digital Object Type: Musical score in a nonproprietary binary format.

- Representation Information: Description of the format in PDF-A, with a pointer to the PDF-A description in a different registry/repository.
- Content Information: Musical score for a synthesizer.
- Designated Community: German readers who wish to generate music using a computer and synthesizer from a digital representation of the score.
- Definition of “understandable”: Binary bitstream reproduced as delivered, Representation Information and PDI delivered in German.
- Repository Access Service: Provides the binary file, Representation Information, and PDI upon request.