



Principles and Good Practice for Preserving Data

Interuniversity Consortium for Political and Social Research (ICPSR)

Principles and Good Practice for Preserving Data

**Interuniversity Consortium for Political and Social Research
(ICPSR)**

December 2009

IHSN Working Paper No 003

Abstract

This document provides basic guidance for managers in statistical agencies who are responsible for preserving data using the principles and good practice defined by the digital preservation community. The guidance in this paper defines the rationale for preserving data and the principles and standards of good practice as applied to data preservation, documents the development of a digital preservation policy and uses digital archive audit principles to suggest good practice for data.

About ICPSR

The Interuniversity Consortium for Political and Social Research (ICPSR) is an international consortium of about 700 academic institutions and research organizations. ICPSR is a unit within the Institute for Social Research at the University of Michigan and maintains its office in Ann Arbor.

The Consortium provides leadership and training in data access, curation, and methods of analysis for the social science research community. It maintains a data archive of more than 500,000 files of research in the social sciences, and hosts 16 specialized collections of data in education, aging, criminal justice, substance abuse, terrorism, and other fields.

ICPSR's educational activities include the Summer Program in Quantitative Methods of Social Research, a comprehensive curriculum of intensive courses in research design, statistics, data analysis, and social methodology. ICPSR also leads several initiatives that encourage use of data in teaching, particularly for undergraduate instruction.

ICPSR-sponsored research focuses on the emerging challenges of digital curation and data science.

Acknowledgments

The document was developed by the Inter-university Consortium for Political and Social Research (ICPSR) for the International Household Survey Network (IHSN), with financial support from the World Bank Development Grant Facility, Grant No 4001009-06, administered by the PARIS21 Secretariat at OECD.

It was prepared by Nancy Y. McGovern, Digital Preservation Officer, ICPSR, with contributions from: Lance Stuchell, Digital Preservation Projects Coordinator, and research assistants Heather Backman, Annelise Doll, Michael Perry and Lindsey Williams.

Olivier Dupriez (World Bank), François Fonteneau (PARIS21), Gaye Parcon (consultant, PARIS21) and a team of reviewers have provided comments on and suggestions for successive versions of the paper.

The guidance refers to *Digital Preservation Management: Implementing Short-term Strategies for Long-term Problems*, with acknowledgement of contributions by Anne R Kenney and the rest of the workshop-development team at Cornell University Library in Ithaca, NY, USA; and funding from the US National Endowment for the Humanities.

© Copyright of this document is shared by OECD, the World Bank and the Regents of the University of Michigan (for ICPSR). The parts of the document defined as background intellectual property (ie, those developed as part of the Digital Preservation Management workshop that have pre-existing copyright held by Cornell University, for which the Regents of the University of Michigan have a non-exclusive licence to use) are licensed for use in this document under Creative Commons - specifically the Attribution, Non-commercial, No Derivative Works licence.

This paper (or a revised copy of it) is available on the web-site of the International Household Survey Network at www.ihsn.org.

Citation

Inter-university Consortium for Political and Social Research (ICPSR). 2009. "Principles and Good Practice for Preserving Data", International Household Survey Network, IHSN Working Paper No 003, December 2009.

The findings, interpretations, and views expressed in this paper are those of the author(s) and do not necessarily represent those of the International Household Survey Network member agencies or secretariat.

Table of Contents

Page Overview of this Guidance	vi
Acronyms	vii
1. Introduction	1
2. Principles and Standards	5
3. Putting the Principles into Practice	13
4. Formulating a Preservation Policy	19
5. Evaluation and Audit for Data Preservation	27

Annexes

A. Glossary	32
B. References	34
C. Milestones of Data Preservation	37
D. Programme and Policy Examples	39
E. Survey of Institutional Readiness	41
F. Technological Obsolescence and Other Threats.....	45

List of Tables

Table 1	Good practice for integrity features of digital content.....	5
Table 2	OAIS roles for data preservation.....	7
Table 3	OAIS functions and implications for data preservation	7
Table 4	OAIS information packages and implications for data.....	8

List of Figures

Figure 1	The high-level OAIS Reference Model	6
Figure 2	The overlap between OAIS and PAIMAS	10
Figure 3	The four phases of PAIMAS.....	11
Figure 4	The TDR attributes applied to an organisational setting.....	18
Figure 5	Three-legged stool for digital preservation	18

Overview of this Guidance

1. Introduction

This document provides basic guidance for managers in statistical agencies who are responsible for preserving data using the principles and good practice defined by the digital preservation community. The guidance defines the rationale for preserving data and the principles and standards of good practice as applied to data preservation; documents the development of a digital preservation policy; and uses digital archive audit principles to suggest good practice for data. This section defines the purpose and scope of the guidance.

2. Principles and Standards

Data archives have been preserving data in digital form since the 1960s and the digital preservation community has been emerging as an international, inter-disciplinary domain since the 1996 report *Preserving Digital Information*. This section reviews the current framework of community principles and standards and the relevance of these for data preservation.

3. Putting the Principles into Practice

The seven attributes of a trusted digital repository provide a framework for an organisation to preserve data that is adaptable to the mandate of the organisation -- a framework that will preserve the data and their nature.

4. Formulating a Preservation Policy

The development and implementation of a policy is essential for effective data preservation. A good policy reflects the mandate of the organisation to preserve data; provides an explicit commitment by the organisation to preserving data in accordance with community principles and standards; and fulfils a fundamental requirement of good preservation practice.

5. Evaluation and Audit for Data Preservation

The digital preservation community is developing a standard for self-assessment, audit and certification of digital archives. These principles provide a framework of recommended practice for any organisation involved in preserving data.

Annexes

- A. Glossary
- B. References
- C. Preservation Milestones for Managing Data
- D. Programme and Policy Examples
- E. Survey of Institutional Readiness
- F. Technological Obsolescence and Other Threats

List of Acronyms

AIC	Archival Information Collection (OAIS)
AIP	Archival Information Package (OAIS)
ASCII	American Standard Code for Information Interchange
CCSDS	Consultative Committee for Space Data Systems
CLIR	Council on Library and Information Resources
CRL	Centre for Research Libraries
DCC	Digital Curation Centre
DDI	Data Documentation Initiative
DIP	Dissemination Information Package (OAIS)
DRAMBORA	Digital Repository Audit Method Based on Risk Assessment
EU	European Union
IASSIST	International Association for Social Science Information Service and Technology
ICPSR	Inter-university Consortium for Political and Social Research
IHSN	International Household Survey Network
iPres	International Conference on the Preservation of Digital Objects
ISO	International Organisation for Standardisation
JISC	Joint Information Systems Committee
METS	Metadata Encoding and Transmission Standard
NASA	National Aeronautics and Space Agency
NDIIPP	National Digital Information Infrastructure Preservation Program, Library of Congress
Nestor	Network of Expertise in Long-term STorage and Availability of Digital Resources
NSF	National Science Foundation
OCLC	On-line Computer Library Consortium
OECD	Organisation for Economic Cooperation and Development
OAIS	Open Archival Information System Reference Model
PADI	Preserving Access to Digital Information
PAIMAS	Producer-Archive Interface – Methodology Abstract Standard (OAIS)
PDI	Preservation Description Information
PLANETS	Preservation and Long-term Access through Networked Services
RAC	Digital Repository Audit and Certification
RLG	Research Libraries Group
SIP	Submission Information Package (OAIS)
SPSS	Statistical Package for the Social Sciences
TDR	Trusted Digital Repository
TRAC	Trustworthy Repositories Audit and Certification
UK	United Kingdom
US	United States
UNESCO	United Nations Educational, Scientific and Cultural Organisation
XML	Extensible Mark-up Language

1. Introduction

This guidance aims to assist official data producers (national statistical offices and line ministries) in defining and meeting their digital preservation requirements and obligations. The recommendations address core preservation requirements. They identify steps for developing an effective preservation approach suited to the needs and requirements of individual official data producers and complying with standards of good practice within the digital preservation community, the latter being committed to the long-term management of digital content of all types. This section defines the purpose and scope of the guidance.

Purpose

Producers of official data (national statistical agencies and line ministries) generate vast amounts of data and information in digital form each year, including microdata from surveys and censuses, the content of administrative recording systems, databases of indicators, sample frames, registers, methodological and analytical reports and publications and maps and other types of content. These digital assets represent significant investment by producers and have considerable value for present and future users. This digital content is part of each country's heritage of its statistical agency's institutional memory and needs to be preserved.

Increasingly, this content is 'born' digitally, though print and other analogue formats are digitised for easier storage, discovery and use. The challenges of maintaining digital content over successive generations of technology have been recognised since the 1960s when the first data archives were established. There are numerous examples of important data being lost in the absence of an effective preservation approach -- for example data have been stored on outdated, unreadable mainframe tapes and other obsolete technology. Also, organisations have expected staff to preserve digital information without providing adequate support and resources to do so, and data often lacks adequate documentation to correctly identify valid versions. Statistical and other data-producing agencies have a responsibility not only to collect and disseminate data but also to guarantee their usability in the long term, either by preserving the data themselves or by entrusting specialist organisations with a mandate for preserving data. Data producers must commit resources to the management of data and related information of

enduring or at least continuing value, and continue to do so until that commitment is formally and explicitly ended. The decision to end that commitment could have political, economic, cultural and potentially legal ramifications, so sound selection criteria are needed to enable sound preservation planning.¹

A sustainable preservation programme addresses organisational issues, technological concerns and funding questions.

Organisational infrastructure includes the policies, procedures, practices and people – the elements that any programme needs to thrive but specialised to address digital preservation requirements. It addresses this key development question:

What are the requirements and parameters for the organisation's digital preservation programme?

Technological infrastructure consists of the requisite equipment, software, hardware, a secure environment and the skills to establish and maintain the digital preservation programme. It anticipates and responds wisely to changing technology. It addresses this key development question:

How will the organisation meet defined digital preservation requirements?

Resources framework addresses the requisite start-up, ongoing and contingency funding to enable and sustain the digital preservation programme. It addresses this key development question:

What resources will be needed to develop and maintain the digital preservation programme?²

The preservation requirements of these three organisational components are discussed in the guidance.

- 1 DPM tutorial, "Digital Assets", available at <http://www.icpsr.umich.edu/dpm/dpm-eng/program/assets.html>
- 2 DPM Workshop, "Conclusion", available at <http://www.icpsr.umich.edu/dpm/dpm-eng/conclusion.html>

Scope

In defining the scope of the guidance, it is necessary to distinguish between *digital preservation*, ensuring long-term access to data, and *data curation*, value-added activities to make specific data as understandable and usable as possible to their community. Data curation practices are not nearly as generalised as preservation practices because they are applied to content type rather than format type. In practice, different domains might curate data differently, but every domain should preserve data using common practices. For example, common data content types, eg, census and surveys, refer to content types that might be created and managed in a wide range of technological environments and might be stored in a wide range of file formats. The content type does not determine the file format, but the file format does inform the selection of an appropriate preservation strategy. *Digital curation* is a term that encompasses both data curation and digital preservation. There are many good sources for data curation and more recently for digital curation, eg, the UK Digital Curation Centre resources and the ICPSR Data Preparation Guide. There are fewer sources that connect good digital preservation practice to the management of data, so this guidance focuses on filling this gap.

The digital preservation community has been developing and promulgating standards and principles of good practice since the mid-1990s. This community has defined the fundamentals of good practice for organisations through a set of standards and principles for preserving digital content. This guidance incorporates the current set of community requirements and directs users of the guidance to preservation standards and principles that are emerging as the current set are extended and revised. The principles of good digital preservation practice are generalised, i.e. applicable to any organisation that preserves digital content of any kind – and local practice is individualised: an organisation is expected to demonstrate how the principles apply to its specific organisational setting based upon local factors, e.g. relevant mandates, requirements, available human and technological resources and the extent of digital content to be preserved.

All organisations that produce and manage data are expected to adhere to these principles. However, all such organisations will not apply the principles in the same way or develop the same approach to preserving data. To address the need for each

organisation to demonstrate compliance with good practice, the guidance distils the current standards and principles of the digital preservation community into recommendations for managers to use in framing their organisation's approach. The desired outcome of this guidance is to encourage organisations to achieve and maintain a state of well-managed data as organisational requirements and technology evolve. This guidance -- ideally in conjunction with relevant training and access to the growing set of suitable organisational examples of approaches and practice (examples of which are included in the guidance) -- should achieve this outcome.

Rationale for Preserving Data

The rationale for effective and compliant preservation of data encompasses a range of incentives. A primary incentive is to meet the obligations when an organisation is required by legislation or another mandate to keep data. A second incentive is to maximise the investment of resources in data production when there is an expectation that they will continue to be available into the foreseeable future due to the cost of producing them. A third incentive is to demonstrate the capability, in relation to peer organisations, to maintain status when there is an implicit or explicit expectation that an organisation is preserving data. A fourth incentive is to avoid the embarrassment of losing data when an organisation is identified as responsible for managing them. Finally, a key incentive is to provide ongoing access to the historical and cultural heritage of a nation.

In early 2010 the Blue Ribbon Task Force on Sustainable Preservation and Access -- convened by the US National Science Foundation, with international support, to examine the cost of long-term preservation -- is expected to complete and publish its report. Its results might offer additional incentives to the rationale for digital preservation costs. For now, these are the most common incentives for preservation and this set has been sufficient for many organisations to justify their programmes.

Digital Preservation Terminology

As a starting point for discussing good practice for digital preservation, a brief review of preservation terminology in use within the community provides a context for the guidance (see Annex A for definitions of terms used). A challenge in discussing and sharing examples of digital preservation practice has been a lack of standardised language. There is yet no international

authoritative source that provides the community with approved definitions of digital preservation and related terms, as they emerge and evolve. Therefore, this guidance uses a working definition reflecting discussion within the community about the scope and meaning of the term. The term *digital preservation* refers to all the activities undertaken by a steward (an organisation or individual) to ensure that the digital content for which the steward has responsibility is maintained in usable formats and can be made available in meaningful ways for current and future uses over time. To understand the term, it is necessary to define *activities* in the context of preservation and to *consider* the implications of managing data *over time*.

In practical terms, preservation activities include explicitly identifying the digital content that needs to be preserved; taking responsibility for bringing that digital content into a sustainable environment with appropriate policies and procedures; and ensuring that the digital content can be made available over time to authorised users. Organisations that are responsible for managing data might include data archives, national statistical offices and line ministries, research programmes that produce data, and any other organisation with responsibility for producing and managing data. Users within these environments might be internal or external and preservation is intended to ensure that data remain useful for any required or desirable use.

In this context, *over time* refers to the length of time the digital content should remain usable for legal, financial, cultural, business or other purposes. This might be measured in months, years, decades or centuries. In practice, *over time* refers to the need to manage data for five years or more because the speed and impact of continual technological change requires that organisations take action to ensure the longevity of data within their care. If an organisation is required or wishes to ensure that digital content is usable and meaningful for five years or more, that equates to preserving the data because they will have to remain useful for one or more generations or iterations of the supporting technology, eg, hardware, software).

Digital content consists of a growing range and combination of information types, e.g. text, images, geospatial data, audio and video. Data for social science and other domains increasingly refer to all these types of digital content. Throughout its lifecycle, digital content relies upon information technology of all kinds

to remain meaningful over time. For digital content to be usable and meaningful, authorised users must be provided with the means — human or automated — to find, open, read and use the material maintained by a digital preservation programme. In this guidance, the digital content referred to is primarily social, economic and other human-related data, e.g. population censuses, labour force surveys, macroeconomic databases and related publications, although the recommendations are relevant to data produced in other contexts.

Digital information is at risk of loss in two important ways: obsolescence of enabling technologies and physical damage. Obsolescence might affect software, hardware and even the arrangement of the data in a stored file, and it might occur at an alarming pace.

- A file format might be superseded by newer versions, which might no longer be supported by the current vendor or relevant standards body.
- A storage medium might be superseded by newer and denser versions of that medium or by new types of media — smaller, denser, faster and easier to read.
- The device needed to read a storage medium might no longer be manufactured.
- Software used to create, manage or access digital content might be superseded by newer versions or newer generations with more capability using the most current technology.
- Computers of every size and scale are continually superseded by faster and more powerful machines that can store and process more and more content.
- Vendors of all technology compete, emerge, merge and fade, making it even more difficult to maintain digital content over time.

Digital information is also vulnerable to physical threat. Like obsolescence, physical damage might affect one of the multiple components required to access digital information, namely hardware and media. Computer components and media can physically fail due to human error, natural events and even just the passing

of time.³ One option for an organisation wishing to be alert to relevant changes is to monitor technological developments and systematically consider potential preservation implications. Currently, there is no central place within the community where all the necessary information has been brought together, but Annex B identifies a number of community-based resources that provide updates of relevant developments, e.g. Preserving Access to Digital Information (PADI), and the guidance notes emerging developments of interest at appropriate points in the document. Annex F offers a more detailed discussion of the nature and implications of technological obsolescence and other threats to digital information from a preservation perspective. The challenges noted in Annex F are addressed throughout the guidance.

Defining Good Practice for Digital Preservation

The digital preservation community has been developing principles and standards for good practice over the past dozen years, beginning in 1996 with the release of *Preserving Digital Information*. Since then, the documents developed by the community have provided a comprehensive framework of good practice for managing data. The framework is presented in this guidance and should be achievable by any organisation with preservation responsibilities. The specific community documents in the framework, which are referenced, include: *Open Archival Information Systems (OAIS) Reference Model*, a standard approved by the International Standards Organisation (ISO) in 2003 that defines the technological framework for preserving data; *Trusted Digital Repositories (TDR): Attributes and Responsibilities*, a report released in 2002 that provides the organisational context for preserving data; the *Data Seal of Approval*, a set of principles for data archives released in 2008 that presumes the use of both TDR and OAIS in demonstrating good practice for producers, archives, and users of data; and *Core Requirements: Ten Basic Characteristics of Digital Preservation Repositories*, a set of high-level principles released in 2007 that reflects consensus on audit and control requirements and practice for digital archives and provides a framework for implementing good practice. The guidance refers to other relevant community documents where applicable

(see Annex B for a list of resources referenced in the guidance).

Good practice does not change whatever type of content to be preserved. The principles of digital preservation define what should be done for any organisation. Each organisation determines how preservation will suit its requirements, technological setting and content. In practice, specific preservation plans are developed to reflect the nature of the materials, especially the file formats; the usage requirements for long-term management, e.g. authenticity and legal requirements; and the technology environment in which the content was created and will be managed over time. Most common file formats have been associated with one or more preservation strategies devised within the digital preservation community. Throughout the guidance, there are examples illustrating the principles of good practice.

This document guides organisations towards an understanding of the standards and principles of the digital preservation community and defining attainable objectives and stages of development for achieving alignment and compliance with those standards and principles. Whenever relevant, the recommendations consider good practice for the long-term management of data in terms of minimum requirements, recommended practice and ideal (or fully compliant) levels, dependent upon the needs and mandate of the organisation.

For organisations not familiar with preservation and which plan to use the guidance to develop a digital preservation approach, a comprehensive self-assessment of requirements, resources and capabilities is a good starting point. For this purpose, Annex E provides a survey of institutional readiness form developed for the Digital Preservation Management workshop series. The survey provides a checklist to assist an organisation in evaluating its digital assets in terms of scope, priorities, resources and overall readiness to address digital preservation concerns.

3 A R Kenney, NY, McGovern and Entlich et al, *Digital Preservation Management: Implementing Short-term Strategies for Long-term Problems*, an on-line tutorial developed and maintained for the Digital Preservation Management workshop series since 2003. See the "Obsolescence and Physical Threats" section of the tutorial at: <http://www.icpsr.umich.edu/dpm/dpm-eng/oldmedia/index.html>

2. Principles and Standards

Data archives have been preserving data in digital form since the 1960s (see Annex C for a complete set of milestones for data preservation). The digital preservation community has been emerging as an international, inter-disciplinary domain for more than a decade. References in this guidance to the *digital preservation community* include any institution or individual in any organisational context in which participants have acknowledged formally or informally the intention of preserving digital content across one or more generations of technology. The publication of the 1996 report *Preserving Digital Information* represented a first step in formalising good practice for organisations engaged in digital preservation. A series of community documents followed that report, in particular a set of approved and emerging international standards that began with the Open Archival Information System (OAIS) Reference Model in 2003. This section defines the current framework of community practice for digital preservation using these documents and indicates how community documentation offers practical guidelines for preserving data.

Founding Document: Preserving Digital Information, 1996

The 1996 *Preserving Digital Information: Report of the Task Force on Archiving of Digital Information* report marked the starting point of the digital preservation community. This proved to be a seminal report identifying the core challenges of digital preservation for the community that manages digital collection -- a set of essential criteria for well-formed digital content, roles and responsibilities for preserving it and a set of principles. The report remains useful to organisations in developing an understanding of digital preservation

as a problem to be solved by the community. It provided a framework for the community documents that followed, many of which incorporated components of the 1996 report.

The report identified these four challenges faced by organisations in managing digital collections over time.

As technology upon which digital content relies is superseded or replaced by new technological developments, effective preservation requires avoiding the impact of this cycle of obsolescence

- As technology evolves, digital content needs to be migrated over generations of technology
- In addition to technological concerns, organisations need to address relevant legal and organisational issues to accomplish digital preservation objectives
- To address the challenges of digital preservation, there is a need for community-wide infrastructure

These challenges remain relevant and continue to shape the development of the community, as well as providing priorities for organisations that preserve digital content.

The report identified five integrity features for digital content that guide preservation efforts. Table 1 indicates good practice associated with each feature.

Table 1 Good Practice for Integrity Features of Digital Content

Integrity Feature	Related Good Practice
Content: ensures that essential elements of digital content are preserved	An organisation is expected to explicitly identify and actively manage data to be preserved
Fixity: requires that changes to content are recorded, ideally from the moment of creation onward	At minimum, this feature might be addressed through routine use of a checksum (a computed value generated by widely-available utilities that uniquely identify a digital file) to detect intentional or unintentional changes to data and notify data managers for action
Reference: ensures content is uniquely and specifically identifiable in relation to other content across time	For example, an organisation is required to adopt and maintain a persistent identifier approach – a system for assigning and managing enduring identifiers that allows digital objects to be consistently and uniquely referred to over time
Provenance: requires digital content be traceable to its origin (point of creation) or, at minimum, from deposit in a trusted digital repository	This feature requires that an organisation records information (captured as metadata) on the creation and action that have affected the content since its creation (e.g. data deposited in an archive, migrated from one file format to another)
Context: documents and manages relationships of digital content	An organisation that preserves data document relationships between its own digital content and, to the extent possible, to data managed by other organisations

It is essential that an organisation responsible for preserving digital content is able to demonstrate how it addresses each of these five integrity features. Apart from the content feature, these features are addressed through the capture, creation and management of metadata for preservation, as indicated in the descriptions of each feature. Preserving metadata is initially addressed in the next section on OAIS.

- The core principles for data preservation that the report described included:
- Data producers (creators, providers and owners) are responsible for archiving their digital content to ensure the objects are preserved
- Data producers might meet this obligation by depositing data to be preserved in a certified digital archive (one that meets the current definition of a trusted digital repository)
- Data producers (or the digital archives they engage) should select data for preservation so that sufficient resources can be designated to maintain the data over time
- Data selected to be preserved must be packaged and prepared for preservation with sufficient metadata and in formats that are able to be preserved (e.g. avoiding software-dependent formats where possible, embracing ASCII and now XML formats)

- Data producers (or the digital archives they engage) should identify a technology environment (software and hardware combination) sufficient in scale and security to manage the data
- Data producers (or the digital archives they engage) should adopt and implement migration and other preservation strategies to ensure that data are usable over time

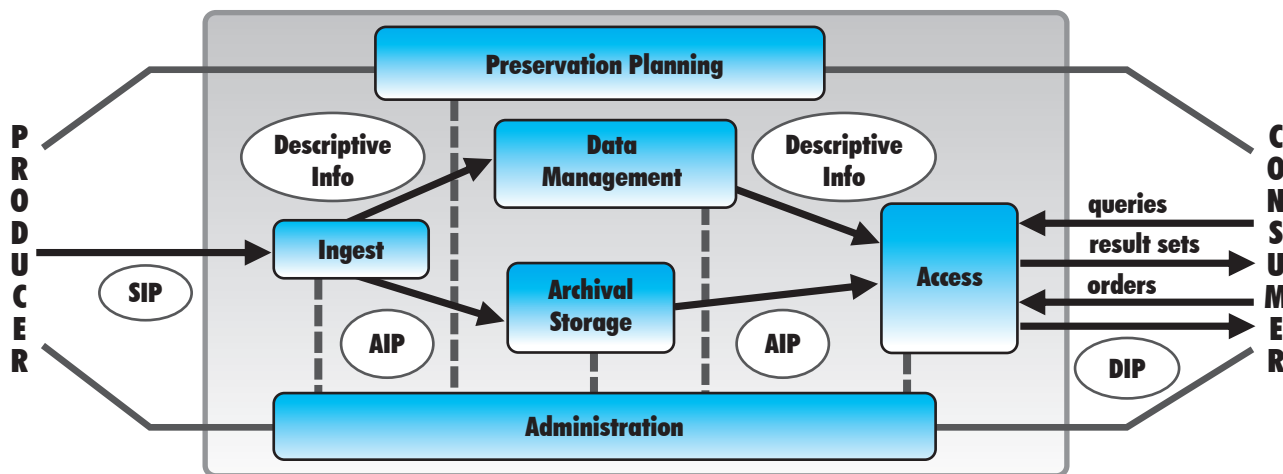
These principles define preservation-related responsibilities for digital content that informed the development of community standards following the release of the 1996 report.

Open Archival Information System (OAIS) Reference Model and Related Standards

The Open Archival Information System (OAIS) Reference Model represents the most formal and comprehensive expression of preservation as a process that is available for community use. OAIS describes the roles, functions and content components involved in preserving digital content.

The international space community initiated the development of the Open Archival Information System (OAIS) Reference Model in 1995 to address its own need to manage data better. It convened an international and representative group of experts. OAIS was approved by the International Standards Organisation (ISO) as ISO 14721 in 2003. Prior to the development of the OAIS standard, there was no language or practice in place within the community to enable productive discussion of good practice across organisations, domains or

Figure 1 The High-level OAIS Reference Model



national boundaries. The promulgation of the OAIS standard has provided a common language for the digital community. The OAIS model was developed to be applicable in any organisational context in which digital content is managed for the long term. Most organisations that manage digital collections have indicated an intention to design and implement their digital repositories in accordance with the OAIS. Any organisation that preserves data should be aware of the OAIS Reference Model, commit to developing its technological approaches in accordance with OAIS, and monitor the community for OAIS-related developments. Section 4 of this guidance, on the technological infrastructure for preserving digital content, uses the OAIS as a framework.

OAIS Roles

OAIS defines three explicit roles (Producer, Manager and Consumer) and one implicit role (Archive) that participate in an OAIS environment, as illustrated in Table 2.

Table 2 OAIS Roles for Preserving Data	
OAIS Role	Implications of Roles
Producer	generates or is responsible for data to be preserved and provides the data to the archive or unit responsible for preservation
Manager	has direct authority for the archive (or unit responsible for preserving data), specifically the approval of the budget and development of high-level policy for preservation initiatives
Consumer	is authorised to use the data, either for internal purposes or general public use
Archive	is the unit responsible for coordinating OAIS functions and managing the content packages defined by the OAIS standard (see Table 4)

In practice, a producer or consumer might be an individual, an organisation or a system authorised to provide or use data. These roles are useful for discussing responsibilities that pertain to preserving data and for raising awareness of it. Organisations that preserve data are required to ensure these roles are assigned and implemented effectively.

OAIS Functions

The OAIS identifies seven areas of activity (defined in Table 3) performed in managing digital content over time: *Ingest*, *Archival Storage*, *Data Management*, *Administration*, *Preservation Planning*, *Access* and *Common Services*. Each area consists of individual functions for performing that activity.

OAIS Content Packages and Metadata

The OAIS defines three types of information packages (described in Table 4) that contain digital content over time as data are preserved.

Table 4. OAIS information packages and implications for data preservation

These packages represent stages of life-cycle management for data: acquiring (SIP), storing and archiving (AIP) and making available to authorised users (DIP). An organisation that preserves digital content must interpret and develop the SIP, AIP, and

Table 3 OAIS Functions	
OAIS Function	Implications for Data Preservation
Ingest: enables secure acceptance and quality control of data for preservation	Documents the identification of data to be preserved, whether by a unit within an organisation that creates data or by a designated data archive. The preserving unit needs to determine what types of data (e.g. file formats) will be preserved, what kinds of validation or quality control will be performed, what steps need to be taken to ensure the data are able to be preserved (e.g. conversion to common file formats or to ASCII-based formats)
Archival Storage: ensures the secure storage, management, and retrieval of preserved data	Organisations might be inclined to equate system back-up with archival storage. For preservation purposes, archival storage is replaced by multiple copies, preferably in more than one media to avoid massive failure, e.g. on-line and off-line, distributed copies stored in more than one location, and enough copies to avoid loss. The trend within the community is towards managing most copies on-line. When that option is not possible, the most effective and feasible means for managing multiple copies should be adopted and the procedures updated as technology and requirements evolve.
Data Management: supports ongoing accumulation and availability of administrative data on the operation of and documentation of the content of an archive	An organisation requires complete and current information about data as units (e.g. data files) and aggregates (collections, studies or series), as well as information about the repository and its management (e.g. policies, standards, procedures and performance measures). More standardised approaches to this administrative documentation are slowly beginning to emerge.
Administration: develops, maintains, and applies policies and procedures to operate and coordinate OAIS functions	Administration is where organisational and technological responsibilities intersect. Preservation decisions are captured in policies and procedures that inform repository operations, e.g. policies on acquisition, storage management, disaster recovery and access.
Preservation Planning: develops and recommends standards, policies, procedures and means for preserving digital content	Provides advice and procedural support for data preservation, e.g. preservation strategies for specific data content types. This should be a shared responsibility within the community: no organisation reinvents the wheel, no single organisation is required to do it all and all organisations are able to benefit from lessons learned by others and share their own experiences.
Access: finds and delivers content in an archive to authorised users	Preservation enables long-term access. The OAIS requires a repository to be able to find and deliver the data over time in effective and efficient ways, whether access is provided for internal or external use.
Common Services: refers to the operating system, network services and security services needed to implement and manage a preservation repository	An organisation needs to ensure these common services are robust and reliable enough to sustain data over time. Adhering to preservation requirements might require supplementing the common services used for current activities, e.g. more levels of security or different approaches to encryption to ensure that data are able to be read and used.

Table 4 OAIS Information Packages and Implications for Data Preservation

OAIS Information Package	Implications for Data Preservation
Submission Information Package (SIP): the state (format and structure) of the digital content when it is provided to the repository or unit responsible for preservation	Effective preservation should make it possible always to be able to demonstrate the state of data as submitted (the original)
Archival Information Package (AIP): the state of digital content when it is placed into archival storage for preservation	An AIP accumulates information on preservation activities (e.g. data are migrated to a new format, preservation metadata are updated) so anything that might affect preservation or use of data is known and documented
Dissemination Information Package (DIP): the state of the digital content when it is made available to users over time	The format and structure of the DIP reflects technology as it evolves. Access is expected to use current technology to deliver data; preservation uses proven and reliable technology

DIP to meet its own requirements in alignment with community practice, especially metadata requirements when an organisation is responsible for providing its data to other organisations.

Preservation Metadata and OAIS

The Open Archival Information Systems (OAIS) Reference Model incorporated the five integrity features identified in the 1996 report into its definition of Preservation Description Information (PDI) within the Archival Information Package, the formal definition of digital content for preservation. The 2009 revision of OAIS acknowledged the importance of documenting and managing rights associated with digital content by adding access rights' information (e.g. copyright information for data, licensing information for data, right to copy and otherwise preserve data) to this list of integrity features for digital content.

Preservation metadata include all the information, including descriptive, structural, administrative and technical metadata, required by an organisation to preserve data. As with other preservation terms, there is no authorised definition of *preservation metadata*, although common approaches to creating and managing metadata for preservation are beginning to emerge within the community. There is also no international standard for preservation metadata.

The Preservation Metadata Implementation Strategies (PREMIS) data dictionary is the only metadata definition specifically developed for preservation. The most recent version of the PREMIS data dictionary was released in 2008 and, since the initial version in 2005, tools and examples for using PREMIS have emerged, making PREMIS more likely to be used by organisations for preserving digital content. PREMIS is a community initiative that recommends good practice for preservation metadata and it might be formalised as a standard.

For data, metadata play a very large role in preservation because the documentation required to understand and use data is a specialised form of

metadata. Documentation that conforms to the Data Documentation Initiative (DDI) generally meets the requirements for descriptive and structural metadata (see below for more information about DDI). In addition, organisations that preserve data should identify a minimal set of metadata for data files (e.g. data, file name, checksum, file format, content type, such as data or documentation) to ensure ongoing access to data.

These are other metadata developments that pertain to or might be useful for data preservation:

- The minimal set of Dublin Core metadata elements is a simple set of metadata widely used internationally to describe resources, including data.
- The Data Documentation Initiative (DDI) has produced DDI 2.0, a fixed format for metadata that might be used to document data for preservation purposes, and DDI 3.0, a new version of DDI that addresses the whole life-cycle for data management, including preservation, and shows promise for broad on-line use when tools to implement it are in production for use by any organisation.
- Statistical Data and Metadata Exchange (SDMX), a technical specification for the exchange of data and metadata, was approved as an ISO standard in 2005 (ISO 17369) and is increasingly used for time-series and other data.
- Information Technology – Metadata Registries (ISO/IEC 11179), a specification for standardising and registering data elements to make data understandable and shareable.

These metadata initiatives are valuable community developments that also support preservation through improved documentation of the structure and content of

data. All these developments are typically implemented using Extensible Mark-up Language (XML), a textual data format designed for simplicity, interoperability and generalised use. XML tags contain internal documentation and description that make XML-based documents easier to use and preserve.

Often, metadata schemas are used in combination because each was developed for a specific purpose and each has strengths and limitations. One step for integrating multiple metadata schemas is a crosswalk that maps one schema to another, e.g. this element on schema A is the same as or relates to this element on schema B. In the preservation context, the Towards Interoperable Preservation Repositories (TIPR) project is an example of a project focusing on the packaging and exchange of metadata for preservation. TIPR is a two-year project funded by the Institute for Libraries and Museum Services (IMLS) in the USA and scheduled to conclude in 2010.

Managing Data Using OAIS

The OAIS Reference Model can be a powerful communication and planning tool for organisations responsible for preserving data. The model can be used to illustrate the assignment of responsibilities, the costs associated with various stages of life-cycle management, the completion of policies and other required documents and the priorities for preservation-related developments. There is a community expectation that an organisation that preserves digital content will be aware of and working to comply with the OAIS standard.

Although the community trend is towards using repository software to preserve digital content, this section does not presume its use for the effective management of data by an organisation over time. An organisation is not required to implement software repository to preserve content, although doing so might make preservation decisions and actions more consistent and scalable. There are no currently-available and in-production software packages fully compliant with the OAIS, particularly in the functional areas of Administration and Preservation Planning, though that is an objective of many developers. There are several options available to organisations for repository software to enable the preservation of digital content:

- *Adopt an open-source package and adapt the package to suit requirements:* the Fedora (Flexible Extensible Digital Object

Repository Architecture) repository software packages (available through the Fedora Commons) is one example of open source⁴ repository software that is increasingly used and has a solution community to help organisations adapt Fedora for preservation;

- *Purchase a software package and adapt it for use:* one example of a software package available for purchase and which acknowledge the OAIS standard is the Rosetta software from *ExLibris*, developed as a result of partnership with the National Library of New Zealand;
- *Develop repository software, alone or in collaboration with partners:* building a repository requires ongoing support for programming time, equipment and software tools for developing the repository that might require licensing.

Each of these options requires an understanding of the OAIS Reference Model standard and related preservation requirements identified in this guidance and an ongoing commitment to maintaining the repository.

If an organisation implements a repository to manage its digital content, a technological platform might need to be assembled, scaled appropriately, documented, monitored for key developments and upgraded or enhanced as necessary (note the discussion of cloud-computing services under OAIS Compliance in the next section of the guidance). There are four likely scenarios for developing a technical infrastructure. Each has its strengths but the correct solution will depend on the capabilities and priorities of the organisation. The four options are:

Build	Opt to build on internal IT for DP program. Focus on developing capacity and storage options.
Extend	Define based upon gap analysis. Add capacity and storage. Ensure adequate redundancy in back-up.

⁴ Open source software is freely available, although there are costs to the organisation in implementing it in terms of programming time, technical skills that might need to be acquired through consulting, and other costs.

- Join** Evaluate alternatives, then outsource or subscribe to service.
Ensure longevity through effective contract/service management.
- Collaborate** Specialise in core areas of expertise or capacity.
Extend capacity through collaboration.
Share well-defined responsibilities.
Leverage extra-repository benefits.⁵

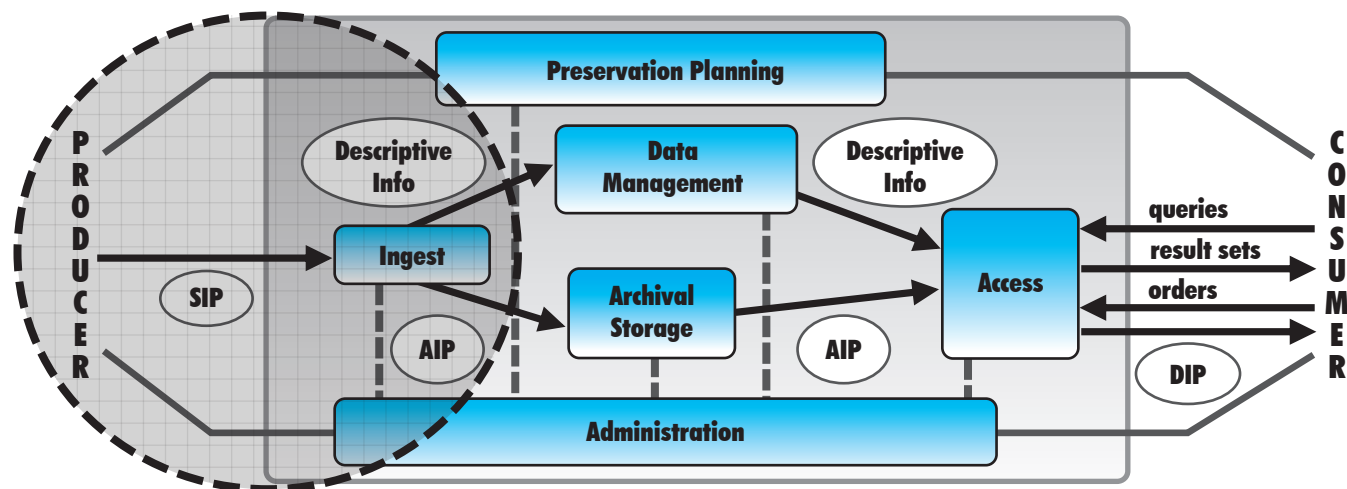
The factors an organisation will need to consider in determining a workable digital repository option for its situation include: the extent of digital content to be managed (scale), the degree of access the organisation has to technical support for implementing the repository, the resources available to implement and sustain a repository, and the time-frame the organisation needs to have a repository in place.

The OAIS Reference Model identified a road map for the development of other related standards including:

Producer-Archive Interface Methodology Abstract Standard (PAIMAS)

In 2003 the OAIS working groups released the *Producer-Archive Interface – Methodology Abstract Standard (PAIMAS)*, which was approved as an ISO standard in 2006 (ISO 20652: 2006). The PAIMAS standard addresses the relationship between producers of data and an archive from the point of deposit through the ingestion of the data into the archival storage component of the archive, including the specification of deposit agreements, requirements for and frequency of deposits and responsibilities of the producer and the archives (e.g. specifying what happens if there are questions about or difficulties with the deposit). As Figure 2 illustrates, PAIMAS overlays the OAIS Reference Model, encompassing the roles of the producer (the organisation or individual producing or owning the data with responsibility for its long-term management) and the archive (the entity that takes on long-term management of the data) -- from the identification of the content for submission to the archive to the ingestion of the data into the OAIS system.

Figure 2 The Overlap Between OAIS and PAIMAS

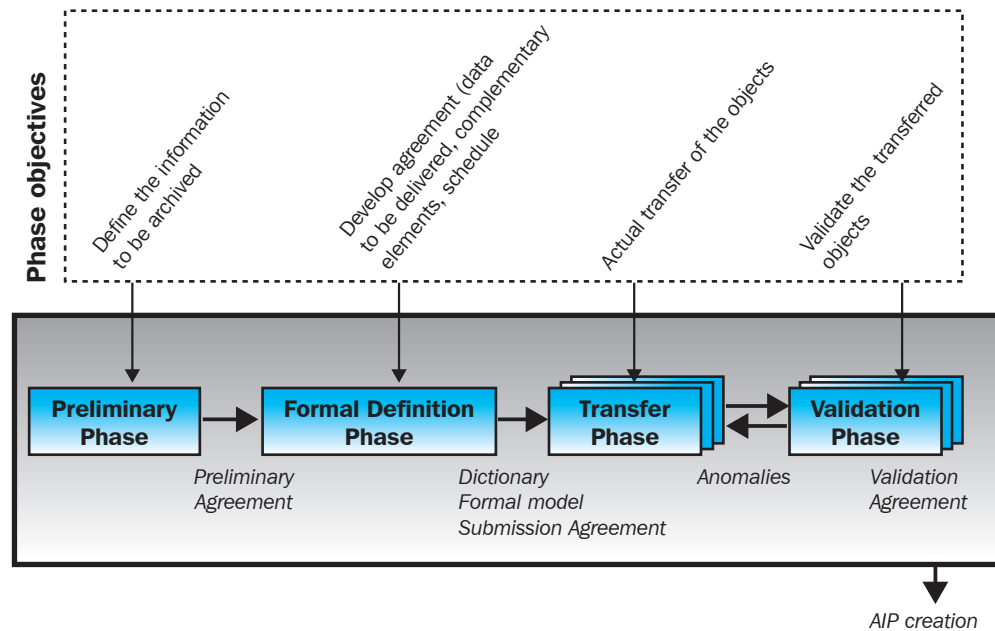


a preservation metadata standard; submission and dissemination standards, for identification of digital content (persistent identifiers); preservation methods; accreditation (or certification); and the need for a producer-archive interface standard. These standards continue to be developed and maintained by the community

The standard delineates the interaction between the producer that creates (or produces) the digital content and the archive which accepts long-term responsibility for preserving this digital content. Managing data over time might involve placing data in a data archive or an organisation establishing its own archival function for preserving its data.

5 DPM Tutorial, "Program Elements", available at <http://www.icpsr.umich.edu/dpm/dpm-eng/program/techinf.html>

Figure 3 The Four Phases of PAIMAS



PAIMAS identifies four phases (Figure 3) to coordinate the transition of digital content from current use to archival storage:

- **Preliminary Phase:** identifies and specifies digital content to be preserved, including the initial contact between the producer and the archive, a preliminary definition of the scope of the interaction and drafts of the Submission Agreement.
- **Formal Definition Phase:** defines requirements and expectations for preserving data and finalises the submission agreement stipulating the terms and means of depositing the digital content into the archive
- **Transfer Phase:** agrees upon the requirements for sending data to an archive (or moving it to archival storage for management) and defines a workflow identifying who will be responsible if there are questions about or issues with a transfer
- **Validation Phase:** defines the quality checking (validation steps) that will be completed for data to be preserved

These phases make explicit the steps an organisation must take when archiving data.

Formalising the interaction between the producer and the archive might make the process of identifying data to be preserved easier and more efficient. When the data are managed internally, the PAIMAS stages serve as a reminder to make the process of identifying content for preservation and placing content into archival storage implicit and intentional. An inventory of data for which a data-producing agency is responsible is a first step in addressing its digital- preservation obligations. The Digital Curation Centre in the UK recently developed a Digital Audit Framework to assist in identifying and assessing data for long-term retention and use; this might be helpful to data-producing agencies in completing this required step.⁶

Ongoing Community Developments of Standards and Principles

Community documents, of which the documents discussed in this section are examples, continue to emerge as good digital preservation practice is defined and standardised. The Blue Ribbon Task Force on the costs of digital preservation, introduced in section 1 of this guidance, is one example of emerging good practice within the community. The pending approval of an ISO standard for the certification of digital archives,

⁶ Digital Curation Centre, *Data Audit Framework*, 2008: <http://www.data-audit.eu/>.

discussed in Section 3, is another example. This section discusses the current core set of community documents. It can be difficult for organisations to keep up with developments in the digital preservation community, but it is essential that conferences and community literature – both ‘grey’ and more formally published – are employed as primary ways for them to keep track. Increasingly, conferences offer proceedings that enable organisations to learn about developments reported at meetings.

Since 2004, several professional conferences have provided ongoing sources of information about current developments and practice by featuring or including digital preservation. The Society for Image Science and Technology (IS&T) Archiving Conference has been held since 2004. This conference series was initiated by the digital-imaging domain and includes both general digital preservation sessions and image-specific preservation sessions. The Digital Curation Centre (DCC) has hosted conferences that include digital preservation topics in the programme, with

international delegates since 2004. The International Conference on the Preservation of Digital Objects (iPres) has been held annually since 2004. iPres is the first regularly-held international conference entirely devoted to digital preservation.

In addition to periodical articles about the challenges of preserving digital content that have appeared in the professional literature of relevant domains – since the late 1960s for archival literature, the 1980s for library literature and the 1990s for museum literature – increasing numbers of publications since 1996 either highlight or are devoted to digital preservation issues, including *Ariadne* in the UK and Preserving Access to Digital Information (PADI) in Australia. These are community-based publications providing information and updates on research and developments pertaining to digital content. The *International Journal of Digital Curation* was launched in 2006. Digital curation includes both data curation and digital preservation, making it particularly relevant to organisations that manage data.

3. Putting the Principles into Practice

The OAIS Reference Model describes the roles, functions and content types of an effective digital archive. The OAIS standard does not provide much guidance on developing the organisational side of operating an OAIS system. The *Attributes of a Trusted Digital Repository: Roles and Responsibilities* report, released in 2002, defined and described expectations and requirements for an organisation intent on being a trusted digital repository (TDR). The document acknowledged that the OAIS Reference Model did not provide specific guidance for organisations on conforming to the standard, on managing the data from an organisational rather than a technological perspective and on achieving and maintaining trust in an organisation that manages data over time.

The TDR document has become a *de facto* community standard for good practice by organisations that manage digital content over time. The seven attributes of a trusted digital repository provide the framework for applying the standards and principles in Section 2 of this guidance. The TDR attributes are adaptable to the mandate and scope of any organisation that preserves data, and every organisation that preserves data is expected to reflect the attributes of a trusted digital repository.

This section uses practical recommendations to illustrate the application of the TDR attributes to an organisation responsible for managing data over time. The seven TDR attributes – *administrative responsibility, organisational viability, financial sustainability, technological and procedural suitability, system security, procedural accountability* and *OAIS compliance* – are each addressed in the following sections.

Administrative Responsibility

The administrative responsibility attribute requires an organisation with responsibility for preserving data to make an explicit commitment to that end. Additional characteristics of this attribute include a commitment to meet or exceed community standards; to share performance measurements with depositors; to involve external community experts in periodic reviews to validate and certify the organisation's processes and procedures; and to thoroughly document and be accountable for preservation-related actions.

The most common way for an organisation to demonstrate compliance with this attribute is through the promulgation of a mission statement encompassing data preservation. The Inter-university Consortium for Political and Social Research (ICPSR) has a mission statement that serves as a good example of for data preservation: "ICPSR provides leadership and training in data access, curation and methods of analysis for a diverse and expanding social science research community". It has become clear, based on accumulated examples within the digital preservation community, that it is difficult or impossible for an organisation to sustain a digital preservation initiative if there is no overt support for the programme at top organisational levels or by key stakeholders.

Organisational Viability

The organisational viability attribute requires that an organisation that accepts responsibility for preserving digital content has the wherewithal to undertake digital preservation – e.g. the legal status, the mandate, the resources, the expertise, the capacity and the equipment. Additional characteristics of this attribute include a commitment to demonstrating the viability and trustworthiness of the organisation's preservation programme; establishing well-documented preservation policies and practices (e.g. procedures and work-flows); defining comprehensive written agreements with depositors; reviewing and updating policies and procedures on a regular basis; and engaging in contingency and succession (trusted inheritors') planning, in case the organisation is no longer able to preserve the data.

The most important way for an organisation to demonstrate conformity with this attribute is by developing and promulgating a digital preservation policy framework that defines the purpose, objectives, scope and desired outcome of an organisation's preservation approach. The digital preservation policy consolidates the organisation's documentation addressing the characteristics of this attribute by explicitly referring to the organisation's relevant policies and related documents, e.g. job descriptions, training plans, submission policies, access policies and policy-development procedures. The development of a compliant digital preservation policy is the focus of Section 4 of this guidance.

Financial Sustainability

The financial sustainability attribute requires an organisation responsible for preserving data to designate adequate resources for implementing and maintaining a data-preservation programme. An organisation should be able to demonstrate that ongoing funding is available and allocated for preserving its data. Additional characteristics of this attribute include the ability of the organisation to demonstrate good business practices through comprehensive documentation and regular audits by outside accounting professionals, to invest limited resources wisely to support the preservation programme, and to seek funding sources to cover data-preservation expenses.

There is currently no specific community document that addresses in detail the cost of preservation. One study of digital preservation challenges noted that:

“Digital preservation is an essentially distributed process including a range of different (and often differently interested) stakeholders who become involved with digital resources at particular phases of their life-cycle. To increase the prospects for digital preservation and reduce the costs, different groups of stakeholders need to become more aware of how their particular involvement with a digital resource ramifies across its life-cycle.”⁷

The report of the US National Science Foundation’s Blue Ribbon Task Force on Sustainable Digital Preservation and Access is expected to provide community-wide recommendations for addressing this attribute. Until the recommendations from the task force can be promulgated within the community, reasonable documentation for this attribute includes appropriate references to data-preservation expenses in an organisation’s budget documentation and narrative reports that discuss data-preservation funding, e.g. the organisation’s annual report.

Technological and Procedural Suitability

The technological and procedural suitability attribute requires an organisation that preserves data to adopt appropriate technology in accordance with community standards, practice and expectations. Technology is

required for the digital preservation programme at many levels to:

- create digital content
- capture content and associated metadata as digital objects
- transfer those objects to and within a digital repository
- process or otherwise interact with digital objects in the repository
- find and deliver stored objects
- build and maintain the repository
- define and implement the policies and protocols that pertain to the repository
- integrate the repository and the digital material into the broader organisational environment and the context in which it operates⁸

This attribute presumes that organisations will comply with the OAIS standard and that adequate and appropriate technology support is in place to manage the data. Additional characteristics of this attribute include a commitment by the organisation to consider and adopt appropriate preservation strategies; ensure there is appropriate infrastructure (e.g. hardware, software, facilities) for the acquisition, storage and access to data over time; establish a technology- management policy for preservation (e.g. procurement, replacement and enhancement of requisite technology, funding for technology investment); comply with relevant standards and best practices supported by adequate technical expertise; and undergo regular external audits of technology components and performance.

Implementing effective preservation strategies is at the core of the technological and procedural suitability attribute. There are three dominant preservation strategies:

- **Migration** typically refers to the conversion of a file format from a less common or obsolete format to a current file format, e.g. converting data from an older

⁷ Arts and Humanities Data Service (AHDS), *A Strategic Policy Framework for Creating and Preserving Digital Collections*, 2001, pg 3.

⁸ Kenney and McGovern, DPM Tutorial, <http://www.icpsr.umich.edu/dpm/dpm-eng/program/techinf.html>

version of an SPSS file format to a newer one or converting a documentation file from one form of PDF to PDF/A, a format defined specifically for preservation.⁹

- **Emulation** refers to the ability of a computer to mimic the capabilities and functionality of older or obsolete computers; this technique might be used to enable a computer to read older file formats, then save them as current file formats (which combines emulation and migration), or to enable a computer in the future to read and use an older, obsolete file format.
- **Normalisation** might refer to limiting file formats for preservation to a small set of common formats (e.g. limiting text files to a version of the Word format or Open Document format), or converting software-dependent file formats (e.g. SPSS system files) to less software-dependent or software-independent file formats (e.g. ASCII or XML-based formats).¹⁰

These strategies were initially developed by the information-technology community to solve common problems caused by technological change. The strategies have been adapted for preservation use and do not yet have community standards defining the protocols for implementing them, and do not have automated protocols for quality control to determine the success of applying one of the strategies.

9 Migration might also be used to refer to migrating content from one storage media to another, e.g. from one computer disk to another when one disk fails, from an off-line storage device to on-line storage for managing multiple copies of digital content. This form of migration is more commonly referred to as media refreshing. File-format migration is a preservation strategy with potentially significant impact because applying the strategy might introduce changes to the content. Media refreshing is a tactic for managing copies of digital content that should not introduce changes to the content. If digital content is stored on storage media that become obsolete, extreme actions (sometimes referred to as digital archaeology) might have to be utilised to rescue the content and save it on current media, but this threat to digital content due to storage media obsolescence occurs outside an effective digital-preservation programme, e.g. before digital content is included in a preservation programme.

10 Discussion of preservation strategies often present the options as either a migration or emulation choice, leaving out normalisation entirely. Some members of the digital preservation community believe that normalisation is not a correct term for this preservation strategy, but no widely-used alternative has been proposed.

Organisations need to identify the type, quantity and risk status (e.g. more at risk due to a decrease in use) of file formats that need to be preserved, and the technology and processes that support them. One prominent data archive had this recommendation regarding their preservation-management approach:

“Routinely review, identify and improve the use of hardware, software, data formats and standards to reduce the risk of storage media deterioration or technological obsolescence... Review and improve policies, plan, and procedures regularly to ensure completeness and coverage for contingencies such as changes in technology, standards and/or project requirements and capabilities, as well as to proactively engage in risk management.”¹¹

The range of file formats in use within the organisation should be consolidated to minimise duplication and eliminate problem formats. As noted above, this process is a form of normalisation. Those formats most at risk, such as those created by obsolete software or by obsolete versions of current software, should be targeted first.

It is common for an organisation to use one or more preservation strategies to adequately preserve its data. ICPSR, for example, uses a combination of migration, to convert data and documentation to accepted formats, and normalisation, to both limit the number of formats that are preserved and convert data and documentation to software-independent formats that are easier to preserve.

The digital preservation community continues to invest significant energy and resources in the development and promulgation of effective digital preservation strategies. For example, the Preservation and Long-term Access through NETWORKED Services (PLANETS) project is a five-year project funded by the EU committed to developing an automated way of matching an appropriate preservation strategy to file formats using the characteristics of the formats and rules defined by an organisation. The PLANETS work is not yet available for production use beyond the project members, but the project is one to watch for useful developments (see Farquhar and Helen Hockx-Yu in Annex B for more information on PLANETS).

11 Centre for International Earth Science Information Network (CIESIN). <http://www.ciesin.columbia.edu/documents/CIESINpreservationpolicy.pdf>

There are several high-profile initiatives in the US, UK and EU that are extending good practice for digital preservation to the management of data. The Digital Curation Centre, for example, has developed the Data Audit Framework, which consists of a four-step process to identify and assess data from a long-term management and use perspective. The Data Information Specialists Committee in the UK has produced guidance for life-cycle management of research data, including a brief section on preservation. Both these documents provide relevant information for preserving data and are listed along with other useful resources in Annex B.

System Security

The system-security attribute requires an organisation that is preserving data to implement adequate controls to appropriately restrict access to data and address potential vulnerabilities in systems that manage and preserve data. Additional characteristics of this attribute include a commitment by the organisation to assure the security of systems used to manage and preserve data; establish policies and procedures to meet security requirements (e.g. policies that address copying data, authentication of system users, firewalls, back-ups, disaster preparedness, emergency response, disaster recovery and training); and stress processes that will detect, avoid and repair loss and document and notify changes and resulting action.

There are yet no community standards for the comprehensive implementation of system security for preservation, but the digital preservation community is able to benefit from the policies and practices of the information-security field. The integrity and security of digital content an organisation preserves need to be assured, and measures to protect the digital content must be comprehensive and well-documented.

Procedural Accountability

The procedural accountability attribute requires an organisation that preserves data to fully document the approval and implementation of decisions, policies, procedures and practices. This documentation serves as the basis for self-assessment, audit and certification of the organisation's preservation programme. Additional characteristics of this attribute include: a commitment by the organisation to enact all relevant policies and procedures for specified tasks and functions and document all preservation practices; establish a means of monitoring systems and procedures to ensure their continued operation; document and provide a rationale

for preservation strategies; and establish the means for feedback from all users of procedures and systems to support problem resolution and negotiate evolving requirements between data producers and data users.

The most important documentation for preservation is the digital preservation policy framework. Section 4 defines the components of and demonstrates the development of a digital preservation policy and Annex D identifies additional policy and procedural documents to serve as examples of creating adequate documentation to meet preservation requirements.

There is yet no community standard providing a specification for digital preservation documentation, although there are increasing examples to use as a starting point for developing an organisation's documentation. Section 4 refers to proposed components of a digital- preservation policy standard that might be used by the community.

The TRAC requirements, introduced in Section 2, are built upon the OAIS and TDR. TRAC provides specific examples of the kinds of documentation needed by an organisation to meet each requirement. A revised version of TRAC is waiting to become an ISO standard -- some time in 2010 if there are no extensive delays. TRAC is further discussed in Section 5.

OAIS compliance

The OAIS compliance attribute requires an organisation that preserves data to commit to designing and implementing its preservation programmes in accordance with the OAIS standard. OAIS compliance may be expressed in very minimal terms using a simple narrative statement of the ways in which an organisation's practice corresponds with components of the OAIS Reference Model (e.g. for ingestion, our organisation has established a policy for accepted file formats; defined quality-assurance measures for content to be preserved; and also defined the elements to be included in each Archival Information Package), or in more meaningful terms by using the OAIS Reference Model to inform and guide the development and enhancement of an organisation's preservation practice.

A challenge for organisations in being OAIS-compliant is to establish a reliable technological infrastructure to sustain an OAIS system in the face of rapid and continuing change. Meeting that challenge requires adequate funding, appropriate expertise,

systematic monitoring of relevant technology and an established process for informed decision-making in technology investment. It is not possible for an organisation to incorporate every technological change, enhancement and upgrade, but it should establish the means of identifying essential changes to implement over time that are needed to sustain the digital preservation programme. Cloud-computing is an option becoming available to organisations as an alternative or supplement to maintaining an internal technological infrastructure. Cloud-computing is a form of technological infrastructure in which a provider of scalable services is made available over the Internet. DuraCloud is a recently-launched provider of cloud-computing services aimed at archives, libraries and museums.¹² A lack of Internet access, legal requirements, jurisdictional constraints on where and how data can be stored and security concerns for confidential data -- all might prohibit organisations from using cloud-computing. But it is becoming a viable and currently-affordable alternative to managing all or some portion of technological infrastructure internally.¹³

Managing a Trusted Digital Repository

To be a trusted digital repository, an organisation must provide continuing evidence that it adheres to community standards and principles and continually demonstrate that it manages its preservation programme in competent and effective ways. Doing so requires an organisation to address each of the seven TDR attributes. Having looked at the individual attributes, an organisation needs to consider the relationship between the attributes as it strives to become a trusted digital repository:

- **OAIS Compliance** is an underlying principle that is an integral part of the other framework components
- **Administrative Responsibility** encompasses all the other components and lays the foundation of a trusted repository; this attribute is influenced by and based upon larger organisational factors (e.g. legislation, mandates)

- **Organisational Viability** encompasses the repository but relies upon aspects of Administrative Responsibility
- **Financial Sustainability** is the most critical of the components within the repository, which cannot exist in its absence; this attribute relates to other financial commitments in the organisation
- **Technological and Procedural Suitability** is the next most essential component within the repository and determines the success of the preservation programme; this attribute should be influenced by and adapted from external experience and practice
- **System Security** is critical to the success of the implementation, but there is known methodology for establishing and maintaining system security; this attribute will be part of larger system-security practices, both within the organisation and externally
- **Procedural Accountability** cuts across and underpins the trusted nature of the repository; this attribute is dictated by and based upon external authorities

The TDR document contains little discussion of and no diagrams to illustrate these relationships between the seven attributes or the relative significance of each, an absence that Figure 4 is intended to fill.¹⁴

First and foremost, an organisation needs to make an explicit commitment to preserving data (administrative responsibility); and, secondly, an organisation needs to have the requisite mandate, training, equipment and other resources (organisational viability) to preserve data effectively. Prior to the release of the TDR document in 2002, the cost of managing digital content (financial sustainability) had not been identified so explicitly for digital preservation. Technology (technological and procedural suitability and system security) was typically the focus of digital preservation discussion, rather than being the means of achieving preservation objectives;

12 DuraCloud is offered by DuraSpace, a partnership of the dSpace Federation and Federa Commons, two repository-software providers. More information about DuraCloud is available at <http://www.duraspace.org/duracloud.php>

13 One group investigating security implications in the cloud is the Cloud Security Alliance. More information is available at <http://www.cloudsecurityalliance.org/>

14 The model was developed at Cornell University Library by Kenney and McGovern for the Digital Preservation Management (DPM) Workshop <http://www.library.cornell.edu/iris/dpworkshop/>, a workshop series funded by the US National Endowment for the Humanities (NEH) <http://www.neh.gov/> since 2003.

and the need to demonstrate through organisational evidence (procedural accountability) the adequacy of an organisation’s operations had not been explicitly required. OAIS compliance is implicit in the attributes and thus not included in the diagram.

The TDR diagram stresses the importance of a strong and well-planned organisational setting for enabling a sustainable approach to long-term preservation, and places technology as an important – but not singular – component within that context. The placement of the attributes in the diagram emphasises that technology should be suited to the scope and requirements of each organisation. The TDR diagram added a *digital archives border* to the TDR attributes because one organisation might maintain more than one repository, in which case the outer layers might be coordinated across the organisation; or a group of organisations might come together to manage one repository, e.g. in a partnership or other collaborative effort, when that is relevant.

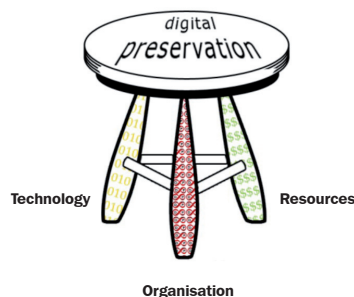
Figure 4 The TDR Attributes Applied to an Organisational Setting



The TDR attributes reflect the three requisite components of a sustained programme for preserving digital content -- organisational infrastructure (“what?”), technological infrastructure (“how?”) and requisite resources (“how much?”) -- as illustrated by the three-legged stool for digital preservation (Figure 5). The organisational leg defines the scope of a digital preservation programme (what data will be preserved?). The technological leg encompasses the approach of the digital preservation programme enabled by technology as it emerges and evolves (how will the data be preserved?). The resources’ leg determines the cost of sustaining digital preservation, based upon the scope and approach (how much will it cost to preserve the data?).

The three-legged stool provides an essential structure for a digital preservation programme. For a programme to be viable and sustainable, the three legs must be equally strong, and therefore balanced, to sustain data over time.

Figure 5 Three-Legged Stool for Digital Preservation¹



1 The three-legged stool concept was developed by McGovern and Kenney to support the DPM workshop series.

4. Formulating a Preservation Policy

The development and implementation of a policy is an essential step to implementing an effective approach to preserving data. A good policy reflects the mandate of the organisation to preserve data, makes an explicit commitment by the organisation to preserve data in accordance with community principles and standards, and fulfils a fundamental requirement of good preservation practice. The process of developing the policy is especially important because it consolidates an organisation's understanding of its digital preservation commitment and brings together the group that will be involved in the process.

Every organisation that preserves data requires a policy, but every organisation should not be required to reinvent the policy's scope. There is no standard or even common format for a preservation policy, but there is an increasing number of policies that provide good examples of formulating one. The outline should produce a digital preservation policy of no more than five pages. It should address the attributes of a Trusted Digital Repository; present a high-level overview of an organisation's digital preservation programme; reflect current not future capabilities of the digital preservation programme; provide links to more detailed and frequently-updated documents, e.g. lower-level policies and procedures; point to the organisation's plan for priorities, time-frames and future development; and document the policy-approval and maintenance process. This section defines the recommended components of such a policy and provides data-specific examples for each component.¹⁵ The policy outline includes the following sections:

OAIS Compliance	[required]
Administrative Responsibility	[required]
Purpose	[required]
Mandate	[required]
Objectives	[desirable]
Organisational Viability	[required]
Scope	[required]
Operating Principles	[optional]
Roles and Responsibilities	[required]
Selection and Acquisition	[desirable]
Access and Use	[desirable]

¹⁵ The policy components were developed for the curriculum of the Digital Preservation Management Workshop (<http://www.icpsr.umich.edu>), then vetted by the more than 500 workshop participants who have attended the workshop since 2003; they represent more than 250 organisations in more than 25 countries on five continents.

Financial Sustainability	[required]
Institutional Commitment	[required]
Cooperation and Collaboration	[optional]
Technological and	
Procedural Suitability	[required]
System Security	[required]
Procedural Accountability	[required]
Audit and Transparency	[required]
Framework Administration	[required]
Definitions and References	[desirable]

OAIS Compliance

The OAIS Compliance section of the policy consists of an explicit statement of the intent of the digital preservation programme to comply with the Open Archival Information System (OAIS) Reference Model. For example:

“In achieving its digital preservation objectives, ICPSR recognises the need to comply with the prevailing standards and practice of the digital preservation community. ICPSR is committed to developing its digital preservation policies, repository and strategies in accordance with the Open Archival Information System (OAIS) Reference Model (ISO 14721:2003). ICPSR tracks and responds to related OAIS initiatives, including developments in digital archives certification, persistent identifiers, preservation metadata and the producer-archive interface. The mapping of ICPSR's preservation process to OAIS is synthesised in Digital Preservation Requirements Applied to ICPSR and explained in greater detail in *OAIS Conformance at ICPSR*.” [ICPSR]

The more detailed plan for making the organisation OAIS-compliant would be contained in a separate document referred to in this section. In addition, this section might refer to the strategic plan for the organisation or the preservation plan, if these documents have been developed.

Administrative Responsibility

The administrative responsibility section of the policy makes an explicit commitment to preserving the data that fall within the scope of responsibility for the organisation and to comply with prevailing standards and practice for digital preservation.

Purpose

The purpose component of the policy makes explicit the intentions of an institution regarding the preservation of data and the role preservation plays in fulfilling the organisation's mission. This component defines the rationale for the policy, identifies responsible parties and stakeholders, indicates the intended audience for the document, and places the document in the context of organisation-wide efforts. The purpose statement might range from broad to narrow, reflecting variations in intention for different types of digital archives. For example:

“This policy statement sets out Statistics New Zealand's approach to the long-term retention and preservation of unit record and aggregate datasets held by Statistics New Zealand and gives guidance on implementing those principles. This statement has been developed from international best practice, in consultation with Statistics New Zealand's subject matter area.” [Statistics New Zealand]

The purpose might refer to the mission statement for the organisation, other high-level policies, and the strategic plan for the organisation, if these documents have been developed.

Mandate

The mandate component of the policy stipulates the authority, jurisdiction or governance based upon which the organisation has undertaken the preservation of the data, e.g. laws, legislation, policies and mission. For example:

“This policy follows from the overarching principle that the statistical datasets created by Statistics New Zealand and other agencies in the Official Statistics System are valuable and irreplaceable, with their utility maximised through ongoing use. The Statement of Principles for the Official Statistics System states this under the maximisation principle as ‘Statistical data are treated as an enduring national resource, with their value increasing through widespread and long-term use’. The utility of official statistics lies in their ability to provide snapshots of society, the economy and the environment, and to show patterns and change over time.” [Statistics New Zealand]

This component might also address requirements not specifically identified as preservation, e.g. legal admissibility, authenticity, Data Protection Acts, copyright legislation, public records acts, E-Government, National Grid for Learning (UK). The mandate component might refer to laws, legislation, contracts, related policies, the organisation's mission statement, regulations or other relevant documents.

Objectives

The objectives' component of the policy states the high-level aims and targets of the organisation for collecting, managing, preserving and sustaining access to its data. This component identifies the benefit of the programme to an institution and its relationship to other objectives, goals and policies. For example:

“The digital preservation function at ICPSR is organised to address these objectives:

- Maintain a comprehensive and responsive digital preservation programme that identifies, acquires, verifies, archives and distributes core social science digital assets
- Adapt preservation strategies to incorporate the capabilities afforded by new and emerging technologies in cost-effective and responsible ways
- Serve the needs of membership organisations by enabling uninterrupted access to digital content over time as the technology for digital content creation and distribution evolves
- Meet the archival requirements of funding agencies and contracting entities committed to the long-term preservation of designated digital content
- Demonstrate auditable compliance with and contribute to the development of the standards and practice of the digital preservation community
- Foster collaborative partnerships with social science researchers and other digital archives to make the best use of available resources and avoid duplicated efforts” [ICPSR]

The objectives' component might refer to the organisation's strategic plan, goals and objectives, budget, preservation plan and technology plans, if these documents have been developed.

Organisational Viability

The organisational viability section of the policy addresses the legal status as well as human and other resources needed to establish and maintain a digital preservation programme. This section might include the following components:

Scope

The scope component of the policy establishes the overall time-frame, levels of responsibility, boundaries, extent, limitations and priorities of the digital preservation programme. This component delineates what the organisation's preservation will include and, as importantly, will not include. The scope statement might be brief or extensive, depending on the nature and scale of the content preserved. For example:

“The policy covers all data captured in the Process, Analysis and Dissemination phases of the statistical process. It covers unit record and aggregate datasets, both published and unpublished. It covers pilot surveys, sampling frames and customised datasets. Statistical datasets created from administrative data and used to create official statistics, or integrated to create official statistics, are also covered by the policy. The treatment of integrated personal data is consistent with the guidelines in the Data Integration Protocol.” [Statistics New Zealand]

The scope component of the policy might refer to the organisation's strategic plans, collection-development policies, preservation plan and definitions of roles, if these documents have been developed.

Operating Principles

The operating principles' component of the policy defines the key principles, models, processes and assumptions upon which the digital preservation programme is developed and implemented. This section is particularly important in establishing system-wide benchmarks for distributed programmes when multiple operational and technical processes are implemented. Common principles include adherence to standards (in particular the OAIS) and other accepted indicators of good practice, support for life-

cycle management, interoperability, evidence-based requirements and preferred methods of preservation. For example:

“The ICPSR digital preservation function operates in accordance with an established set of principles. ICPSR will strive to:

- comply with the OAIS and other digital preservation standards and practice
- ensure that digital content at ICPSR can be provided to users and exchanged with partner and other digital archives so that it remains readable, meaningful and understandable
- participate in the development and promulgation of digital preservation community standards, practice and research-based solutions
- develop a scalable, reliable, sustainable and auditable digital preservation repository
- manage the hardware, software and storage media components of the digital- preservation function in accordance with environmental standards, quality control specifications and security requirements.” [ICPSR]

The operating principles' component might refer to community and organisational standards, documents that define good practice, work-flow and process documents and procedures.

Roles and Responsibilities

The roles and responsibilities' component of the policy describes key stakeholders and their respective roles in digital preservation, including creators, producers, digital repository staff, administrators, financial managers, user groups, advisors, other repositories and collaborators. The “policy should define who will undertake the work and assign responsibility. By connecting policies with current activities, proposed tasks can be aligned with existing work-flows and staff skills. When assigning new responsibilities, changes should be thought through to ensure they are practical and implementable. Defining roles will also be important for accountability.”¹⁶ This section makes an explicit statement that digital preservation is a shared responsibility requiring participants within and beyond the organisation.

¹⁶ Jones, S *DCC Curation Policies Report*. Digital Curation Centre (DCC), 2009. http://www.dcc.ac.uk/docs/reports/DCC_Curation_Policies_Report.pdf

It describes broad categories of roles and responsibilities and cites documents containing more specific descriptions. For example:

“As an organisation acting for its member institutions, funding bodies, and depositors, ICPSR has accepted responsibility for preserving its digital assets. Within ICPSR, the Director, the Digital Preservation Officer, the Computer and Network Services unit, the Collection Development unit, the topical archive managers and the Collection Delivery unit all contribute to the management of the digital preservation function and the life-cycle of digital content at ICPSR. The ICPSR Council, an elected advisory board, evaluates high-level policy documents and reviews programmatic plans and progress. *Roles and Responsibilities for Digital Preservation at ICPSR* provides descriptions of the roles and current assignments.”

The roles and responsibilities’ component might refer to role definitions with explicit specification of responsibilities, documentation of current role assignments, job descriptions and organisational charts.

Selection and Acquisition

The selection and acquisition component of the policy provides the rationale and processes for preserving data based on specific parameters (e.g. formats, types of records, geographic scope). A clear articulation of the data preserved is critical in ensuring that preservation supports the organisation’s mission and priorities and that requisite funding is made available for preserving data. For example:

“The *ICPSR Collection Development Policy* sets forth the priorities and criteria for acquiring digital content. The *ICPSR Deposit Form* reflects the priorities and criteria that are defined in the policy. The *Guide to Social Science Data Preparation and Archiving* provides guidance and templates for depositors to encourage complete and well-documented deposits. The *Collection Development Policy* and the *Guide* are available on the ICPSR website and the related policies are available upon request.” [ICPSR]

The selection and acquisition component might refer to the organisation’s collection-development policy, submission guidelines and the workflow for acquiring data or other digital content, if these documents are relevant or have been developed.

Access and Use

The access and use component of the policy identifies the designated communities for the digital preservation programme and the barriers and/or restrictions to use of the digital content for which the programme is responsible. Specific policies should be developed to further articulate access and use requirements and restrictions. Preservation requires demonstrable proof that an organisation is able to provide meaningful long-term access to the content it is responsible for preserving. For example:

“The designated community at ICPSR, as described by the OAIS, includes traditional users, e.g. social science researchers and graduate students at member institutions; and newer categories of users, e.g. undergraduates, policy-makers, practitioners and journalists. To protect the identity of human subjects who might be represented in the deposited data, ICPSR devotes significant resources to developing and implementing the means to ensure confidentiality.

“ICPSR uses current technology and tools to provide a range of access services. The *ICPSR Data Access Policy* defines the principles and criteria for access to data in the ICPSR collections. ICPSR has developed lower-level policies and procedures to manage access to digital content, including *Procedures for Processing Requests for Restricted Data*, the *Privacy Policy* for handling information about users, *Release Management Procedures* that specifies the preparation of digital content for release, and the *Requests for Permission to Redistribute ICPSR Data* policy that addresses the use of ICPSR digital content by other data archives and distributors.” [ICPSR]

The access and use component might refer to the organisation’s access policy, deposit agreements, digital rights’ management rules and

practice and user agreements, if these documents are relevant or have been developed.

Financial Sustainability

The financial sustainability section of the policy documents the tangible basis for sustaining the digital preservation programme and might include these components:

Institutional Commitment

The institutional commitment component confirms and synthesises the support for the programme and the resources available to sustain digital preservation. For example:

“To sustain its digital preservation function, ICPSR has allocated a portion of its membership support to digital preservation services. In addition, ICPSR continually seeks external research funding to extend its digital preservation scope and capabilities and has secured contracts to fund specific initiatives. Detailed information about digital preservation funding is available in the *ICPSR Annual Report* and in the annual budget of ICPSR.” [ICPSR]

The institutional commitment component might refer to budgets, financial reports, fiscal policies, annual reports, succession plans and contracts.

Cooperation and Collaboration

The cooperation and collaboration component of the policy acknowledges that the organisation’s effort exceeds or will exceed available resources and might not guarantee the safety of all vital assets. This section places digital preservation programmes in a broader context that recognises their dependency on other partners and the community at large. Collaboration and partnership might require formal, legally-binding agreements that delineate the explicit roles and responsibilities of each party. For example:

“The active and collaborative research programme at ICPSR integrates digital preservation requirements and competencies into its priorities and acknowledges digital preservation as a shared community

responsibility. ICPSR has long-standing and emerging partnerships with other data archives, other digital repositories, data producers, and data providers in the United States and internationally for digital preservation cooperation and collaboration.” [ICPSR]

The cooperation and collaboration component might refer to partnership agreements, operating principles and good practice for collaborative projects.

Technological and Procedural Suitability

The technological and procedural suitability component of the policy summarises the preservation approach, strategy and techniques employed by the digital preservation programme to achieve stated objectives. This section states the general philosophy of the digital preservation programme and points to relevant requirements, policies, standards, guidelines and practice. It makes a tangible link with the preservation-planning component of the digital- preservation programme and to the organisation’s preservation plan.

“The majority of digital content in the collections at ICPSR currently consists of social science research data, requisite documentation to use and understand the data, and associated files. Upon receipt of a deposit, ICPSR processes the digital content to ensure that confidential information has not been included in the data; corrects errors; fills gaps in the documentation; and produces distribution versions of the data. Technicians digitise documentation that is received only in hard copy format. ICPSR archives the original digital content received, the normalised versions of processed data and superseded versions of data that have been distributed. The archived files will enable ICPSR to retain the ability to regenerate distribution formats over time. For files submitted on physical storage media, ICPSR makes archival copies but does not preserve the original transmission media. In terms of documentation, ICPSR has been an active participant in the development of and is an adherent to the Data Documentation Initiative (DDI) standard.

“ICPSR has adopted normalisation and migration as its primary digital preservation strategies. Normalisation produces file formats

for data and documentation that are as close as possible to ASCII for text, or TIFF for images, to enable preservation, and reduces the range of file formats to be preserved to ensure that the digital preservation load is manageable. Migration converts digital content to current file formats as software and related technology evolve and copies digital content from older to newer storage media as part of a systematic programme. ICPSR is investigating appropriate preservation strategies for the expanding range of digital content types in its collections. The *ICPSR Digital Preservation Plan* identifies the priorities, objectives, and action plans for the next three years.” [ICPSR]

The technological and procedural suitability component might refer to preservation strategies, preservation plans for the organisation and for specific content, and deposit agreements.

System Security

The system-security component of the policy specifies the organisation’s commitment and approach to ensuring the accuracy, completeness, authenticity, integrity and long-term protection of the organisation’s digital assets.

“The processing procedures for digital content at ICPSR actively address the need for ensuring the accuracy and completeness of digital content through the careful comparison of documentation and data submitted and the generation of metadata and documentation for data. The implementation at ICPSR of an automated deposit form addressed the need for ensuring the authenticity of digital assets by requesting detailed information and signatures for submission. ICPSR ensures the authenticity and integrity of its digital content through the active and ongoing use of checksums from receipt of the digital content onward. In addition, ICPSR conducts periodic reviews and audits of its digital content in archival storage. *Information Security for Digital Assets at ICPSR* provides more detailed information about the principles and good practice for ICPSR’s approach to systems security.

“ICPSR has developed several lower-level policy and procedural documents that address

specific aspects of the long-term protection of its digital assets. For example, Secure Package Handling Procedures at ICPSR stipulates the procedure for receiving and sending physical packages containing sensitive data; Procedures for Processing Requests for Restricted Data stipulates the steps for responding to requests for restricted data; and ICPSR Policy on Receiving Data with Direct Identifiers defines appropriate actions when data received contain direct identifiers. *Disaster Planning for Digital Assets at ICPSR* provides a set of principles ICPSR has embraced for protecting digital assets within the broader context of disaster planning for ICPSR.” [ICPSR]

The system security component might refer to security policies and procedures or to the organisation’s disaster-planning documentation. Disaster planning is an essential component of digital preservation. Examples of the policies and documentation needed by an organisation are available on the ICPSR website (<http://www.icpsr.umich.edu/icpsrweb/ICPSR/curation/disaster/index.jsp>) and Annex F contains additional information on disaster planning.

Procedural Accountability

The procedural accountability section of the policy acknowledges the need and stipulates the means for ensuring the transparency and accountability of the digital preservation programme’s policies and operations. This section might include these components:

Audit and Transparency

The audit and transparency component of the policy explicitly commits the organisation to periodic self-assessment and audit to evaluate, measure and adjust the policies, procedures, preservation approaches and practices of the digital preservation programme. Transparency enables self-assessment and audit. Self-assessment and audit improve internal operations, facilitate external reviews and contribute to the development of effective partnerships and collaboration.

“ICPSR participated as a test audit in the Certification of Digital Archives research project conducted by the Centre for Research Libraries. ICPSR is committed to a two-year cycle of

self-assessment and a five-year audit cycle to evaluate, measure and adjust the policies, procedures, preservation approaches and practices of the digital preservation function. A complete set of current policies is available on the ICPSR web-site and procedural documents are available upon request.” [ICPSR]

The audit and transparency component might refer to audit and self-assessment schedules and results, strategic plans and preservation plans.

Framework Administration

The framework administration component of the policy describes the organisation’s policies and practice pertaining to the development, approval and maintenance of the policy framework over time, e.g. frequency of updates and reviews, maintenance roles and expiration dates. The framework has little value if it has not received the appropriate approval and has not been implemented. At a minimum, the date and source of approval and the review cycle should be provided.

“The digital preservation policy framework was completed in April 2007; approved by the ICPSR Directors Group on May 1, 2007; and approved by the ICPSR Council at the June 2007 meeting. The March 2006 predecessor document, ICPSR Preservation Policy, served as a starting point for developing the framework. ICPSR will review the framework every two years to ensure that it remains current and comprehensive as the digital preservation function at ICPSR evolves.” [ICPSR]

The framework administration component might refer to the organisation’s policy administration procedures and policy-approval documentation.

Definitions

The definitions’ component of the policy identifies terms and concepts that might be needed to understand the framework and could be instrumental in strategies for securing institutional commitment. This is an optional section but one that can be very important. It is particularly important to include legally-required and other mandated terminology and definitions.

The section might either provide or point to requisite definitions.

“*Digital Preservation and Data Archiving Terms for ICPSR* provides a current set of definitions of terms used in this digital preservation policy framework and employed by the digital preservation function at ICPSR.” [ICPSR]

The definitions’ component might refer to definitions developed by the organisation and glossaries adopted by it.

References

The references’ component of the policy provides citations for or pointers to key resources that informed the development and application of the framework. This section identifies more detailed documents, both internal and external, that provide a deeper expression of the mission, underlying principles, illustrative processes and sustaining roles. It might contain citations for these documents or point to a current list of relevant community standards and guidance.

“*Citations for Digital Preservation at ICPSR* provides a list of ICPSR, digital preservation community and other documents that have informed the development and maintenance of the framework.” [ICPSR]

The references component cited resources, community lists of standards and practice.

Developing the Policy

This section emphasises the process of developing the policy because it not only fulfils a core requirement of preservation but also provides a focal point for establishing the organisation’s data-preservation initiative. The importance of a preservation policy cannot be overemphasised. If the organisation carefully considers the mandatory, desirable and optional policy components and develops a policy that reflects the priorities and context of the organisation, the result will be a solid foundation on which the organisation can build and sustain its data-preservation programme.

There are a number of factors for the organisation to consider in initiating and completing the policy-development process, including:

- Who should lead and coordinate the development of the policy?
- Who should be involved in developing the policy, both internally and externally?
- Who needs to be aware of the policy initiative before, during and after it is developed?
- Who will need to approve the policy when it is completed?
- How will the organisation make the policy available and to whom?

In developing the policy, the ideal is for one or two people to draft the policy document or sections of it, then share it with a larger group for feedback and suggestions. Attempting to actually write the policy as a group is often not successful and adds to the time required to develop the policy. When the group convenes, some effort should be devoted to agreeing

on the time-frame for developing the policy, specific milestones to be met and the roles of the group in developing the policy.

It is worth endeavouring to develop a sound and comprehensive preservation policy because it is essential documentation to demonstrate that an organisation is engaged in preserving data, and it enables the organisation to, at least in part, address a number of the requirements to provide evidence of its preservation action. The policy-development process should be explicit, intentional and transparent within the organisation, because a key benefit of the process is raising awareness of preservation, its importance and its requirements. The policy-development process is not complete until the policy is finalised, approved and implemented. It is necessary for an organisation not only to have a preservation policy but also to demonstrate that it is acting in accordance with it. The policy – both the process to create it and the completion of that process – is instrumental in achieving good digital preservation practice.

5. Evaluation and Audit for Data Preservation

This guidance defines the expectations for organisations that preserve data. This section concludes the guidance by considering two approaches an organisation might consider as alternatives in demonstrating good practice: the Data Seal of Approval, which is particularly useful for organisations that are data archives, and the Characteristics of a Trusted Digital Repository, which is useful for any organisation responsible for preserving data. This section also defines a set of stages that have proven useful to organisations in achieving incremental but continuing progress in developing their preservation approach.

As a starting point for this final portion of the discussion, it is useful to provide some context for the emergence of performance measures for digital repositories. Both the 1996 *Preserving Digital Information* report and the OAIS Reference Model identified the need for digital archives to demonstrate compliance with standards and practice through a commitment to ongoing audit and certification. To address this need, the Research Libraries Group (RLG) and the US National Archives and Records Administration convened a working group that produced the *Trustworthy Repositories Audit & Certification (TRAC): Criteria and Checklist* in 2007. In early-2007, the certification of digital archives became the focus of an international working group to develop an ISO standard. The ISO working group used TRAC as a starting point for its work. The work on the certification standard is also informed by the Digital Repository Audit Method Based on the Risk Assessment (DRAMBORA) toolkit, it is a risk-management approach for assessing digital collections, developed by the Digital Curation Centre and Digital Preservation Europe (DPE), and also embracing the work of the Nestor project in Germany, which produced a catalogue of criteria for a digital archive audit and a coaching approach for assisting smaller organisations in addressing such criteria.

The TRAC requirements have been a focal point for the community and formed the basis of most community efforts in audit and certification. The requirements are, however, extensive and might prove daunting to organisations just starting to develop a data-preservation approach. Therefore, this section presents two more concise options. Both options presume that organisations adhere to both TDR and OAIS. Each of the audit and certification approaches builds upon the accumulated good practice of the community, as

defined in sections 2, 3 and 4 of this guidance. These sections provided examples and descriptions of what an organisation is expected to do to meet community requirements as defined by TDR and the OAIS. Organisations are encouraged to read and refer to the TRAC document as a useful resource, even if they find one of the suggested options to be more helpful as a beginners' guide.

Option 1: Data Seal of Approval

The first community framework that might aid organisations in developing their data-preservation approach is the *Data Seal of Approval*, a concise set of higher-level principles that defines expectations for producers of digital content, repositories that manage digital content and consumers of digital content. The principles were proposed by the National Data Archives in the Netherlands (DANS) in 2008. The principles reflect the fundamentals of both TDR and OAIS and represent a workable approach to digital-content management. The Data Seal of Approval has since been accepted for international use by data archives and the purpose of the principles is being extended to apply to digital content of any kind. There is an increasing community expectation that data archives will adhere to the principles and seek the Data Seal of Approval. The principles are summarised here. The principles' aim is to be simple and straightforward, so no additional discussion of them is provided. Organisations should refer to the recommendations provided in preceding sections and review the guidelines and assessment process provided for the Data Seal of Approval at <http://www.datasealofapproval.org/>

The data-producer deposits the research data in a data repository with sufficient information for others to assess the scientific and scholarly quality of the research data and compliance with disciplinary and ethical norms.

1. The data-producer provides the research data in formats recommended by the data repository.
2. The data-producer provides the research data together with the metadata requested by the data repository.
3. The data repository has an explicit mission in the area of digital archiving and promulgates it.

4. The data repository uses due diligence to ensure compliance with legal regulations and contracts including, when applicable, regulations governing the protection of people.
5. The data repository applies documented processes and procedures to managing data storage.
6. The data repository has a plan for long-term preservation of its digital assets.
7. Archiving takes place according to explicit work-flows across the data life-cycle
8. The data repository assumes responsibility from the data producers for access and availability of the digital objects.
9. The data repository enables the users to utilise the research data and refer to them.
10. The data repository ensures the integrity of the digital objects and the metadata.
11. The data repository ensures the authenticity of the digital objects and the metadata
12. The technical infrastructure explicitly supports the tasks and functions described in internationally-accepted archival standards like the OAIS.
13. The data consumer complies with access regulations set by the data repository.
14. The data consumer conforms to and agrees with any codes of conduct that are generally accepted in higher education and research for the exchange and proper use of knowledge and information.
15. The data consumer respects the applicable licences of the data repository regarding the use of the research data.

Option 2: Ten Characteristics of a Digital Preservation Repository

These characteristics were agreed at a meeting convened just before the start of the ISO working group in 2007. The meeting included representatives of the major international audit and certification initiatives

for digital archives: the Digital Curation Centre, DigitalPreservationEurope and Nestor met at the Center for Research Libraries in Chicago to consider core criteria based on this cumulative set of audit and certification developments. The purpose of the meeting was to compare the progress of the groups' work to that point and avoid redundant or conflicting results that might complicate the process of defining good practice for audit and certification of digital archives or confuse the community. They identified these ten basic characteristics of digital preservation repositories; these provide a convenient and manageable set of criteria for an organisation to address.

1. ***The repository commits to continuing maintenance of digital objects for identified community/communities.*** This characteristic represents an organisation's fundamental commitment to preserving its digital content over time.
2. ***The repository demonstrates organisational fitness (including financial, staffing structure and processes) to fulfil its commitment.*** This characteristic equates to demonstrating the organisational viability to preserve digital content.
3. ***The repository acquires and maintains requisite contractual and legal rights and fulfils responsibilities.*** This characteristic addresses the need to take full responsibility for the digital content that is preserved.
4. ***The repository has an effective and efficient policy framework.*** This characteristic reflects the need to document and share the intention of the organisation to preserve digital content.
5. ***The repository acquires and ingests digital objects based upon stated criteria that correspond to its commitments and capabilities.*** This characteristic addresses the need to accurately define and maintain the scope of preservation by an organisation.
6. ***The repository maintains/ensures the integrity, authenticity and usability of digital objects it holds over time.*** This

characteristic reflects the essential purpose of preservation. At a minimum, an organisation needs to have control over its digital content (know where it is and who has or is able to gain access to it), associate minimal metadata data with the content and create and control more than one copy of the digital content.

7. ***The repository creates and maintains requisite metadata about actions taken on digital objects during preservation as well as about the relevant production, access support, and usage process contexts before preservation.*** This characteristic reflects the development of an appropriate preservation approach for the scope and content to be preserved.
8. ***The repository fulfils requisite dissemination requirements.*** This characteristic ensures that appropriate access is provided to preserved digital content and that intended users are aware of access policies and practices.
9. ***The repository has a strategic programme for preservation planning and action.*** This characteristic acknowledges the need for organisations to plan for the impact of technological change, in order to avoid loss.
10. ***The repository has technical infrastructure adequate to continuing maintenance and security of its digital objects.*** This characteristic refers to the need for each organisation to adopt technology suited to their requirements and resources for storing media used for preservation, storage options and file formats preserved.

The purpose of these characteristics is to provide a basic guide for repositories that are preserving digital content, regardless of type and size. The principles acknowledge that an organisation's preservation approach will be developed at an appropriate scale to address relevant requirements and resources.

Building Capacity and Measuring Progress

Every organisation should engage in a thorough self-assessment process using one or more of the community frameworks to identify its strengths and

areas for development. Some organisations might also need or want to undergo an external audit, either an informal process with a peer organisation or a more formal one with a recognised authority, if there is one for the organisation. Certification of digital archives is only known as a current option for organisations in Germany, but the efforts of the ISO working group might result in broader certification options. In the meantime, organisations have the set of self-assessment and audit tools developed by the community and identified in this guidance.

None of the audit and certification initiatives presumes that an organisation is able to meet all the requirements the first time it completes a self-assessment or audit process. Organisations all go through stages of incremental development as they continue to progress their capacity to achieve preservation objectives. Five stages of development for digital preservation have been identified and formalised:¹⁷

Stage 1. Acknowledge: understanding that digital preservation is a local concern—an organisation acknowledges and determines to address preservation

Stage 2. Act: initiating digital preservation projects—an organisation takes action to address preservation, often through a project

Stage 3. Consolidate: moving smoothly from projects to programmes—an organisation recognises that projects are limited so develops a basic programme

Stage 4. Institutionalise: incorporating the larger environment and rationalising programmes—an organisation formalises the preservation programme as a mainstream initiative

Stage 5. Externalise: embracing inter-institutional collaboration and dependency—an organisation might choose to partner and collaborate with others

These stages provide a means for the organisation to identify next steps for *developing* its digital-

17 Anne R Kenney and Nancy Y McGovern, "The Five Organisational Stages of Digital Preservation", in *Digital Libraries: A Vision for the Twenty First Century*, a festschrift to honour Wendy Lougee, 2003.

preservation programme, to provide a way of *communicating* progress and plans with peers and others in the community, and to enable *measuring* progress towards goals.

Organisations that preserve data can learn several lessons from the audit and certification developments in the digital preservation community to date. Every organisation responsible for preserving data should regularly employ self-assessment to demonstrate alignment with good practice, document progress towards objectives, and identify priorities for further development. Changes within or outside the organisation might require it to revisit its preservation programme status, for example:

- a change in leadership in the organisation, especially a change that removes an advocate for preservation
- a major shift in technology (e.g. the impact of the World-Wide Web on computing and communication practice)
- a change in accepted preservation practice (e.g. the shift from off-line storage of preservation copies to on-line management of preservation copies)

Although there is not yet a comprehensive set of standards for digital preservation, there is a known and growing base of good practice to guide an organisation's efforts. As with the development of a preservation policy, organisations should use the opportunity of self-assessment via a working group process. Some organisations, especially those with a national or international mandate for managing data over time, have community examples to follow in conducting an audit that will demonstrate adherence to good practice. Any organisation responsible for preserving data should be able to use the framework of good practice outlined in this guidance to develop a better understanding of digital preservation, and use that understanding to establish a system that meets their needs, requirements and scope and nature of data to be preserved. There is no single approach to preserving data but there is a range of documents and examples to guide organisations to good practice.

Annexes

A. Glossary

B. References

C. Milestones for Data Preservation

D. Programme and Policy Examples

E. Survey of Institutional Readiness

F. Technological Obsolescence and Other Threats

Annex A. Glossary

This glossary defines terms used throughout the guidance and provides a stand-alone reference to relevant terms and concepts for ongoing use by readers.

Access: The act of making information available. Digital preservation ensures long-term access to digital content. The OAIS requires that an archive should supply and deliver digital content to authorised users (consumers); delivery might be to an individual or to an access-delivery system.

Source: OAIS: CCSDS 650.0-B-1, p. 1-7.

Administration: The OAIS entity that coordinates and controls day-to-day operations of an OAIS system. The OAIS identifies the policies and other documents that are the responsibility of Administration and are required by an OAIS.

Source: OAIS: CCSDS 650.0-B-1, p. 1-7.

Archival Information Collection (AIC): An AIC is an Archival Information Package (AIP) that represents an aggregation of other Archival Information Packages.

Source: OAIS: CCSDS 650.0-B-1, p. 1-7.

Archival Information Package (AIP): An AIP is an Information Package composed of Content Information and related Preservation Description Information (PDI). An AIP is a core concept of OAIS and the basis for storing and preserving digital content. There is yet no standard package, but active research and development within the digital preservation community has produced examples for organisations to use.

Source: OAIS: CCSDS 650.0-B-1, p.1-7.

Archival Storage: The OAIS entity responsible for the storage and retrieval of Archival Information Packages.

Source: OAIS: CCSDS 650.0-B-1, p. 1-8.

Archive: An organisation that identifies information it is committed to preserving; takes responsibility for that content for a specified period of time; uses approaches that ensure the preserved content remains

meaningful to users; and makes the content available to authorised users. Most archives now commit themselves to designing their digital repositories in accordance with the OAIS Reference Model.

Source: ICPSR.

Checksum: “A checksum or hash sum is a fixed-size datum computed from an arbitrary block of digital data for the purpose of detecting accidental errors that may have been introduced during its transmission or storage. The integrity of the data can be checked at any later time by re-computing the checksum and comparing it with the stored one. If the checksums do not match, the data were certainly altered (either intentionally or unintentionally).”

Source: <http://en.wikipedia.org/wiki/Checksum>

Common Services: The OAIS services that include the operating system, network services and security services.

Source: OAIS: CCSDS 650.0-B-1, p. 1-8

Consumer: The OAIS role that refers to individuals, organisations or other systems that interact with an OAIS system to retrieve and use preserved content in that system.

Source: OAIS: CCSDS 650.0-B-1, p. 1-8

Content Information: Content Information comprises a Data Object (one or more files or data-streams) and its Representation Information (information that allows the object to be rendered by a computer). For example, Content Information could be a single table of numbers representing the responses to a survey but excluding the documentation that would explain its history and origin, how it relates to other surveys, etc.

Source: Trusted Digital Repositories: Attributes and Responsibility

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

Source: Trusted Digital Repositories: Attributes and Responsibility

Data Management: The OAIS entity that captures and coordinates the information required to operate the OAIS system, including information on individual objects, aggregations of objects and the repository system itself. Data Management receives updates from other functions and provides reports to other functions in response to requests.

Source: OAIS: CCSDS 650.0-B-1, p. 1-9.

Designated Community: An OAIS concept describing the constituency for which the archived information should be relevant and understandable.

Source: ICPSR

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.” An archive works with Consumers over time to ensure DIPs remain useful.

Source: OAIS: CCSDS 650.0-B-1, p. 1-10.

Information Package: Content Information and associated Preservation Description Information needed to aid preservation of Content Information. The Information Package has associated Packaging Information used to specify the Content Information and Preservation Description Information.

Source: Trusted Digital Repositories: Attributes and Responsibilities.

Ingest: The OAIS entity that accepts Submission Information Packages from Producers; ensures the quality of the content adheres to established criteria; generates the Archival Information Packages for storage; and confirms that AIPs and additional information to find and use the AIPs (Descriptive Information) are received and secured by Archival Storage.

Source: OAIS: CCSDS 650.0-B-1, p. 1-11.

Management: The OAIS role referring to individuals or organisations that establish policies and approve the budget for an OAIS system.

Source: OAIS: CCSDS 650.0-B-1, p. 1-11.

OAIS: The Open Archive Information System (OAIS) Reference Model, an ISO standard that formally expresses the roles (producer, management, consumer and, implicitly, archives), functions (common services, ingest, archival storage, data management, administration, preservation planning and access), and content (submission information package, archival information collection, archival information package and dissemination information package) of an archive. It was approved as an ISO standard in 2003. OAIS underwent a five-year review beginning in 2007.

Source: The Consultative Committee for Space Data Systems (CCSDS) of NASA coordinated the development of the OAIS. Its final version, produced by CCSDS before the OAIS became an ISO standard was CCSDS 650.0-B-1. This is the version often cited and referenced because it is in the public domain. The OAIS is an extensive document full of descriptions and examples of additional concepts. It was accepted by the International Standards Organisation in 2003 (ISO 14721: 2003) and updated in 2009.

Persistent Identifier: A persistent identifier refers to enduring identifiers that allow digital objects to be consistently and uniquely referred to over time. For preservation purposes, over time refers to the life of the object.

Notes: often, persistent identifiers are discussed in relation to resources that are available via the World Wide Web, but even off-line digital objects need to have persistent identifiers for preservation purposes. There are a number of options for implementing persistent identifiers – this article is one concise source that reviews available options.

Source: <http://www.ariadne.ac.uk/issue56/tonkin/>

Preservation Planning: The OAIS entity that monitors the environment of the OAIS to ensure it is able to be maintained across generations of technology and provides recommendations to Administration regarding standards and policies; the implication of technological developments; migration plans; and other strategies for maintaining preserved content, information packaging appropriate to requirement and content types, test plans and associated prototypes for implementing recommendations.

Source: OAIS: CCSDS 650.0-B-1, p. 4-2.

Producer: The OAI role that refers to individuals, organisations and systems that provide information to be preserved by the OAI.

Source: OAI: CCSDS 650.0-B-1, p. 1-12.

Submission Information Package (SIP): An Information Package provided by the Producer to the OAI for long-term management.

Source: OAI: CCSDS 650.0-B-1, p. 1-13.

Annex B. References

This set of references identifies sources used in developing the guidance, sources for additional information on digital preservation and specific topics within the scope of the guidance, and sources of information for ongoing community updates.

Good Practice – Data-specific or related

Atkins, D E. *Revolutionizing Science and Engineering through Cyberinfrastructure*. US **National Science Foundation (NSF)**. 2003. <http://www.nsf.gov/od/oci/reports/atkins.pdf>

Beagrie, N, Chruszcz, J and Lavoie, B. *Keeping Research Data Safe: A Cost Model and Guidance for UK Universities*. **Joint Information Systems Committee (JISC)**. 2008. <http://www.jisc.ac.uk/media/documents/publications/keepingresearchdatasafe0408.pdf>

CUGIR Data Management and Distribution Policy. **Cornell University Geospatial Information Repository (CUGIR)**, 2006. <http://cugir.mannlib.cornell.edu/CugirDataMgmtPolicy.20060828.pdf>

Data Archive Policy. **US National Radio Astronomy Observatory (NRAO)**. Undated. <http://www.nrao.edu/admin/do/dataarchive.shtml>

Data Archiving and Networked Services (DANS, Netherlands). *Data Seal of Approval*. Undated. http://www.datasealofapproval.org/sites/default/files/Data_Seal_of_Approval_1-4.pdf

Data Archiving Policy. **Statistics New Zealand**. Undated. http://www.stats.govt.nz/about_us/policies-and-guidelines/data-archiving-policy.aspx

Data Audit Framework. Digital Curation Centre (DCC, UK). 2009. <http://www.dcc.ac.uk/tools/daf/>

Data Bank Role. **Greek Social Data Bank (GSDB)**. Undated. http://www.gsdb.gr/databank_role_en.html

Data Documentation Initiative (DDI). <http://www.icpsr.umich.edu/DDI/>

Data Management and Communications Plan for Research and Operational Integrated Ocean Observing Systems. **National Office for Integrated and Sustained Ocean**

Observations Ocean Report. 2005. http://dmac.ocean.us/dacsc/docs/march2005_dmac_plan/dmac_partI_3.15.05.pdf

Data Management Policy. **Rural Economy and Land Use (RELU, UK)**. 2006. <http://www.relu.ac.uk/about/data.htm>

Data Policy. **Earth Observing Laboratory (EOL)**. Undated. <http://www.eol.ucar.edu/data/data-policy>

Data Policy. **Economic and Social Research Council (ESRC, UK)**. 2001. http://www.esrcsocietytoday.ac.uk/ESRCInfoCentre/Images/DataPolicy2000_tcm6-12051.pdf

The Dissemination of Government Geographic Data in Canada: Guide to Best Practices. **Geo Connections**. 2008. http://www.geoconnections.org/publications/Best_practices_guide/Guide_to_Best_Practices_Summer_2008_Final_EN.pdf

Dusa, A. "A Data Archive for Social Sciences in Romania." *IASSIST Quarterly*. **Summer** 2001. <http://iassistdata.org/publications/iq/iq25/iqvol252dusa.pdf>

Establishment of a National Policy to Archive Surface-Geophysical Data Memo. **United States Geological Survey (USGS)**. 2009. <http://water.usgs.gov/admin/memo/GW/gw09.02.html>

Gold, A. "Cyberinfrastructure, Data, and Libraries: Part 1 and Part 2 - A Cyberinfrastructure Primer for Librarians." *D-Lib Magazine*. *September/October 2007*. <http://www.dlib.org/dlib/september07/gold/>

Govett, M W, M Doney and P Hyder. *The Grid: An IT Infrastructure for NOAA in the 21st Century*. **National Oceanic and Atmospheric Administration (NOAA)**. 2004. <http://nomads.ncdc.noaa.gov/docs/NOAAGRIDcomputing.pdf>

Guide for IOSS Data Providers. **Integrated Ocean Observing System (IOOS)**. 2006. http://www.sccoos.org/interactive/dmac/bestpractices/Guide_for_IOOS_Data_Providers_060206.doc

- ICPSR Guide to Social Science Data Preparation and Archiving Guide: Best Practice Throughout the Data-Like Cycle* (4th ed). **Inter-university Consortium for Political and Social Research (ICPSR)**. 2009. <http://www.icpsr.umich.edu/icpsrweb/ICPSR/access/dataprep.pdf>
- ICPSR Digital Preservation Policy Framework*. **Inter-university Consortium for Political and Social Research (ICPSR)**. 2007. <http://www.icpsr.umich.edu/DP/policies/dpp-framework.html>
- Jones, S and M Donnelly. *Data Management Plan Template*. **Digital Curation Centre (DCC)**. 2009. http://www.dcc.ac.uk/docs/templates/DMP_checklist.pdf
- Jones, S, S Ross and R Ruusalepp. *The Data Audit Framework: A Toolkit to Identify Research Assets and Improve Data Management in Research-led Institutions*. Presentation at **iPRES**, 2008. http://www.data-audit.eu/docs/DAF_iPRES_paper.pdf
- Kanyengo, C. *Managing Digital Information Resources in Africa: Preserving the Integrity of Scholarship*. **University of Zambia**. 2006. <http://www.ascleiden.nl/pdf/elecpubliconfkanyengo.pdf>
- Krejci, J. "The Czech Sociological Data Archive." *IASSIST Quarterly*. **Summer** 2001. <http://iassistdata.org/publications/iq/iq25/iqvol252krejci.pdf>
- Laaksonen, H, S Borg and J Stebe. "Setting up Acquisition Policies for a New Data Archive." *IASSIST Quarterly*. **Spring** 2006. <http://www.iassistdata.org/publications/iq/iq30/iqvol301laaksonen.pdf>
- Lesaoana, M A. "Data Archiving in Africa: The South African Experience." *IASSIST Quarterly*. **Winter** 1998. <http://iassistdata.org/publications/iq/iq21/iqvol211lesaoana.pdf>
- Murakas, R and A Rämmer. "Estonian Social Science Data Archives: Past and Future Perspectives." *IASSIST Quarterly*. **Summer** 2001. <http://iassistdata.org/publications/iq/iq25/iqvol252murakas.pdf>
- National Research Council (NRC)**. *Environmental Data Management at NOAA: Archiving, Stewardship, and Access*. 2007. http://www.nap.edu/catalog.php?record_id=12017#toc
- National Science Board (NSB)**. *Long-Lived Digital Data Collections Enabling Research and Education in the 21st Century*. **National Science Foundation (NSF, US)**. 2005. <http://www.nsf.gov/pubs/2005/nsb0540/nsb0540.pdf>
- National Science Foundation Cyberinfrastructure Council. *Cyberinfrastructure Vision for 21st Century Discover*. **National Science Foundation (NSF)**. 2007. http://www.nsf.gov/od/oci/ci_v5.pdf
- NewDISS - A 6- to 10-Year Approach to Data Systems and Services for NASA's Earth Science Enterprise*. **National Aeronautics and Space Administration (NASA)**. 2002. http://esdswg.eosdis.nasa.gov/pdf/ND_Reprt.pdf
- Paris, J. "Collaboration In Africa." *IASSIST Quarterly*. **Summer** 2002. <http://iassistdata.org/publications/iq/iq26/iqvol262paris.pdf>
- Preston, J and S Moses. *A Public Digital Repository for Jamaica*. *eJamaica.org*. 2007. <http://pubs.or08.ecs.soton.ac.uk/121/3/ejamaica2.pdf>
- Policy-making for Research Data in Repositories: A Guide*. Data Information Specialists Committee UK. 2009. <http://www.disc-uk.org/docs/guide.pdf>
- Rural Economy and Land Use: Guidance Data Management*. **Economic and Social Data Service**. 2006. <http://www.data-archive.ac.uk/relu/reluaug2006.pdf>
- Stebe, J and I Vipavc. "The Social Science Data Archive in Slovenia." *IASSIST Quarterly*. **Summer** 2001. <http://iassistdata.org/publications/iq/iq25/iqvol252stebe.pdf>
- Stewardship of digital research data: a framework of principles and guidelines*. January 2008: <http://www.rin.ac.uk/our-work/data-management-and-curation/stewardship-digital-research-data-principles-and-guidelines>
- A Strategic Policy Framework for Creating and Preserving Digital Collection*. **Arts and Humanities Data Service (AHDS)**. 2001. <http://ahds.ac.uk/strategic.pdf>
- Treloar, A, D Groenewegen and C Harboe-Ree. "The Data Curation Continuum." *D-Lib Magazine*. 2007. <http://www.dlib.org/dlib/september07/treloar/09treloar.html>
- Woolfrey, L. "The Establishment of the African Association of Statistical Data Archivists (AASDA)." *IASSIST Quarterly*. **Summer** 2007. <http://www.iassistdata.org/publications/iq/iq31/iqvol312woolfrey.pdf>
- Woolfrey, L. "A Survey Data Archive Network in Africa - Possibilities and Practicalities." *IASSIST Quarterly*. 2007. <http://www.iassistdata.org/publications/iq/iq31/iqvol311woolfrey.pdf>
- Woollard, M. *UK Data Archive Preservation Policy*. **United Kingdom Data Archive**. 2009. <http://www.data-archive.ac.uk/news/publications/preservationpolicy.pdf>

Good Practice – relevant, but not data-specific

- Beagrie, N. "Digital Preservation: Best Practice and its Dissemination," *Ariadne* (43). 2005. <http://www.ariadne.ac.uk/issue43/beagrie/>
- British Library Digital Preservation Strategy*. **British Library**. 2006. <http://www.bl.uk/aboutus/stratpolprog/ccare/introduction/digital/digpresstrat.pdf>
- Consultative Committee for Space Data Systems, *Reference Model for an Open Archival Information System (OAIS)*, draft recommended standard. October 2009. <http://public.ccsds.org/sites/cwe/rids/Lists/CCSDS%206500P11/CCSDSAgency.aspx>
- Core Requirements: Ten Basic Characteristics of Digital Preservation Repositories. Centre for Research Libraries (CRL). 2007. <http://www.crl.edu/archiving-preservation/digital-archives/metrics-assessing-and-certifying/core-re>
- Digital Repository Audit Method Based on Risk Assessment (DRAMBORA)*. Digital Curation Centre and Digital Preservation Europe (DPE). Revisions 2006-2009. <http://www.dcc.ac.uk/tools/drambora/>
- Farquhar, A and H Hockx-Yu. Planets: Integrated Services for Digital Preservation. *International Journal of Digital Curation* 2.2 (2007): 88-99. <http://www.ijdc.net/ijdc/article/view/46/59>
- Guidelines for the Preservation of Digital Heritage*. **National Library of Australia**. 2003. <http://unesdoc.unesco.org/images/0013/001300/130071e.pdf>
- Invest to Save: Report and Recommendations of the NSF-DELOS Working Group on Digital Archiving and Preservation*. 2003. <http://delos-noe.iei.pi.cnr.it/activities/internationalforum/Joint-WGs/digitalarchiving/Digitalarchiving.pdf>
- It's About Time, a joint report by NSF and LC on research challenges in digital archiving*. 2003. http://www.digitalpreservation.gov/library/resources/pubs/docs/about_time2003.pdf
- Jones, M, and N Beagrie. *Preservation Management of Digital Materials: A Handbook*. First edition 2001, current version maintained by the Digital Preservation Coalition. <http://www.dpconline.org/graphics/handbook/index.html>
- Jones, S. *DCC Curation Policies Report*. **Digital Curation Centre (DCC)**. 2009. http://www.dcc.ac.uk/docs/reports/DCC_Curation_Policies_Report.pdf
- Keakopa, S M *Trends in Long-Term Preservation of Digital Information: Challenges and Possible Solutions for Africa*. **University of Botswana**. 2008. http://www.codesria.org/Links/conferences/el_public8_eng/segomotso_keakopa.pdf
- LIFE: Life Cycle Information for E-Literature <http://www.life.ac.uk/>
- NSF Blue Ribbon Task Force on Sustainable Digital Preservation and Access, *Interim Report*. December 2008. <http://brtf.sdsc.edu/bibliography.html>
- National Library of Australia. *Preserving Access to Digital Information (PADI) since 1995*. <http://www.nla.gov.au/padi/>
- Parse Insight Deliverable D2.1. Draft Road Map. Permanent Access to the Records of Science in Europe (PARSE) Insight*. 2009. http://www.parse-insight.eu/downloads/PARSE-Insight_D2-1_DraftRoadmap_v1-1_final.pdf
- Peters, D "DISA: Insights of an African Model for Digital Library Development." *D-Lib Magazine*. 2001. <http://www.dlib.org/dlib/november01/peters/11peters.html>
- Preservation Metadata Implementation Strategies (PREMIS). <http://www.loc.gov/standards/premis/>
- Preserving Digital Information: Report of the Task Force on Archiving of Digital Information. Commission on Preservation and Access and The Research Libraries Group*. 1996. <http://www.clir.org/pubs/reports/pub63watersgarrett.pdf>
- RLG-OCLC, Trusted Digital Repositories: Attributes and Responsibilities, An RLG-OCLC Report. May 2002*.
- Resource Centre. Digital Curation Centre*. <http://www.dcc.ac.uk/resource/>
- Ribes, D, et al. "The Long Now of Technology Infrastructure: Articulating Tensions in Development." **National Science Foundation (NSF)**, *Journal of the Association for Information Systems (JAIS)* 10 (5). 2009.
- Suleman, H. *An African Perspective on Digital Preservation*. **University of Cape Town**, 2008. http://www.husseinsspace.com/research/publications/iwdph_2007_african.pdf
- Webb, C. *UNESCO Guidelines for the Preservation of Digital Heritage*. 2003. <http://unesdoc.unesco.org/images/0013/001300/130071e.pdf>

Annex C. Milestones of Data Preservation

These milestones in the development of standards and practice provide context and background for the guidance, and complete the comprehensive community landscape for preserving data.

- 1881:** J S Billings, then Director of what was to become the US National Library of Medicine, suggests to Herman Hollerith that a mechanical system based on cards be used to tabulate the Census. Hollerith develops a punch-card system used with the 1890 Census, resulting in the first data in an automatically-stored format
- 1928:** IBM introduces a rectangular hole punch-card that becomes the industry standard
- 1939:** A committee at the US National Archives determines that federal agencies (rather than archivists) can determine whether records stored in punch-cards have historical value and should be preserved. Following this decision, few agencies retain any punch- card records for historical purposes
- 1950:** The US Census Bureau begins using the first non-military computer
- 1960s:** The establishment of data archives and national data programs for preserving data
- 1961:** National Archives of Australia
- 1962:** Inter-University Consortium for Political and Social Research (ICPSR)
- 1968:** National Archives and Records Administration (US), custodial program for electronic records
- 1968:** Statistical Package for the Social Sciences (SPSS) released
- 1970:** General Household Survey (UK) started
- 1971:** Statistical Analysis System (SAS) released
- 1975:** UK Data Archive begins affiliation with ICPSR
- 1976:** Council of European Social Science Data Archives (CESSDA) established
- 1980:** US Census Bureau establishes the State Data Center Program, a system to facilitate public access to data stored on computer tapes. All US states joined the program by mid-1980s
- 1985:** Ronald Reagan's White House issues National Security Decision Directive 189, stating: "[i]t is the policy of this Administration that, to the maximum extent possible, the products of fundamental research remain unrestricted."
 - The US Census Bureau becomes the first government agency to make information available on CD-ROM
- 1985-1986:** UK Data Archive and the BBC are partners in producing a new Domesday Survey, released on interactive video disc
- 1991:** Australian Centre for Remote Sensing (ACRES) rescues aging space data from disintegration by migrating from high-density magnetic tapes to optical tape
 - The "Bromley Principles" on full and open access to global change data are published <http://www.gcric.org/USGCRP/DataPolicy.html>
 - The UK Data Archive releases its first compact disk (CD)
- 1992:** GenBank launched by the US National Centre for Biotechnology Information
- 1993:** The Aboriginal Studies Electronic Data Archive (ASEDA) is launched by the Australian Institute of Aboriginal and Torres Strait Islander Studies
 - The South African Data Archive (SADA) is established by the Centre for Science Development (CSD) of the Human Sciences Research Council (HSRC)
- 1994:** UK Data Archive becomes web-based resource
- 1996:** The US Commission on Preservation and Access (CPA)/Research Library Group (RLG) publishes a seminal report on preserving digital information <http://www.oclc.org/programs/ourwork/past/digpresstudy/final-report.pdf>
- 1997:** The Social Science Data Archive (ADP) is established in Slovenia
- 1998:** The Sociological Data Archive (SDA) of the Institute of Sociology in Prague opens to the public
- 1999:** The UK's Arts and Humanities Data Service (AHDS) begins "Preservation Management of Digital Materials," a project to develop a handbook giving guidance on digital preservation
- 2000:** The US Library of Congress receives funding for the National Digital Information Infrastructure and Preservation Program (NDIIPP) to "provide a national focus on important policy, standards and technical components necessary to preserve digital content"

- The Dutch Digital Preservation Testbed is established as a part of the *Digitale Duurzaamheid* programme with the goal of achieving lasting accessibility of digital government information
 - The Internet becomes the principal mode of dissemination for the 2000 US census data
- 2001:** The UK Data Archive begins providing downloadable data from the Internet
- 2002:** Trusted Digital Repositories: Attributes and Responsibilities and Preservation Metadata and the OAIS Information Model are both published by RLG/OCLC
- Initial Open Archival Information System (OAIS) standards are released, providing a framework for long-term digital information preservation and access, including terminology and concepts for describing and comparing archival architectures [where, by whom?]
 - The Globus Toolkit is released by the Globus Alliance, establishing grid computing as a major component of the scientific computing infrastructure <http://www.globus.org/>
- 2003:** National Academy of Science releases an assessment of the US National Archives and Records Administration's proposed digital-archiving plan
- The US National Institutes of Health (NIH) adopts a data-sharing policy http://grants.nih.gov/grants/policy/data_sharing/
- 2004:** Digital Curation Centre launched in the UK, introducing the term *digital curation* as the combination of data curation and digital preservation
- 2005:** The US National Science Board (NSB) issues the report *Long-Lived Data Collections: Enabling Research and Education in the 21st Century* <http://www.nsf.gov/pubs/2005/nsb0540/>
- 2006:** The Accelerated Data Programme (ADP) is launched. The project "provides training and assistance in data curation to National Statistics Offices (NSOs) in developing countries"
- 2007:** The US National Science and Technology Council establishes the Inter-agency Working Group on Digital Data
- An interest group for people interested in data management in Africa is formed at IASSIST's conference
- 2008:** Data Seal of Approval launches <http://www.datasealofapproval.org/>
- First meeting of AASDA, the African Association of Statistical Data Archivists, at the DataFirst Survey Data Archive at the University of Cape Town, South Africa
- 2009:** CCSDS publishes a revision of the Reference Model for an Open Archival Information System (OAIS) for public examination and comment <http://cwe.ccsds.org/moims/docs/MOIMS-DAI/Draft%20Documents/OAIS-candidate-V2-markup.pdf>
- CCSDS publishes Metrics for Digital Repository Audit and Certification: <http://wiki.digitalrepositoryauditandcertification.org/pub/Main/WebHome/MetricsForDigitalRepositoryAuditAndCertificationWBv03a.doc>

Sources for Milestones

- Gold, Anna. "Cyberinfrastructure, Data, and Libraries: Part 1 - A Cyberinfrastructure Primer for Librarians." *D-Lib Magazine*. September/October 2007. <http://www.dlib.org/dlib/september07/gold/09gold-pt1.html>.
- , "Cyberinfrastructure, Data, and Libraries: Part 2 - Libraries and the Data Challenge: Roles and Actions for Libraries." *D-Lib Magazine*. September/October 2007. <http://www.dlib.org/dlib/september07/gold/09gold-pt2.html>.
- The Inter-university Consortium for Political and Social Research and Cornell University.
- "Digital Preservation Management Workshops and Tutorial." <http://www.icpsr.umich.edu/dpm/index.html>.
- Krejci, Jindrich. "The Czech Sociological Data Archive." *IASSIST Quarterly*. Summer 2001. <http://iassistdata.org/publications/iq/iq25/iqvol252krejci.pdf>.

- Lesaoana, Maseka A. "Data Archiving in Africa: The South African Experience." *IASSIST Quarterly*, Winter 1998. <http://iassistdata.org/publications/iq/iq21/iqvol211lesaoana.pdf>.
- McDonald, John. "Towards a National Digital Information Strategy: A Review of Relevant International Initiatives." Libraries and Archives of Canada. 2005. <http://www.collectionscanada.gc.ca/obj/012033/f2/012033-400-e.pdf>
- Stebe, Janez and Irena Vipavc. "The Social Science Data Archive in Slovenia." *IASSIST Quarterly*. Summer 2001. <http://iassistdata.org/publications/iq/iq25/iqvol252stebe.pdf>
- Suber, Peter. "Timeline of the Open Access Movement." 2009. <http://www.earlham.edu/~peters/fos/timeline.htm>.
- The Task Force on Archiving of Digital Information. "Preserving Digital Information." The Commission on Preservation and Access and the Research Libraries Group. 1996. <http://www.clir.org/pubs/reports/pub63watersgarrett.pdf>
- Woolfrey, Lynn. "The Establishment of the African Association of Statistical Data Archivists (AASDA)." *IASSIST Quarterly*. Summer 2007. <http://www.iassistdata.org/publications/iq/iq31/iqvol312woolfrey.pdf>.
- UK Data Archive. "Across the Decades of the Archive." 2009. <http://www.data-archive.ac.uk/ukda40/about/timeline.asp>.
- United States Census Bureau. "History." 2009. http://www.census.gov/history/www/through_the_decades/overview/.

Annex D. Programme and Policy Examples

These examples of policy and practice from a range of programmes, including data archives and other institutions, identify appropriate size and scale for organisations that preserve data. The guidance itself addresses organisations of any size or with any complexity of structure (e.g. a programme that produces data, a unit that preserves data, a data archive).

Arts and Humanities Data Service (AHDS). *A Strategic Policy Framework for Creating and Preserving Digital Collection*. 2001. <http://ahds.ac.uk/strategic.pdf>

Note: This report provides an overview of a high-level approach, case studies and best practice for creating strong digital preservation policies.

Centre for International Earth Science Information Network (CIESIN). *CIESIN Policy for Preservation of Digital Resources*. 2002. <http://www.ciesin.columbia.edu/documents/CIESINpreservationpolicy.pdf>

Note: CIESIN is a well-known programme that has provided a number of policy documents for the community. This policy is brief but offers an example of concise language and the identification of roles and duties of the CIESIN staff.

Cornell University Geospatial Information Repository (CUGIR). *CUGIR Data Management and Distribution Policy*. 2006. <http://cugir.mannlib.cornell.edu/CugirDataMgmtPolicy.20060828.pdf>

Note: This document provides a good overview of policies and procedures with respect to the institution's management and access policies, including data and metadata management, data security, distribution, use and rights.

Data Archiving and Networked Services (DANS). *Data Seal of Approval*. Undated. http://www.datasealofapproval.org/sites/default/files/Data_Seal_of_Approval_1-4.pdf

Note: This document provides "guidelines for the application and verification of quality aspects with regard to creation, storage and (re)use of digital research data in the social sciences and humanities", and identifies what each stakeholder must address to ensure data are safeguarded.

DataFirst

<http://www.datafirst.uct.ac.za/home/>

Note: DataFirst is a survey data archive and training facility in South Africa.

Earth Observing Laboratory (EOL). *Data Policy*. Undated. <http://www.eol.ucar.edu/data/data-policy>

Note: This policy outlines how EOL will deal with data from differing sources, access and levels of service provided.

Economic and Social Research Council (ESRC). *Data Policy*. 2001. http://www.esrcsocietytoday.ac.uk/ESRCInfoCentre/Images/DataPolicy2000_tcm6-12051.pdf

Note: This policy integrates subjects such as Principles, Collaboration with other Agencies, Charging Policy, Ethics and Intellectual Property. Although this might be too broad for an overall data-archive policy, it could be useful in constructing subordinate policies.

Greek Social Data Bank (GSDB). *Data Bank Role*. Undated. http://www.gsdb.gr/databank_role_en.html

Note: Although this is not a policy, it could be helpful in drafting an organisation's role, goals and founding principles.

Laaksonen, Helena, Sami Borg & Janez Stebe. "Setting up Acquisition Policies for a New Data Archive." *IASSIST Quarterly*. Spring 2006. <http://www.iassistdata.org/publications/iq/iq30/iqvol301laaksonen.pdf>

Note: Provides a detailed framework for establishing acquisition policies based on experience from the Finnish Social Science Data Archive (FSD) and the Slovene Social Science Data Archives (ADP).

National Centre for Atmospheric Research (NCAR). *The Computational and Information Systems Laboratory Strategic Plan*. 2009. <http://www.cisl.ucar.edu/dir/StrategicPlan/CISLSP2009-2014.pdf>

Note: This document provides an example of an organisation enumerating its strategic goals and identifying areas of growth and development.

Rural Economy and Land Use (RELU) *Data Management Policy*. 2006. <http://www.relu.ac.uk/about/data.htm> and <http://relu.esds.ac.uk/reluaug2006.pdf>

Note: This high-level overview provides a gateway to a more detailed data-management plan with an extremely comprehensive identification of roles and responsibilities.

Statistics New Zealand. *Data Archiving Policy*, undated. http://www.stats.govt.nz/about_us/policies-and-guidelines/data-archiving-policy.aspx

UK Data Archive. *UK Data Archive Preservation Policy*. 2009. <http://www.data-archive.ac.uk/news/publications/preservationpolicy.pdf>

Note: This comprehensive policy also serves as a guide to some procedural aspects of the archive such as IT architecture and security.

University of Zambia. Christine W Kanyengo. *Managing Digital Information Resources in Africa: preserving the integrity of scholarship*. 2006. <http://www.ascleiden.nl/pdf/elecpublconfkanyengo.pdf>

Note: Identifies issues facing digital preservation in Africa, specifically a lack of policies and infrastructure.

Annex E. Survey of Institutional Readiness

The Digital Preservation Management: Implementing Short-term Strategies for Long-term Problems workshop series developed this checklist to help organisations consider their digital assets in terms of scope, priorities, resources and overall readiness to address digital preservation concerns. The survey is available in the Digital Preservation Management tutorial that supplements the workshop curriculum <http://www.icpsr.umich.edu/dpm/dpm-eng/foundation/tdr/readiness.pdf> The checklist incorporates the themes of organisational, technological and resource concerns, which are introduced in the guidance.

At the start of planning, it is very common for many organisations to answer “No” or “Don’t know” to many of the questions. This survey is intended to identify requisite components of a digital preservation programme to enable effective planning.

A. Organisational Infrastructure

Organisational readiness is best reflected in the development and adoption of explicit policies that address digital preservation commitments and decisions. Often an organisation undertakes a programme without first ensuring necessary policies and controls are in place.

Mission

1. Can your institution’s mission statement be interpreted as supporting a long-term commitment to the preservation of valuable digital materials that your agency has acquired or created?

- Yes
- No
- Don’t know

Policies and Procedures

2. Do you have written policies and procedures that address long-term access (as opposed to those covering digitization)? If “No” or “Don’t know”, skip to section B. If “Yes”, continue to 2a.

- Yes — go to question 2a
- No — skip to section B
- Don’t know — skip to section B

- 2a. Do you have a written agreement with principal stakeholders on defined roles and responsibilities?

- Yes
- No
- Don’t know

- 2b. Do you have policies and guidelines covering selection, de-selection and acquisition?

- Yes
- No
- Don’t know

- 2c. Have you defined and promulgated quality-creation requirements and procedures?

- Yes
- No
- Don’t know

- 2d. Are comprehensive deposit guidelines in place?

- Yes
- No
- Don’t know

- 2e. Do you have written transfer requirements?

- Yes
- No
- Don’t know

- 2f. Have you explicitly defined preservation strategies appropriate to digital collections and objects that you have committed to preserving?

- Yes
- No
- Don’t know

If yes, please describe:

Authority

3. If you answered “Yes” to any part of question 2, were these digital preservation documents vetted by senior management?
- Fully
 - Partially
 - No but in progress
 - No
 - Don’t know

Implementation

4. Has your organisation implemented the policies and practices contained in these documents?
- Fully
 - Partially
 - No but in progress
 - No
 - Don’t know

B. Technological Infrastructure

Organisations tend to rely on or create digital content first and address long-term access later. This section addresses current and planned digital objects and collections, storage management and depositories.

Digital Collections

5. Do you currently have the following types of digital objects that your institution has committed to maintaining over time? (Some objects might fall into multiple categories. Please check all that apply.)

Licensed e-journal files (articles, issues, journals)

- have now will have no

Institutional records (in any format)

- have now will have no

Web-sites

- have now will have no

E-mail

- have now will have no

Word-processing files

- have now will have no

Digital-image files

- have now will have no

PDF files

- have now will have no

Numeric data files

- have now will have no

Databases and spreadsheets

- have now will have no

Geographic information systems (GIS)

- have now will have no

Audio-visual files

- have now will have no

Other (please list)

6. Is there any digital material in your holdings for which you lack the operational and/or technical capacity to mount, read and access?
- Yes
 - No
 - Don’t know

If “Yes”, please describe:

Archival Storage

7. Are you using any of these kinds of file storage?
(Please check all that apply.)

On-line (i.e. spinning magnetic disk)

- access copies master files backup

Magnetic tape

- access copies master files backup

CD, DVD or other optical or magneto-optical disk

- access copies master files backup

Solid state

- access copies master files backup

Other (please list)

Storage Practice

8. Does your storage programme include (please check all that apply):

- multiple copies of digital content managed online?
- the use of high-quality storage media?
- on-line or off-line copies stored in geographically-distributed locations?
- an access-controlled area for the machines and media on which files are stored?
- an environmentally-controlled area for storage media?
- a disaster-recovery plan?
- a media-testing programme?
- a media-refreshing/migration plan?
- back-up?
- off-site storage for back-ups?

Other (Please describe.)

Obsolescence

9. Have you undertaken any action to extend the life of digital content threatened by obsolescence of file formats, storage media and the supporting hardware to access it or other associated hardware and software?

File Formats Yes No

Storage Media Yes No

Storage Drive Yes No

10. What action have you taken?

Depository

11. Have you established any kind of digital depository arrangements for managing your digital collections over time?

- Yes
- No
- Don't know

Depository Development

12. If yes, have you:

- developed a depository in-house?
- acquired proprietary software to implement the depository?
- acquired open-source software to implement the depository?
- outsourced the development/maintenance of the depository?
- contracted with a third party organisation for depository services?
- joined a consortium for developing/delivering depository services?

- made other arrangements? (describe)

If you have ticked any of these, please bring whatever documentation you have on the arrangements.

Security

13. Does your depository have security and other mechanisms in place to ensure the integrity of objects against intentional or accidental security threats?

- Yes
 No
 Don't know

C. Resources

Once the need to establish a digital preservation programme is recognised and there is the will to do so, the organisation must be ready to build and sustain the programme. This section covers resources: financial, human and technical.

Sustainable Funding

15. Does your institution currently have funding dedicated to the long-term maintenance of your digital collections?

- Yes
 No
 Don't know

If "Yes", please describe the extent and nature of funding.

If "No", please outline potential sources of and plans for acquiring funding.

Staffing

16. Are there staff at your institution specifically charged with digital preservation responsibility?

- Yes
 No
 Don't know

16a. If "Yes", how many?

16b. If "Yes", please list their titles.

OAIS Compliance

14. Is your organisation committed to the development or use of an OAIS-compliant depository for the long-term preservation of digital objects that are or will be within the scope of your preservation programme?

- Yes
 No
 Don't know

17. Is there adequate organisational expertise to develop a digital preservation programme?

- Yes
 No
 Don't know

18. Is there adequate technical expertise to develop a digital preservation programme?

- Yes
 No
 Don't know

19. Does senior management view digital preservation as a key priority?

- Yes
 No
 Don't know

20. Is there adequate support for staff training in digital preservation?

- Yes
 No
 Don't know

21. Does your institution currently use outside sources of expertise for digital preservation (e.g. consultants, contracts)?

- Yes
 No
 Don't know

Technological Infrastructure

22. Is the current technological infrastructure of your organisation adequate to build and/or sustain a digital preservation programme, with requisite upgrading and enhancement over time?
- Yes
 - No
 - Don't know

Please describe the status of the infrastructure and potential concerns.

23. Has your institution dedicated funds annually for technology development, replacement and upgrading?
- Yes
 - No
 - Don't know

Administrative Structure

24. Would the digital preservation programme be:
- incorporated into existing units?
 - a separate unit?
25. How would you rank the following factors as threats to the loss of digital material at your institution within the next three years? 1=greatest threat, 5=smallest threat:
- Technological obsolescence
 - Insufficient policies or plans for preservation
 - Insufficient resources for preservation
 - Inadequate support from senior staff
 - Lack of expertise
 - Other; please describe:

Annex F. Technological Obsolescence and Other Threats

This annex includes a discussion of file format obsolescence, hardware obsolescence and physical threats to digital content. The content was extracted from section 3 of the Digital Preservation Management: Implementing Short-term Strategies for Long-term Problems on-line tutorial. The full content of the tutorial is available at <http://www.icpsr.umich.edu/dpm/> The primary author of this text was Richard Entlich at Cornell University Library, with periodic additions and updating since the tutorial's launch in 2003.

Obsolescence: File Formats and Software

Computer files, the objects normally thought of as the main target of digital preservation, are presented according to pre-defined structural and organisational principles. These principles, usually referred to as a file format, are typically laid out in a document called a format specification. A format specification provides the details necessary to construct a valid file of a particular type and to develop software applications that can decode and render such files.

Although some file format specifications are largely independent of specific software (for example, encoding schemes such as ASCII and Unicode), most are tied to individual or related groups of software. The software and its related file format specification usually evolve together and their fates are closely linked. Therefore, it makes sense to discuss software obsolescence and file format obsolescence together.

What Factors Contribute to File-Format Obsolescence?

File formats can become obsolete for a number of reasons:

- Software upgrades fail to support legacy files.
- The format itself is superseded by another or evolves in complexity.
- The format "take-up" is low or industry fails to create compatible software.

- The format fails, stagnates or is no longer compatible with the current environment.
- Software supporting the format fails in the marketplace or is bought by a competitor and withdrawn.

Why are File Formats a Challenge to Digital Preservation?

A number of factors have contributed to the challenge presented by digital file formats. During the early decades of computing, the threat of file format obsolescence to the long-term maintenance of digital objects was not widely recognised. No systematic efforts were made to collect software documentation or file format specifications. Without proper documentation, the task of trying to interpret an old file, or even determine what format it was written in, is daunting. Thousands of file formats and their variants have been created. Only recently has an effort been made to catalogue and document them and understand their relationships and variations. Tools are beginning to emerge to automate the process of identifying and characterising files by their formats.

Most software is upgraded on a regular basis. Although most applications can read files created in the previous version and perhaps the one before that, the ability to read older versions is often lost. Files that have not been migrated might not be readable by the latest software version, and the older-version software might no longer be available or run on a current computer, or under a current version of the operating system.

Also, due to the complexity and dynamic nature of many file formats, it can be extremely difficult to determine whether a file moved from one format to another (or to a newer version of the same format) has retained all its characteristics and functions.

Are Some File Formats Less Vulnerable to Obsolescence than Others?

Since all software is subject to obsolescence, all file formats used by that software are also vulnerable. On the surface, it might seem that the files used by software that is more stable (i.e. not undergoing much change) would be less subject to obsolescence, and that is true in the short term. But software that stands still inevitably becomes obsolete also, because it fails to adapt to the changing computing environment that it must operate in (e.g. CPU architecture, operating systems, encoding

schemes, and data-transfer protocols). So users must be watchful of files that either rapidly evolve or stagnate, since both are prone to obsolescence.

To decode an old file format, the format specification must be available. Therefore, the degree of control the creator of a format specification exerts over its publication has a significant impact on the format's vulnerability to obsolescence. Specifications tend to fall into one of three categories:

Proprietary, closed specifications. Proprietary and closed specifications represent some of the most enduring and successful software in use. However, these also tend to evolve quickly and exist in many different versions for different platforms, with only limited backward compatibility provided. In fact, there is substantial commercial incentive to avoid good backward compatibility, since the need to share files ultimately forces all users, including those who would prefer to keep using older versions, to upgrade to newer versions. Commercial vendors must regularly release new versions of their software with added features and functions in order to entice users to upgrade and provide a revenue stream. Unfortunately, experience has shown that even very old specifications for versions of commercial file formats withdrawn from the market might never be released. Also, as one might expect, proprietary and closed file formats are interpreted with the highest accuracy by the manufacturer's own software. Therefore, such formats are the most vulnerable to obsolescence since they face the dual risk of rapid specification change and being tied to a single product or company.

Furthermore, today's wildly successful software can be tomorrow's also-ran or distant memory. There has been tremendous consolidation in the commercial software industry and many products have disappeared following mergers and acquisitions. Others have succumbed to competition from superior or more cleverly-marketed products.

Proprietary, open specifications. Some proprietary formats have a lower risk because the specification has been publicly released, allowing other companies (and non-commercial entities) to produce software that can read them. However, commercial entities can and sometimes do change their minds about leaving specifications open. For example, the DjVu image format was an open specification for a while before its owner decided to make changes and not release them to the public.

Such formats can provide a compromise between closed specifications and international standards by combining commercial clout with some degree of openness. There is even the potential for middle ground, as when a subset of a proprietary format is adopted as a standard, such as the case of PDF/A, an archival version of PDF grounded in Adobe's proprietary but open specification. PDF/A differs from PDF in the requirement of XML-based metadata and the elimination of elements likely to complicate decoding and accelerate obsolescence, such as audio and video, JavaScript, unembedded fonts and device-dependent colour spaces.

Most proprietary but open specifications are still vulnerable to the whims of market forces. In addition to being subject to arbitrary withdrawal, they can be abandoned for commercial reasons.

Non-proprietary, open specifications. In terms of guaranteed long-term availability, published specifications produced by international standards bodies are the safest. Generally, representatives from many different constituencies are involved in creating the standard, helping to ensure that it balances the needs of a wide variety of users and is not beholden to any particular commercial interest. Broad participation also helps provide an incentive for wide support once the standard is completed. Backward compatibility with older, related standards is usually a priority and there are no commercial pressures for rapid obsolescence. One of the most recent examples of this is the standardisation of the OpenDocument Format (ODF) as an open-source format. ODF stemmed from the open XML-based OpenOffice.org specification and was approved by ISO as a standard in 2006.

On the other hand, not all standard formats should be assumed to be the best choices. Standards must become widely adopted by both user and developer communities to acquire reduced vulnerability to obsolescence. That does not always happen.

Choosing File Formats for Reduced Vulnerability to Obsolescence

The following factors should be considered in assessing a file format's long-term stability:

- wide adoption
- history of backward compatibility

- good metadata support (in open format such as XML)
- good range of functionality but not overly complex
- available interchange format with usable target
- built-in error checking
- reasonable upgrade cycle

Obsolescence: Hardware and Media

Rapid obsolescence of computer hardware has been a particular characteristic of the industry since its inception more than 50 years ago. A one-or-two order of magnitude improvement in power, speed, efficiency or cost per value has occurred every several years in areas such as CPU speed, memory chip density, storage device capacity, video-processing rate and data-transmission rate.




Such monumental changes have a powerful obsolescent effect. New computers replace older ones not just because they are quantitatively faster, more productive or higher in capacity (though this impact alone is a considerable incentive to upgrade), but because they enable qualitative changes in the function of the device. Entire classes of computer uses, and the software and file formats with which they were implemented, would not exist today had computing hardware not advanced to such a degree. These include uses such as CAD, digital-imaging, audio and video production, simulation, and graphic web surfing.

Thus, new computing hardware opens the door to new and improved software, leading to software and file-format obsolescence. The new software will not run on old hardware, further exacerbating hardware obsolescence. At the same time, the new hardware introduces other new technology, such as peripheral connections (e.g. Firewire and USB have replaced RS-232 serial and Centronics parallel ports) and storage devices (e.g. USB keys and CD drives have replaced floppy disks). These changes force older peripherals into retirement along with their compatible computers.

Types of Digital Storage Media

The nature of the physical media on which digital data are stored presents a major challenge to the preservation of digital content. The great variety of media types, their often rapid obsolescence due to technological changes and their vulnerability to physical degradation all contribute to the problem.

There are three commonly-used categories of digital-storage media: disk, tape and solid state. Within each category are many levels of sub-categories, representing both integrated storage (drive and media as a single unit) as well as removable media.

 Disk
Magnetic (fixed hard drive)
Magnetic (removable)
Hard disk packs
Floppy
Zip, Jaz, etc
Magneto-optical (write-once, read/write)
Optical (read-only, write-once, recordable, read/write)
 Tape
Open reel
Cassette
Cartridge
 Solid State
CompactFlash, Memory Stick, Smart Media (digital camera memory)
USB memory key or stick, pen drives, keychain drives (portable storage up to 2 GB)
Flash drives (IDE and SCSI) using standard hard disk form factors; often for industrial or military use in adverse temperatures or extreme or dust conditions (capacity up to 61 GB)

Trends Contributing to Hardware Obsolescence

Several technological trends encourage storage media obsolescence. These trends include:

Decrease in physical size

- hard drives (24" → 1" over a 40 year period)
- floppy disks (8" → 5.25" → 3.5" over a 10 year period)
- optical media (14" → 2" over a 20 year period)

Increase in storage capacity

- hard drives (5 MB → 400 GB, 2 TB in 2009)
- tape cartridges (1 TB cartridges coming)
- 12 cm optical media (650 MB → 50 GB in 2009)

Declining cost per unit of storage

- hard drives (most rapid)
- tape cartridges
- optical media (least rapid)

Other trends are less uniform for all media

- reliability (generally improving)
- fragility (variable)
- stability (generally improving)
- time to obsolescence (variable)

Physical Threats to Storage Media and Hardware

Digital storage media and hardware are subject to numerous internal and external forces that can damage or destroy their readability:

- material instability
- unsuitable storage conditions (temperature, humidity, light, dust)
- over-use (mainly for physical contact media)
- natural disaster (fire, flood, earthquake)
- infrastructure failure (plumbing, electrical, climate control)
- inadequate hardware maintenance
- hardware malfunction
- human error (including improper handling)
- sabotage (theft, vandalism)

Storage Issues

Improper storage might be the most common reason for premature media failure. Moderation of temperature and humidity are well known to extend the usable life of most storage media, but many other factors can help, too.

Suggested Preservation Action for Managing Storage Media

- Maintain consistent temperature ~ 20 deg C (68 deg F) (see the IPI Media Storage Quick Reference for specific guidelines).
- Maintain relative humidity around 40%.
- Avoid large and rapid fluctuations in temperature/humidity.
- Control dust (maintain a slight positive pressure environment).
- Avoid exposure to magnetic fields (for magnetic media).
- Avoid exposure to fumes.

- Establish a “No food, drink or smoking” policy in media-storage areas.
 - Store media in closed metal cabinets, electrically grounded.
 - Shelf media vertically (not stacked).
 - Store media in their original cases.
 - Minimise exposure to sunlight and UV from light fittings.
 - Allow media to acclimatise to new temperatures and humidity before using.
 - Return to controlled storage immediately after use.
- CDs should be labelled only on the top surface and with approved markers.
 - Avoid flexing CDs and DVDs.
 - Do not leave media in drive after use.
 - Re-tension magnetic tapes after use and every 1-3 years even if not used.
 - Limit media access to properly-trained staff.
 - Use write-once or recordable (rather than rewritable) media.
 - Set write-protect tabs, if available.

Preparation Matters

Physical threats from natural disasters, infrastructure failure and malicious destruction cannot usually be predicted, but it is possible to lessen their occurrence and minimise the damage they cause by proper preparation. Awareness the moment a hazardous condition arises allows the fastest possible response. Sensors and alarm systems to detect and report the presence of fire, heat, smoke, water leaks and unauthorised entry are available. Fire-suppression systems, floor drains and use of heat-resistant and/or waterproof storage can all help minimise damage to sensitive media and equipment. Media storage areas should be locked and accessible only to properly-trained personnel. All media, no matter how reliable, need to be backed up. Creation of multiple back-ups and the use of off-site storage for one set of copies provides the best protection against catastrophic loss.

Handling Issues

Another major threat to storage media comes from improper handling. Though many digital media give the impression of sturdiness and durability, they can be damaged by too casual an approach to use. Observe these recommendations:

Suggested Preservation Action

- Don't open shutters designed to protect media in cartridges.
- Handle media with lint-free gloves to minimise dust.
- Clean and dry hands before handling media.
- Do not touch exposed media surfaces (e.g. handle CDs at edges).
- Keep media in their cases except when in use.
- Place labels only in approved areas; write on label *before* applying.

Life Expectancy

A lot of ink has been spilt over the issue of media longevity. Media consumers and producers have placed a great deal of emphasis on seeking and promoting long-lasting media. Ultimately, however, for a great many reasons, longevity is overrated as a desirable media characteristic.

Media life expectancy claims are statistical averages based on accelerated aging tests and can only provide a rough estimate of how long any particular piece of media will last under certain storage and handling conditions..

Longevity provides no protection against many media threats, including theft, natural disasters, infrastructure failures and accidental handling damage.

Media technology changes so rapidly that long-lasting media are likely to be threatened by obsolescence before their useful life is over.

Suggested Preservation Action

- A better strategy is to take steps to maximise the intrinsic longevity of standard-longevity media:
- Adhere to good storage and handling practices.
- Buy quality media.
- Take note of media manufacturer and batch numbers so performance and quality trends can be tracked.
- Do not over-buy.
- Remember that some unrecorded media have a shorter shelf-life than recorded media (optical and magneto-optical in particular).
- Buy media designed for the speed and capacity of the drives in which it will be used.

All media need periodic testing to confirm data integrity. At a minimum this should include procedures to:

- Confirm fidelity of all media immediately after recording.
- Once recorded, read samples (by batch code and manufacturer and/or storage location) of the entire media and samples of files from several media periodically..
- Detect problems with specific batches, manufacturers or storage conditions and test them more extensively.
- Test blank media (can be expensive and time-consuming).
- Monitor use of error correction and replace media before errors become impossible to correct..

Specific Media Issues

Hard disk drives

- Highly commoditised. Expect to pay more for quality drives with above-average reliability.
- Do not buy excess capacity. Prices and technology change extremely rapidly.
- Do not expect more than five years' use from any hard-disk drive.
- High temperature can dramatically reduce life expectancy. Control the environment and make sure fans are working and not clogged with dust.

Magnetic tape

- Still the least expensive back-up medium per unit.
- New technology offers extremely high-density storage.
- High-density cartridge technology (SDLT, LTO, AIT) considered the most reliable.
- New generations of tape formats appear regularly. Backward compatibility is usually only offered for 1-2 generations.

Optical (CD/DVD variants)

- These media have several possible failure modes
 - o Dye layer (for recordable media)
 - o Reflective layer
 - o Substrate separation
- Unrecorded media have a 5-10 year shelf-life

- The top surface (label area) of CDs requires extra care, as it is more vulnerable than the bottom layer
- CD/DVD cleaning should be done axially (i.e. outer edge to inner edge), not radially or along the tracks
- DVDs are more vulnerable to flexing damage due to closer track spacing. Use special DVD carriers to minimise flexing upon removal

Disaster Recovery

The methods and procedures already mentioned are designed to minimise casual loss of data and maximise media longevity. However, even if you had perfect storage conditions and impeccable handling protocols, some media still fail. Therefore, valuable data must be stored redundantly, that is, backed up, on more than one piece of media. In addition, back-ups and disaster-recovery plans are needed to avoid catastrophic media loss from causes such as:

- sabotage (theft, vandalism, malicious modification/erasure, viruses, terrorist attack, etc)
- natural disaster (fire, flood, earthquake, hurricane, tornado, infrastructure failure)

A disaster-recovery plan that deals specifically with information technology infrastructure is needed. Developing such a plan is not a one-off process; it has to be tested and modified as changing circumstances dictate. Revisit the disaster recovery plan for occurrences such as new staff, new or reorganised physical plant or new equipment. Once a disaster-recovery plan is in place, take steps to prevent catastrophes and minimise damage from them.

NIST (National Institute of Standards and Technology) publishes an excellent guide to developing and implementing plans to cope with disaster entitled "Contingency Planning Guide for Information Technology Systems".

Back-ups

The maintenance of redundant copies of valued digital content is an essential component of any digital-preservation programme, and a key element in the prevention of catastrophic loss. A great variety of back-up solutions is available. Which to use depends on:

- quantity of data
- rate of change

- degree of automation desired
- available budget

In addition to backing up data files, application software and operating systems might also need back-up. In some cases, it might be necessary to purchase additional licences or obtain special permission from the software vendor in order to back up applications.

In addition to testing back-up media periodically to ensure the data are still readable and have not been altered, restore procedures should also be tested to ensure the hardware, software and any outside vendors involved in maintaining back-up are all functioning as expected.

A prudent back-up strategy places at least one copy of all critical data at a sufficient distance from the main data store so both are not likely to succumb to the same disaster. This is called off-site storage. Institutions should check with regulatory agencies for their records-retention requirements. Medical and financial records might have more rigorous requirements for the distance of the off-site storage facility from the main one. Cooperating on a reciprocal storage arrangement with a another institution could be an inexpensive way of managing off-site storage. If outsourcing, make sure you are getting true data management, not just

warehousing. Generic storage facilities are unlikely to know how to store and handle digital media properly. Environmental controls and handling protocols should be at least as rigorous as those at the main facility.

An important consideration in selecting a back-up strategy is the possibility of disastrous loss of the entire primary facility -- all equipment and data. In such a situation, it will be necessary to replace the entire IT infrastructure and to restore the secondary data store to new equipment. If new equipment that can handle the back-up media, the restoration software and the applications software and operating environment needed to access the data cannot be purchased, it might be time to think again about the back-up plan.

Emergency Rescue

Poor planning or circumstances beyond one's control can lead to emergency rescue of critical data. Anything from budget woes to bad luck can provide the trigger. Companies exist that specialise in salvaging data from badly-damaged media, when no backup exists, and reading data from obsolete storage technology. These services can often be quite expensive but also be a lifesaver. A web search for "data recovery" should produce a plethora of links to such specialised companies.

About the IHSN

In February 2004, representatives from developing countries and development agencies participated in the Second Roundtable on Development Results held in Marrakech, Morocco. They reflected on how donors can better coordinate support to strengthen the statistical systems and monitoring and evaluation capacity that countries need to manage their development process. One of the outcomes of the Roundtable was the adoption of a global plan for statistics, the Marrakech Action Plan for Statistics (MAPS).

Among the MAPS key recommendations was the creation of an International Household Survey Network. In doing so, the international community acknowledged the critical role played by sample surveys in supporting the planning, implementation and monitoring of development policies and programs. Furthermore, it provided national and international agencies with a platform to better coordinate and manage socioeconomic data collection and analysis, and to mobilize support for more efficient and effective approaches to conducting surveys in developing countries.

The IHSN Working Paper series is intended to encourage the exchange of ideas and discussion on topics related to the design and implementation of household surveys, and to the analysis, dissemination and use of survey data. People who wish to submit material for publication in the IHSN Working Paper series are encouraged to contact the IHSN secretariat via info@ihnsn.org.

www.ihnsn.org
E-mail: info@ihnsn.org