CENDI

icsti
international council for scientific and technical information
conseil international pour l'information scientifique et technique

CENDI – 2004-3: Rev. 05/04

DIGITAL PRESERVATION AND PERMANENT ACCESS TO SCIENTIFIC INFORMATION: THE STATE OF THE PRACTICE

Gail Hodge
Information International Associates, Inc.

Evelyn Frangakis
CENDI Digital Preservation Task Group/
National Agricultural Library

A Report Sponsored by
The International Council for Scientific and Technical Information (ICSTI)
And
CENDI
US Federal Information Managers Group

February 2004

Revised April 2004

DISTRIBUTION STATEMENT A
Approved for Public Release
Distribution Unlimited

20040607 047

The authors wish to thank the following for their contributions to this report:

William Anderson (US CODATA)
Neal Beagrie (DPC/JISC)
Robert Chen (CIESEN)
Julie Esanu (US CODATA/NRC)
Linda Hill (UC Santa Barbara)
Eleanor Frierson (NAL)
Harold Frisch (NASA Goddard Space Flight Center)
Barry Mahon (ICSTI, Executive Director)
Kurt Molholm (DTIC/ICSTI President)
Marc Nikolaus (National Cancer Institute/NIH)
Kent Smith (NLM/CENDI Chair)
Deborah Woodyard (The British Library)

The CENDI Digital Preservation Task Group
and
The representatives of the organizations/projects who responded to the survey.

ICSTI *is a unique forum for interaction amongst organizations that create, disseminate and use scientific and technical information. ICSTI's mission cuts across scientific and technical disciplines as well as international borders to give Member organizations the benefit of a truly global community. ICSTI seeks to reduce or eliminate barriers to effective transfer of information by:*

- *Promoting the value of scientific and technical information to the world's economic, research, scholarly and social progress.*
- *Enhancing access to and delivery of information to academia, business, government and the public.*
- *Forging better relations among different communities involved in information transfer, from generator to disseminator to user.*
- *ICSTI is a non-profit association which derives its operational budget essentially from membership dues.*

Web: http://www.icsti.org
Email: icsti@icsti.org

CENDI *is an interagency cooperative organization composed of the scientific and technical information (STI) managers from the Departments of Agriculture, Commerce, Energy, Education, Defense, the Environmental Protection Agency, Health and Human Services, Interior, Government Printing Office, National Archives and Records Administration, and the National Aeronautics and Space Administration.*

CENDI's mission is to help improve the productivity of federal science-based programs through the development and management of effective scientific and technical information support systems. In fulfilling its mission, CENDI member agencies play an important role in helping to strengthen US competitiveness and address science- and technology-based national priorities.

Web: http://www.dtic.mil/cendi

TABLE OF CONTENTS

Executive Summary **1**

1.0 Introduction *3*

2.0 Scope and Methodology *3*

3.0 Highlighted Systems *4*

4.0 Setting the Stage *8*
- 4.1 Archiving Concepts and Definitions 8
- 4.2 The Scientific Environment 8
- 4.3 The Technological Environment 9
- 4.4 Scientific Publishing and Communications 9
  - 4.4.1 Open Access 9
  - 4.4.2 Institutional Repositories 11
- 4.5 Legal Deposit and Copyright 12

5.0 Stakeholder Roles *15*
- 5.1 Publishers 15
- 5.2 National Libraries 17
- 5.3 Institutions 18
- 5.4 Museums 19
- 5.5 National, State and Regional Archives 19
- 5.6 Trusted Third Parties 20
- 5.7 The Role of Government 21
- 5.8 Foundations and Other Private Funding Sources 22

6.0 Preservation by Document Type *23*
- 6.1 Electronic Journals 23
- 6.2 Theses and Dissertations 24
- 6.3 Scientific Data Sets 25
- 6.4 Technical Reports 26
- 6.5 Conferences, Meetings and Lectures 26
- 6.6 E-Records 27

7.0 Standards by Format Type *28*
- 7.1 Text 29
- 7.2 Images 30
- 7.3 Numeric Data 32
- 7.4 Video and Audio 32
- 7.5 Output from Design, Modeling and Visualization Tools 33

8.0 The Workflow *34*
- 8.1 Selection Criteria 34
- 8.2 Metadata Creation 36
- 8.3 Archiving and Transformation 38
  - 8.3.1 Transformation to a Preservation Format 38
  - 8.3.2 Migration 39
  - 8.3.3 Migration On-Request 39
- 8.4 Storage 40
- 8.5 Dissemination 40

9.0 The Introduction of "Off-The-Shelf" Systems *41*
- 9.1 DSpace Institutional Digital Repository System 41
- 9.2 Digital Information Archive System 42
- 9.3 OCLC Digital Archive 44
- 9.4 PANDORA Digital Archiving System (PANDAS) 45

9.5 Lots of Copies Keep Stuff Safe (LOCKSS) 46
9.6 Fedora™ (Flexible Extensible Digital Object Repository Architecture) 47

10.0 Standards Activities *48*
10.1 Metadata 48
10.1.1 Descriptive Metadata 48
10.1.2 Preservation Metadata 49
10.1.3 Technical Metadata 50
10.1.4 Structural Metadata 51
10.2 Permanence Ratings 51
10.3 Open Archival Information System Reference Model (OAIS RM) 52
10.4 Producer-Archive Interface Methodology 52
10.5 Persistent Identifiers 53

11.0 New Issues and the Research Agenda *56*
11.1 Authenticity 57
11.2 Rendering Objects for Permanent Access 57
11.3 Saving the Dynamic Web 58
11.4 Appraising and Retaining Scientific Data 58
11.5 Preserving Government Information 59
11.6 Archiving the Archive 61
11.7 Interoperable Archives 62
11.8 Partnerships 62
11.9 Costs and Sustainability 63

12.0 Findings and Trends *65*
12.1 Systems solutions are being developed by a variety of stakeholders and partnerships 65
12.2 The Open Archival Information System (OAIS) Reference Model has been widely adopted 66
12.3 Organizations are focused on capturing and acquiring digital information, rather than preservation or permanent access 66
12.4 Efforts for digital depository legislation are gaining momentum 66
12.5 Migration remains the preservation strategy of choice; it is still too soon for most archives to have undergone a significant technological change 66
12.6 There are increased standards-related activities 67
12.7 Open standards developed for interoperability hold promise as the basis for preservation formats 67
12.8 Key technical issues remain 67
12.9 Partnerships are increasingly important 67
12.10 Key social, political, and economic issues remain, including the need to develop a "will to preserve and provide permanent access" within the scientific and technical community and society in general 67

13.0 Recommended Next Steps *68*

14.0 References *70*

15.0 Appendix I: Follow Up Discussion Questions *82*

Executive Summary

In 1999, the International Council for Scientific and Technical Information (ICSTI) and CENDI jointly sponsored a report on **Digital Electronic Archiving: The State of the Art and Practice** (Carroll & Hodge 1999). ICSTI and CENDI remain interested in digital preservation as they represent large repositories, publishers, and libraries of scientific and technical information. This report is an update to that 1999 report.

This report focuses on operational digital preservation systems specifically in science and technology (S&T). It considers the wide range of digital objects of interest to S&T, including e-journals, technical reports, e-records, project documents, scientific data, etc. The report also discusses archiving based on format types – text, data, audio, video, etc. It is, of course, international in scope, and as much as possible crosses organizational sectors (academic, government, commercial, etc.).

However, this report does not attempt to provide a comprehensive survey of systems, but, rather, to highlight selected systems/projects that can help to identify trends, remaining issues and activities that ICSTI, CENDI, and other organizations interested in the preservation and permanent access to the record of science can consider when developing their own systems and policies. More than 50 projects and systems were identified from the surveys, from experts, or from the literature. From these, 21 were selected for highlighting in this report. However, references are made to other projects throughout the report as appropriate.

The major findings are as follows:

Systems solutions are being developed by a variety of stakeholders and partnerships.
The advent of off-the-shelf solutions shows advancing maturity in the area of digital preservation. The library model with shared cataloging tools and service providers is apparent. The six key systems, the OCLC Digital Archive, DSpace, LOCKSS (Lots of Copies Keep Stuff Safe), Fedora™, PANDAS, and the Digital Information Archive System (DIAS) from IBM, come from different types of organizations – a library service provider, a university repository, a large academic research library paired with a provider of publishing services, a university repository teamed with another university's digital library research group, a national library system, and a national library working with a commercial company, showing the need for partnerships and the interactions among a variety of stakeholders.

The Open Archival Information System (OAIS) Reference Model has been widely adopted.

The OAIS Reference Model, which became an International Standards Organization (ISO) standard in June 2003, has been adopted widely. All types of archives use the OAIS terminology and conceptual model. However, it is not as prevalent in the scientific data community for which it was initiated, partly because these organizations already had systems, customers, producers, and processes of a legacy nature. Efforts are underway among some data archives to minimally ingest Submission Information Packages (SIPs) and to produce Dissemination Information Packages (DIPs) in order to respond to the spirit of the standard.

As systems are redesigned and the need for interoperability increases, it is likely that the OAIS Model will become more prevalent as the conceptual basis for scientific archives.

Organizations are focused on capturing and acquiring digital information, rather than preservation or permanent access.

Even if they use the term archive or have preservation in their mission, the initial goal is to get a critical mass of material, to promote a culture of deposit/submission/harvesting and sharing, and to provide access to the currently collected materials. While many of the institutional repository activities are committed to long term preservation and access, the technical and metadata aspects required are not yet well incorporated into their systems.

Efforts for digital depository legislation are gaining momentum.

There are significant activities on the part of national libraries and other stakeholder groups with regard to changing existing laws or adding new laws that would require deposit of digital materials. This has gained significant momentum over the last several years, and most recently, the United Kingdom and New Zealand have passed such legislation. Digital deposit legislation may be more accepted, now that there have been major pilot projects involving national libraries and large commercial publishers. In addition, voluntary arrangements are already in place, so the legislation more closely reflects current practice rather than leading it.

Migration remains the preservation strategy of choice; it is still too soon for most archives to have undergone a significant technological change.

Other than the large data archives, which have existed for many years, archives have not yet faced large-scale technological changes. This means that migration remains the strategy for most of the materials of interest to libraries, archives, and publishers. The prevalence of migration, particularly from one version of software to another, also indicates the prevalence of commercially available products, such as Microsoft Office and Adobe products, in the scientific environment. While concerns were expressed about outdated software, hardware, and media, these issues are not the current focus as the institutions grapple with collecting and ingesting the flood of current archival content.

There are increased standards-related activities.

There are standards-related activities underway in the areas of producer-archive interaction, permanence ratings, persistent identifiers as critical components of digital preservation systems, preservation metadata, and preservation formats (e.g., PDF-A for text). These activities are likely to produce significant results, because they are codifying many of the best practices that have been identified over the last several years of pilot projects.

- ***Open standards developed for interoperability hold promise as the basis for preservation formats.***

While the main rationale for development of open standards is interoperability among software environments, these standards also may be applicable for long term archiving. Open formats such as those for geographic information systems (OpenGIS), product design and

manufacturing (STEP), open office documents (OpenOffice), and chemical structures (Molfiles and SMILES) are working toward hardware and software independence. The potential for using these formats for preservation should be investigated further.

Key technical issues remain.

There are several key technical areas requiring future research that have been identified in recent studies funded by the National Science Foundation. Additional research is needed into the automatic generation of metadata, through self-describing objects or the provision of archiving mechanisms in authoring tools. Registries, perhaps of a global nature, are needed to maintain authoritative, computer-actionable information about metadata tag sets, reference information for formats and hardware/software behaviors. Research into the archiving and preservation of dynamic, non-HTML and database-driven Web content is a major research activity for several groups. Other technical issues include creating interoperable archives and best practices for archiving and preserving the archive itself.

Partnerships are increasingly important.

Over the last several years, there has been an increasing realization that partnerships are the only way to ensure that digital information will be preserved. In addition to ensuring some measure of comprehensiveness over the wide spectrum of scientific information in digital form, partnerships have the benefit of providing some measure of redundancy, sustainability and sharing of the cost for preservation which is likely to exceed the revenues that can be made on the reuse of any particular object. A workable infrastructure will result from a multi-pronged approach involving publishers, libraries, archives, institutions, and trusted third parties, with appropriate support from governments, foundations and other funding sources, users and creators during the life cycle of the material to be preserved.

- ***Key social, political, and economic issues remain, including the need to develop a "will to preserve and provide permanent access" within the scientific and technical community and society in general.***

There are several outstanding social and political issues that require further discussions by the various stakeholder groups involved in preserving and providing permanent access to scientific and technical information. For example, the social, political and legal aspects of creating federated archives and working partnerships that cross stakeholder groups and object types (data, publications, multimedia, etc.) must be resolved. The archiving and preservation of and long term access to government information pose special challenges in this regard. Sustainable business models that will survive for the long term also remain elusive. Collecting information about the cost of digital archiving and preservation proved to be as difficult as in the first report, with most of the respondents unable or unwilling to provide cost information. However, several major organizations (OCLC, DSpace, National Library of Australia) are trying value-added services and licensing of software to other organizations as ways of offsetting the cost of preservation activities. Overriding these social, political and economic issues is the need to develop within the scientific and technical community and society in general a culture that encourages the "will to preserve and provide permanent access".

Recommended Next Steps

The work on digital preservation is continuing apace with significant developments in off-the-shelf, generalized digital preservation systems; legal deposit legislation; partnerships and federations; and standards activities. However, much remains to be done. There are activities in which both ICSTI and CENDI can take a lead or become involved that will move preservation practice forward.

ICSTI:

1) Continue to work with the Committee on Data (CODATA), the International Council for Science (ICSU), the individual scientific unions, institutional repositories, and university management on issues related to the archiving and preservation of data and its relationship to publications. A key component of this effort should be identifying the similarities and differences between preserving data (of various types) that results from scientific research and the textual documentation such as journal articles and technical reports. Another key area of investigation is the identification of similarities and differences between preserving data in various disciplines. It will be important to determine what standards can be shared and what must be different.

2) Analyze the impact of Open Access (including author self-archiving), institutional repositories, and e-Science initiatives on digital preservation and permanent access, and identify a framework in which all these initiatives can be successfully achieved.

3) Investigate the usefulness of interoperability standards such as OpenGIS, OpenOffice and STEP for long term preservation formats.

4) Promote the "will to preserve and provide permanent access" as well as best practices by encouraging the incorporation of preservation concepts in science education and work-based training. This might be done in collaboration with science education organizations such as the American Association for the Advancement of Science, learned societies, university management and academic faculty.

5) Produce a list of foundations that are interested in supporting digital preservation in science in order to help those looking for funding. This could be done by polling members of ICSTI and those involved in digital preservation at national libraries and academic research libraries.

CENDI:

1) Work with the US Government Printing Office (GPO), the National Archives and Records Administration (NARA), the LOCKSS-DOCS Project, the Library of Congress (particularly the National Digital Information Infrastructure and Preservation Program [NDIIPP]), and others to develop effective and sustainable preservation guidelines for government scientific and technical information in the context of the R&D (research

and development) component of the Federal Enterprise Architecture, e-Government, the federal research process, and the environments of the federal science agencies.

2) Host a follow-on workshop to the previous workshops sponsored by the National Science Board and the National Archives and Records Administration to continue the discussions about the selection, retention, organization, and on-going preservation of scientific data produced as a result of government funding.

3) Support the development of technical and social solutions to the federation of archives, which are likely to be needed to address the preservation of government information. This would involve the development of core metadata standards for different digital objects, the implementation of high level Open Archival Information System Reference Model functions, and producer-archive interaction checklists specific to the federal science environment.

4) Continue involvement with standards efforts; specifically, review the Journal Publishing DTD (document type definition) and the DTD for Technical Reports, and determine how these efforts might be addressed by the agencies and as part of the Federal Enterprise Architecture.

5) Support the development of technical solutions for archiving the dynamic and deep Web.

1.0 Introduction

ICSTI and CENDI have been interested in electronic archiving and other issues related to the management of digital information since 1996. In 1998, a synopsis of relevant projects, which focused on science, was international in scope, addressed various types of digital objects and included projects at all stages of development, and was determined to be beneficial to the members of these organizations and to the digital preservation community as a whole. Therefore, the International Council for Scientific and Technical Information (ICSTI) and CENDI jointly sponsored a 1999 report on ***Digital Electronic Archiving: The State of the Art and Practice*** (Carroll & Hodge 1999).

Since 1999, both organizations have remained active in issues and discussions related to digital preservation. Based on the findings of the 1999 report, ICSTI and CENDI sponsored a variety of workshops, presentations and articles on digital preservation (Hodge 2000, Hodge 2002, Mahon & Siegel 2002). ICSTI's President, David Russon, made recommendations concerning the importance of preservation in the sciences to the World Science Congress (Russon 1999). CENDI and the Federal Library and Information Center Committee (FLICC) sponsored a workshop on the Open Archival Information System Reference Model in relation to the management of US government information (CENDI & FLICC 2001). Most recently, the need for preservation was included in ICSTI's input to the World Summit on the Information Society (ICSTI 2003). On an ongoing basis, CENDI's Digital Preservation Task Group monitors and reviews best practices and standards as they relate to preservation of the results of science and technology research in the science mission agencies of the US federal government.

Once again, both organizations are joining together to produce this report on the state of digital preservation. The purpose of this report is to determine the new advances and issues in the preservation of scientific and technical information by focusing on operational systems, specifically in the sciences. The goal is to advance the thinking and practice of ICSTI and CENDI members and to provide a basis for further work by others, particularly in the scientific community.

The report begins with a statement of the scope and methodology and an overview of the highlighted systems. The subsequent sections use the highlighted systems and information gathered from experts and from the literature to discuss stakeholder roles, archiving and preservation practices by document type and format type, the workflow established by operational systems, standards activities, and the availability of "off the shelf" systems. The report concludes with a discussion of trends and issues and possible next steps for CENDI and ICSTI.

2.0 Scope and Methodology

The scope of this report is purposefully quite narrow:

- Focus on operational digital preservation systems.
- Focus on science and technology and the digital objects that they create or use, including text, images, video, primary data sets, etc.

- Include, where possible, projects from a variety of countries, disciplines, and sectors.
- Where projects are not solely science- and technology-oriented, attempt to identify the degree to which the archive contains scientific and technical information.
- Discuss relevant activities in standards and best practices, even if they are not solely related to the preservation of science and technology.

The call for participation was sent to several listservs, and the members of the CENDI and ICSTI communities were asked to contribute suggestions for operational systems. In addition, the investigators identified key people involved in digital preservation, attended several meetings on the topic, and performed literature searches. Over 50 systems or projects were identified from these various sources. After initial information was collected, follow up discussion questions (see Appendix I) were used to gather more detailed information.

The survey of operational systems is not intended to be comprehensive. Inclusion or exclusion from the report should not be taken as an endorsement or lack thereof on the part of the investigators, CENDI, ICSTI, or any of their member organizations. The goal is to see what these representative systems can tell us about the state of the practice of digital preservation in science and technology and the outstanding issues, lessons learned, and next steps.

3.0 Highlighted Systems

From the more than 50 systems or projects identified, 21 systems were selected to highlight because of the operational nature of their systems and the potential interest to the scientific community. The highlighted systems represent several countries and international organizations. They are from the government, academic and private sectors. Commercial, learned society, and gray literature publishers are represented. The highlighted systems manage a wide range of scientific resources including e-journals, e-theses, scientific datasets, technical drawings and photographs.

The following table provides key information about the highlighted systems. The information from these more detailed interviews is used throughout the report, along with selected information from other non-highlighted sources.

Highlighted Project	Brief Description	Special Archive Characteristics
American Institute of Physics www.aip.org	Learned society publisher.	Well-established policy and procedures for archiving e-journals. Policy used as model by others.
Aerospace Industries Association/Boeing Co. www.aia-aerospace.org/	Preserving engineering drawings (CAD/CAM) in the aerospace industry.	Developing standards for interoperability of engineering drawings. Working within the consortium to develop standards for preserving the STEP files as the basis for a preservation format.

Highlighted Project	Brief Description	Special Archive Characteristics
Digital Information Archiving System (DIAS) Dutch National Library www.5.ibm.com/nl/dias/	System developed by IBM for the Dutch National Library for deposit of e-journals from multiple publishers.	Dutch National Library set requirements and sponsored development of IBM's DIAS System. Implemented at KB in December 2002. Established as the official archive by Kluwer and Elsevier Science.
DiVA Electronic Publishing Centre, Uppsala University Library www.diva-portal.se/	An electronic publishing system that treats the digital version at the master and creates an archive for long term preservation. Creates institutional archives for theses and dissertations, working reports and other types of born digital documents. Currently used by universities of Uppsala, Umeå, Stockholm, Örebro and Södertörn in Sweden and Statsbibliotek, Århus in Denmark. All full text publications are available through a common interface, known as the DiVA portal.	Archiving is an outgrowth of DiVA's electronic publishing system. Local repositories transmit archival packages directly to the Royal Library (the Swedish National Library) and in particular to the National Bibliography and to satisfy requirements for e-deposit of theses and dissertations. Data originally entered by the author is the basis for the metadata. Metadata is stored in the DiVA document format, a locally developed schema. Transformations of this schema provide metadata in a variety of other formats and support various services including OAI-PMH.
DSpace at MIT www.dspace.org	Institutional archives; MITs implementation is primarily science and technology and is being developed to share the university's intellectual assets.	MIT's implementation of a generalized system for institutional repository development. Heavy use of existing lessons learned and standards such as Open Archives Initiative, Open Archival Information System Reference Model, and Dublin Core. Looking to establish an Alliance that will work on federating the D-Space repositories across institutions. Software is open source.
Elsevier Science Direct – also part of the Dutch National Library Not available	E-journals from a single publisher	First publisher to establish an agreement with the KB to permanently archive the Science Direct journals.
Earth Resources Observation Systems (EROS) Data Center earthexplorer.usgs.gov	Data center with the mission to preserve the remotely sensed, cartographic and topographic records entrusted to the US Geological Survey.	Developed check lists, procedures and an Advisory Committee to help in selection and appraisal of datasets. Currently completing an online decision support tool.

Highlighted Project	Brief Description	Special Archive Characteristics
Fedora™ (Flexible Extensible Digital Object Repository Architecture) Cornell University and the University of Virginia Library www.fedora.org	Can handle a variety of objects; all MIME types. Current applications focus on E-books, XML objects, images at multiple resolution.	Used as the underlying architecture by several systems including DSpace and the University of Virginia Library's Centralized Digital Repository. Flexible container architecture with option to default or customize various aspects of the system. Newest version includes content versioning critical for preservation activities.
International Union of Crystallography www.iucr.org	Learned society publisher with online journals and supplementary data.	Policy for archiving online journals, available with frequently asked questions. Also working with members of their community to ensure better archiving of the data component.
JSTOR www.jstor.org	Journals and conference proceedings (digitized from paper, new initiative on e-journals)	New Electronic Archiving Initiative is bringing together the organizational elements necessary to ensure the long term preservation of and access to e-journals. Initial work is focused on business model and technical infrastructure development.
Life Science Data Archive www.lsda.nasa.gov	Archive of data from NASA's work in the life sciences.	Developing the archive to conform to the Open Archival Information System Reference Model.
LOCKSS (Lots of Copies Keep Stuff Safe) lockss.org	Software system to create preservation copies of journals at various library sites.	New release of software to support synchronization of redundant archives. Generally works within the publishers' current business models. LOCKSS-DOCS project would extend the technology into the US Government Printing Office's Federal Depository Library Program.
NASA Goddard Space Flight Center Library Not available	As part of the Library's support for knowledge management activities, a series of digital preservation projects have resulted in a partially operational system for internal information at the NASA Goddard Space Flight Center.	Capture and storage of a variety of project-oriented materials including web sites, images, project documents and videos. Operational video system that includes webcasts, digital storage, video indexing and segment retrieval. Development of a Goddard Core set of descriptive metadata and a single metadata repository across document/format types.
National Motor Museum Images available at www.heritage-images.com/	Operational system to scan and preserve photos for a technology museum.	Digitizing photos in the collection; following same approach as Profiles in Science
OCLC's Digital Archive www.oclc.org/digitalpreservation/	Text and images submitted by the subscriber	Subscription-based system for making available and preserving a variety of materials via the OCLC's Connexion system and WorldCat.

Highlighted Project	Brief Description	Special Archive Characteristics
PANDORA - National Library of Australia pandora.nla.gov.au/index.html	Long running project to capture web-based publications of Australia.	Development of the PANDAS system, selection criteria and other infrastructure components to support the capture and preservation of Australian publications online. PANDAS system will soon be available to others with trial access. Revised collection guidelines. New efforts in agreements with Australian publishers and government.
Profiles in Science, National Library of Medicine profiles.nlm.nih.gov/	Digital library of papers, photos, audio and video clips and memoirs for noteworthy scientists, particularly Nobel Laureates.	Various digital object types, including audio, video, manuscripts, letters, e-mails, etc. Organizes these into collections for each scientist and, in some cases, links across collections.
PubMed Central, National Library of Medicine www.pubmedcentral.gov	Systems hosted by the NLM to archive journals in the life sciences.	Currently archiving over 120 journals in the biomedicine and life sciences. National Center for Biotechnology Information developed a Journal Archiving & Interchange DTD and a Journal Publishing DTD. Various terms and conditions with publishers – agreement that the current issue can be free either immediately or after a certain period of time.
The Internet Archive/Alexa www.thearchive.org	Not-for-profit organization that takes periodic snapshots of the Internet. About 10% might be scientific and technology related depending on the definition	Snapshots of the Internet as well as focused crawls based on institutional criteria. Working on issues related to the dynamic web and copyright.
US Government Printing Office www.gpo.gov	Government agency responsible for the printing, preservation and distribution of government publications. Includes responsibility for the Federal Depository Library Program that includes regional depositories and a network of various types of libraries to ensure access by the public. Now includes requirements for moving toward a more electronic depository library program.	System to harvest metadata and capture content for government publications from agency web sites. Also involved in helping set requirements for the OCLC Digital Archive. Working with LOCKSS-DOC. Implementing digital signatures to support authenticity.
Victorian Electronic Records Strategy (VERS) – Australia www.prov.vic.gov.au/vers/	Responsible for setting the strategy for electronic records systems in the state of Victoria, Australia.	Well-established system for ingesting and managing e-records. Development of standards.

4.0 Setting the Stage

Before reviewing and analyzing the findings of this research, it is helpful to look at the world in which digital preservation of science occurs. Several aspects of the environment are highlighted below, including current archiving concepts, the scientific environment, technology trends, scientific communications, and the legal deposit and copyright regimes.

4.1 Archiving Concepts and Definitions

A significant shift in the terminology of archiving has taken place since the first report in 1999. The term "electronic" has been replaced with the word "digital", perhaps indicating a shift from concern about electronic journals to the full range of material represented in bits and bytes. While major efforts toward digitizing paper materials continue, there is a clear emphasis on objects that are "born digital." The technical issues of long term preservation are similar once the analog materials have been digitized, but the fact that there is no analog original to preserve makes the "born digital" information all the more fragile.

Another significant shift in terminology is the move away from the word "archiving". This term was problematic from the outset. Those involved with digital information were concerned that "archiving" was too closely identified with records management storage. In addition, the term "archive" had taken on new meanings from e-print and preprint archives, which are primarily repositories with no inherent responsibility for or commitment to long term preservation. The more common term now is preservation, which links this activity to the long history of preservation in paper.

The phrase "permanent access" is usually paired with the term "digital preservation," indicating that preservation is only half the battle. The more difficult issue in the digital environment is how to provide for permanent access and adequate rendering of the object, given the technological changes that have and will continue to occur.

4.2 The Scientific Environment

The goal of e-science is to take advantage of high speed computing and networking to provide virtual laboratories, collaboratoria, and informatics methods to enable scientific discovery. E-science activities by their very nature require digital input and result in digital output that must be managed. Instead of physical laboratory experiments, the investigation is conducted via modeling and simulation approaches that are only available in digital environments and that require systems and networks capable of massive, distributed computer processing. These large network systems are generally referred to as the Grid (National Science Foundation 2003). E-science initiatives are often government sponsored; major initiatives are underway in Japan, the US and the UK.

A global network of e-science centers would result in massive amounts of information. However, the Grid may also provide the basis for a distributed system for archiving and preserving the data, perhaps resulting in more comprehensive data curation (Pothen 2002). The National Science Foundation's Advanced Cyberinfrastructure Program (ACP) emphasizes

the connection between e-science (or digital science) and the need to preserve data and other outcomes from the R&D process. Discussions are underway as to how the various stakeholder groups and new communication mechanisms, such as institutional repositories, might provide the backbone for supporting data preservation and curation (Messerschmitt 2003). In September 2002, the Library of Congress and the San Diego Super Computer Center announced a project to evaluate the Storage Resource Broker Data Grid for preservation of LC's digital holdings (Tooby & Lamolinara 2002; Mayfield 2002; Shread 2002).

4.3 The Technological Environment

Since the publication of the 1999 report, there have been continuing advances in both hardware and software technologies. New processors and operating systems are on the market. Microsoft Office Suite has undergone several upgrades. Windows has seen Windows 2000, Millennium, and XP. Oracle has introduced several versions including 10G. Even in a time of global economic slowdown, technology pressures are ever advancing, causing increased concerns about the future of digital information in a time of limited resources.

Meanwhile, the Internet becomes ever more pervasive. While the rate of growth of available content on the public Web has slowed and recent research suggests that the Web may have decreased in size (OCLC Office of Research 2003), the Web still includes a vast amount of information. One might speculate that some of the scientific information has gone underground; i.e., into the deep Web. More scientific information may be Web-enabled, but hidden in databases, behind firewalls, or on institutional intranets. Concerns about national security, cyberterrorism, and the frequency of cyber attacks have made the archiving process more difficult (Kahle 2003).

4.4 Scientific Publishing and Communications

There are many factors of scientific communication and publishing that impact the digital preservation environment. Open access and institutional repositories are highlighted here.

4.4.1 Open Access

One of the major changes impacting the future landscape of scientific communication and publishing is the advent of open access initiatives. Open access asserts that scholarly materials, particularly those in the sciences, should be available for free to users and institutions, with the need for new business models on the part of publishers.

"By 'open access' to this literature, we mean its free availability on the public internet, permitting any user to read, download, copy, distribute, print, search, or link to the full texts of these articles, crawl them for indexing, pass them as data to software, or use them for any other lawful purpose, without financial, legal, or technical barriers other than those inseparable from gaining access to the internet itself. The only constraint on reproduction and distribution, and the only role for copyright in this domain, should be to give authors control over the integrity of their work and the right to be properly acknowledged and cited. Open access eliminates two kinds of access barriers: (1) price barriers, and (2) permission barriers associated with

restrictive use of copyright, licensing terms, or DRM [digital rights management]." (Budapest Open Access Initiative 2002)

The early efforts in open access to biomedical literature (eBioMed), led by Dr. Harold Varmus of the US National Institutes of Health and others, spurred a number of open access statements and initiatives including the Budapest Open Archives Initiative (2002) quoted above, the Public Library of Science (2003), and the Bethesda Statement on Open Access (2003). Open Access initiatives in the sciences are particularly strong in developing countries (CODATA 2003).

US legislative actions, such as H.R. 2613 - "The Public Access to Science Act" (also called the Sabo Bill), may have a major impact on open access to science in the US. The Sabo Bill requires authors who receive federal funds to deposit their work in an open depository and make the information generated from these efforts free of copyright. In a related bill passed in 2001, researchers who receive federal grants must make their data sets publicly available within certain time limits.

The impact of open access can already be seen in the sciences. The Directory of Open Access Journals, maintained by Lund University Libraries and sponsored by the Information Program of the Open Society Institute and SPARC (Scholarly Publishing and Academic Resources Coalition), includes over 350 open access journals in 15 subject categories. The scientific categories include Agriculture & Food Sciences, Biology & Life Sciences, Chemistry, Health Sciences, Earth & Environmental Sciences, Mathematics & Statistics, Physics & Astronomy and Technology & Engineering (DOAJ 2003). Some of these journals are alternatives to the more expensive commercial journals in various disciplines developed by open access publishers such as BioMed Central, SPARC partners, and some institutional repositories. These organizations may also act as trusted third parties for other publishers who are willing to deposit their materials in an open access arrangement with terms and conditions.

Open access may appear to be a boon for digital preservation in the sciences. However, many open access initiatives are based only on the immediate desire for access. "The major open-access initiatives differ on whether open access includes measures to assure long term preservation. For example, the definitions used by BMC [BioMed Central] and the Bethesda Statement include this element, but the BOAI [Budapest Open Access Initiative] and PLoS [Public Library of Science] definitions do not. Taking steps to preserve open-access literature directly answers an objection often raised against open access. This makes it both desirable and important for open-access initiatives to take steps to preserve their literature and to say so prominently. The need for prominent mention often brings the mention right into the definition of "open access". But none of this means that preservation is part of open access, merely that it is desirable. Is preservation an essential part of openness or a separate essential?" (Suber 2003)

Suber (2003) advocates that long term preservation is only one of several desirable requirements for open access, along with deposit in an archive or repository, but that preservation and openness are not inherently linked. "...By bundling them [preservation and openness] all under the concept of openness, we risk blurring or over-burdening our simple concept and we risk delaying progress by multiplying the conditions that our initiatives must

meet." So while open access can act as a catalyst for addressing long term preservation without the restrictions of copyright, open access may also focus on immediate dissemination rather than long term preservation goals.

4.4.2 Institutional Repositories

A definition generalized from Lynch (2003) defines an "institutional repository" as a set of services that the institution offers to the members of its community for the management and dissemination of digital materials created by the institution and its community members. The primary impact of the institutional repository movement has been in academia, spearheaded, in part, by university management. After collecting theses and dissertations, many academic institutions have begun to broaden the types of materials included in their repositories to include virtually all materials of long term value that are produced by faculty, staff, or employees. The Association of Research Libraries produced a position paper on the growth of institutional repositories and the types of infrastructure, which gives six examples of institutional repositories (Crowe 2002).

In the government sector, there have been institutional (or one might call them enterprise) repositories for decades. In the US, science mission agencies with institutional repositories include the Defense Technical Information Center, the Department of Energy's Office of Scientific and Technical Information, and the NASA Center for AeroSpace Information. On a government-wide scale, the National Technical Information Service and the Government Printing Office also have responsibility. In France, Institut de l'Information Scientifique et Technique (INIST-CNRS) is responsible for similar activities. Other countries have similar organizations with varying authorizations to collect, preserve and disseminate scientific and technical information for their respective enterprises.

In the 1990s, these organizations began to collect technical reports, reprints, and other text materials in electronic form, and to add certain types of non-print materials to their collections. In the last several years, dissemination of the full text has shifted from print and microfiche to e-mail or FTP downloads and Web access. Many of these materials are now received and stored electronically, and a wider range of materials is being collected, resulting in large repositories of digital information that must be preserved.

While many of these institutional and enterprise repositories have a history of preserving paper, they are increasingly conscious of the responsibility of being a repository in the digital environment. "[An institutional repository] is most essentially an organizational commitment to the stewardship of these digital materials, including long term preservation where appropriate, as well as organization and access or distribution." (Lynch 2003)

The arguments and issues related to the long term preservation of e-print and other institutional archives are outlined in Penfield & James (2003). They conclude that depending on the circumstances both filling the repositories, i.e., focusing on content, and long term preservation should be part of building an open archive. The issues, feasibility and requirements for e-print preservation have been identified by the Arts and Humanities Data Services (AHDS) SHERPA (Securing a Hybrid Environment for Research Preservation and Access) Project sponsored by

the Joint Information Systems Committee (JISC) and CURL (Consortium of University Research Libraries) in the UK (James 2003).

4.5 Legal Deposit and Copyright

The goal of legal deposit is to ensure that access to a nation's published works is preserved in libraries and archives. "A statutory obligation which requires that any organization, commercial or public, and any individual producing any type of documentation in multiple copies, be obliged to deposit one or more copies with a recognized national institution." (Lariviere 2000) Its principle is established in international convention and in the national legislation of many countries.

Digital information requires active management to ensure that a complete record of a nation's published material exists for the future. If legal deposit is applied to digital information, the protection of publishers' rights and investments needs to be considered, since the potential for multiple accesses to a single digital information object is an issue. Another issue is the differing nature of digital information from that of its traditional physical counterpart (PADI 2003).

"Legal deposit legislation in many countries predates the current information age and requires a new legal framework in order to encompass digital publications. The complications associated with the collection and control of electronic materials, together with the lack of a comprehensive legal model, have made drafting appropriate legislation problematic and slow. Major issues to be considered include copyright, preservation requirements, public access, scope of coverage, method of collection, protection of publishers' rights, penalties, and implementation of revised legislation." (PADI 2003)

Digital legal deposit has undergone significant change over the last several years. The major initiatives are outlined below. Information and quotes are from PADI (2003) unless otherwise noted.

Countries that have enacted legislation that covers physical format and online forms of digital publications or that have a legislative process in place include: Canada, Denmark (static online publications), New Zealand, Norway (static online publications), South Africa, and the United Kingdom.

- Canadian legal deposit legislation has been extended to include electronic publications issued in physical formats. The National Library of Canada has continued to collect electronic publications on a voluntary deposit basis, with emphasis on publications not available in any other format.

- Denmark's deposit legislation states that "all published material is subject to legal deposit, regardless of the production technique or type of carrier." Emphasis has shifted from printers of documents to publishers of documentary materials in the broadest sense, including physical format digital and static Internet publications. "The Royal Library of Denmark acts as the deposit institution for Danish maps, electronic products and Internet

publications. A legal deposit registration system for downloading deposit documents has been created in collaboration with UNI-C, a government data research institute."

- New Zealand's 2003 legislation applies to public documents issued in print or in electronic physical or online form. It specifically provides for the copying of Internet documents. "Until the Requirement relating to electronic documents comes into force, electronic documents in physical format continue to be purchased or obtained by voluntary deposit through standard acquisition processes. Currently the [National] Library is developing its processes for the selection, acquisition, harvesting, description, storage and provision of access to physical format and online electronic documents."

- Norway's Legal Deposit Act has cultural preservation as its primary intent. Physical format electronic documents and static Internet documents are included, but dynamic electronic resources are not. However, the legislation includes any works which can be read, heard, broadcast or transmitted, and is written in a way to be applicable to future electronic formats. (Van Nuys 2003; PADI 2003)

- Norway is one of five countries involved in Web archiving that can base its work on legal deposit legislation. Countries that have started some type of Web archiving activity include: Denmark and Australia (selective collection strategy); Sweden, Iceland, and Finland (harvested entire national web spaces); while the National Library of the Netherlands has made an agreement with the Dutch Publishers' Association (NUV) for deposit of electronic publications offline and online (Van Nuys 2003).

- The National Library of Norway is investigating ways to fulfill the intent of the act as applied to digital documents and is considering using a combination of different collection approaches. The Paradigma Project, which began in August 2001 and will end in December 2004, will "develop and establish routines for the selection, collection, description, identification, and storage of all types of digital documents and to give users access to these publications in compliance with the Legal Deposit Act." (Van Nuys 2003)

- In South Africa's Legal Deposit Act of 1997, the definition of 'document' and interpretation of the term 'medium' enables the Act to apply to electronic publications available in both physical format and online. Due to the technical and administrative challenges associated with the deposit of dynamic electronic publications, online electronic materials are presently only subject to deposit when specifically requested by the State Library of Pretoria."

- In the UK, the Code of Practice for the Voluntary Deposit of Non-Print Publications came into effect in 2000, endorsed by various UK publisher trade bodies and legal deposit libraries. The arrangement provided for the deposit of microfilms, physical format digital publications and other offline electronic media, but the challenges for the deposit of static and dynamic online publications were also recognized in the guidelines. Subsequently, the Legal Deposit Libraries Act 2003 became law in October and will ensure that works published in non-print format will be collected. Categories of non-print

materials that will be collected and saved include electronic journals and other materials accessed over the Internet; a limited range of research-level web sites; microforms such as film and fiche; as well as CDs, DVDs and other "hand-held" electronic media. The Act will be implemented through a series of regulations, and it is anticipated that the first set of regulations will deal with offline publications such as CDs and microform material (British Library Press & Public Relations 2003).

Countries that have legislation in place that currently applies to physical formats but not to online digital publications include: Austria, France, Germany, and Sweden. Physical format digital material refers to information that is digital and stored on transportable media such as floppy disks, magnetic tape, CDs, and DVDs. Further detail follows:

- Austria's response to a legislative gap for the deposit of online and networked digital material is the AOLA (Austrian Online Archive) project, established to investigate the challenges associated with the collection and archiving of online publications.

- The revised French legal deposit legislation requires legal deposit of documents regardless of the technical means of production, as soon as they are made accessible to the public by the publication of a physical carrier. Legal deposit of CD-ROMs has been enforced since 1994, but, to date, "deposit provisions do not cover online electronic publications, and no incentives exist for the voluntary deposit of non-physical format digital materials."

- In Germany, publishers are required to deposit copies of their publications, including physical format digital materials.

- Sweden's legislation requires legal deposit of electronic documents available in physical format, such as optical disks. Online electronic documents, like those found on the Internet, are not covered by this legislation. The Royal Library of Sweden's Kulturarw3 (Cultural Heritage Cubed) project is investigating preservation of published electronic documents; it collects electronic information through harvesting.

Where legislation is not in place, national libraries and publishers are negotiating voluntary deposit schemes as a means of collecting digital publications. "Current trends suggest that in some instances these voluntary codes will become permanent, especially where governments prove reluctant to change laws and if legal deposit is afforded a low priority for amendment."

The Netherlands does not have legal deposit legislation and relies on voluntary deposit based on bilateral agreements with publishers. These deposit arrangements have also been negotiated for digital information. Other voluntary deposit efforts either currently operate or are under development in Canada, Germany, the United Kingdom and Australia. "In addition, a model code has been developed by the Conference of European National Librarians and the Federation of European Publishers to facilitate the drafting of locally-endorsed voluntary deposit arrangements."

- In Australia, the Copyright Amendment (Digital Agenda) Act 2000 made no changes to the existing provisions. To cover the gap in federal legal deposit law, the National Library of Australia (NLA) has implemented an interim Voluntary Deposit Scheme for Electronic Publications, together with a Policy on the Use of Australian CD-ROMs and Other Electronic Materials Acquired by Deposit. While Commonwealth statutes don't include electronic publications, some states, such as Tasmania, have legislation that includes some digital components.

A recent study by Charlesworth (2003), sponsored by The Wellcome Trust and the JISC, addresses the legal issues related to archiving the Web. Charlesworth notes that the most obvious "legal stumbling block" is copyright law, but cautions that there are also hazards regarding defamation law, content liability and data protection depending on the countries regime in these areas. However, he believes that the issues are not insurmountable, with careful selection of the sites to be archived, effective rights management policies, and good access rights mechanisms.

5.0 Stakeholder Roles

Previous investigations of digital preservation have identified numerous stakeholder groups involved in digital preservation. Flecker (2002) identified discipline-based models, commercial services, government agencies, research libraries, and passionate individuals. Lavoie (2003) reduces the stakeholder roles to rights holders, archives, and beneficiaries. The previous ICSTI report identified creators/producers, publishers, libraries and library consortia, funding agencies and users (Carroll and Hodge 1999).

The following section describes the preservation activities by publishers, national libraries, institutions and their libraries and museums, archives, and trusted third parties. It also discusses the role of governments.

5.1 Publishers

A study sponsored by the Association of Learned and Professional Society Publishers showed that 52 percent of commercial and 45 percent of not-for-profit publishers interviewed have formally addressed long term preservation of their publications with most taking on the responsibility themselves. Third party archives such as JSTOR, OCLC, and HighWire LOCKSS are used. Discipline-specific depositories such as PubMed Central were found to play only a minor role at present (Cox 2003).

Commercial publisher initiatives are coming from two major impetuses. First, these materials have intellectual property value that benefits the publisher if the materials remain under the control of the publisher. Secondly, publishers have begun to realize the economic benefits of the reuse of the content. This is especially true as mark-up languages and XML schema are used that allow material to be extracted, merged, integrated and even provided to users on-demand through Web-based content models. Many publishers have SGML/XML-based

systems that provide preservation-oriented formats as a natural outcome of their publishing processes.

Wiley's DART (Digital Assets Repository Technology), for example, has three major priorities (Morgan 2000). These are digital printing (including distributed and on-demand), creation of electronic versions of existing paper products so that they can be more easily provided on the Web or to online retailers, and creation of new products, such as coursepacks, based on the re-use of previously published material. The specific goal of the metadata designed into the DART system is to support Wiley's commercial priorities.

Many learned society publishers consider preservation to be an extension of their mission to preserve the knowledge of their discipline, justifying the resources committed to these activities. Many of these society publishers have been at the forefront of preservation activities for both text and data and instrumental in raising the awareness among the researchers in their respective disciplines.

The American Institute of Physics (AIP) advertises archiving services as one of the Composition Services in its Electronic Journals Platform (OJPS) (American Institute of Physics 2003). AIP performs rich mark-up in SGML or to the customer's specific DTD. In the near-term, AIP can supply files in a variety of formats including Postscript, PDF, and SGML. The files include graphics and RGB files for color work. Dissemination is available via FTP, CD-ROM, or 8mm tape. In addition, authors can request their articles in a variety of formats appropriate for inclusion in conference proceedings, books or other reprint vehicles. In the long term, AIP's ASCII-based format is reliable for future preservation and reuse.

Based on the AIP model, the International Union of Crystallography (IUCr) has published a policy on long term preservation and access. It also utilizes the concepts and terminology of the OAIS Reference Model (International Union of Crystallography 2001). The policy specifically covers IUCr's online journals, but the intent is to extend it to other types of materials available from the union's web site. The policy is only partially applied; the IUCr has taken steps to create local offline copies of the journals in SGML as well as in HTML and PDF. However, this is primarily aimed at short-term disaster recovery. IUCr intends to pursue partnerships with major public crystallographic databases for preservation of the data, since there is a close relationship between the text publishing and the data activities. This involves working with CODATA to raise awareness of the need for these databases to develop their own preservation strategies.

In 2001, the International Union of Pure and Applied Physics (IUPAP) held a conference that brought together publishers, researchers and librarians to discuss the long term preservation of digital documents in physics. Two recommendations resulted from the meeting -- the development of a registry of physics archives that would include information about hardware and software so that it could serve as an early warning about possible need for migration or data at risk, and the creation of a subgroup to investigate the use of XML and other format standards as applied to physics documents (Smith 2001). In addition, IUPAP has encouraged its member societies to develop XML schema and standards appropriate to their disciplines (Butterworth 2003).

5.2 National Libraries

National libraries were given a major role in the 2002 joint statement between the International Publishers Association (IPA) and the International Federation of Libraries and Archiving Institutions (IFLA 2002). This statement sets out several key points, including the importance of digital information and the fact that it is severely at risk under the current circumstances. Successful, long term archiving and preservation will require a partnership and neither the libraries nor the producers of the information can adequately archive alone. Ultimately, the most appropriate stakeholder to manage the long term preservation of digital materials is the national library infrastructure. National libraries are already trusted third parties and digital preservation is an extension of the mandate of legal deposit in the analog environment. IFLA and the IPA have also agreed to continue joint activities including technology research and searching for funding opportunities.

The new relationships between publishers and national libraries may be the result of publishers, particularly commercial publishers, determining that their missions are better served by focusing on the initial publication and dissemination of the material than on long term preservation. The initial wariness on the part of publishers may have subsided, particularly among those publishers who participated in pilot projects over the last several years. The majority of these pilot projects have proven successful and seemed to have produced a symbiosis of the needs of these publishers and the needs of libraries. Also, long term preservation became such an issue with the publishers' constituents, primarily the libraries, that preservation arrangements were necessary.

Many major publishers have signed agreements with national libraries as trusted third parties. After developing its own electronic warehouse, Elsevier determined that it needed to partner with others (Hunter 2003). Elsevier identified KB (The National Library of the Netherlands) as its official archive based on KB's technical competence. The formal arrangement addresses permanent retention and international access. The archive holds those articles that are withdrawn as well as those that are active. The archive will be available for KB walk-in users only. Elsevier emphasizes that this archive is not a hot backup for the company's data recovery, but it could be used to support recovery from a truly catastrophic event. The intent is to use the KB agreement as a model for two to three negotiations with other national libraries.

In May 2003, Kluwer Academic announced an agreement with KB to serve as the archive for the journals featured on Kluwer Online. Kluwer Online contains over 235,000 articles from over 670 journals. In September 2003, an agreement was signed with BioMed Central to archive its 100 open access journals and the other deposited materials (BioMed Central 2003). Unlike agreements with other publishers, KB's remote users as well as walk-in users will have access in accord with BioMed Central's open access philosophy. The KB is seeking to enter into agreements with other major scientific publishers.

The National Library of Australia was an early investigator of digital preservation methodologies and support tools. PANDORA (now officially known as PANDORA: Australia's Web Archive) is national in scope with all the mainland State libraries, ScreenSound Australia, and the Australian War Memorial as partners. (The State Library of

Tasmania continued to develop its own archive, Our Digital Island, (http://odi.statelibrary.tas.gov.au/)). PANDORA now contains over 4,000 titles and over 8,000 instances (Phillips 2003). (An 'instance' is a single gathering of a title that has been added to the archive. Many titles are re-gathered on a regular basis to capture changing content, for example, when serial titles add new issues.) The Archive consists of approximately 16 million files and the display copies alone occupy almost 500 gigabytes of storage space. (There are two additional copies for preservation purposes, as well as back up copies.) The Archive covers the full range of material published online in Australia, including science and technology (266 titles), agriculture (210), health (214), computers and the internet (157).

A major area of preservation for some national libraries is electronic theses and dissertations. Major operational systems are in place in Denmark, Sweden, and India. (The activities in Denmark and Sweden are described in Section 6.2.) Since July 1998, Die Deutsche Bibliothek (the National Library of Germany) has collected online dissertations and theses. The university libraries report electronic dissertations to Die Deutsche Bibliothek and then they are stored to the library's archive server DEPOSIT.DDB.DE. Since February 2001, Die Deutsche Bibliothek has hosted the "Co-ordination Agency DissOnline". Die Deutsche Bibliothek is planning an e-deposit system that will eventually hold and preserve not only dissertations but electronic journals, web pages, and other materials considered to be of preservation value (Steinke 2003).

5.3 Institutions

Institutions, particularly major research universities and their management, are becoming major players in preservation activities (Lynch 2003), perhaps as an outgrowth of the development of institutional repositories and the availability of open source software such as DSpace and the Open Archives Initiative-Protocol for Metadata Harvesting (OAI-PMH). While not every institutional repository is committed to a long term archive, there are key relationships between producers and the repository that are similar to those identified in the Producer-Archive Interface Methodology Abstract Standard draft (CCSDS 2002) that can create a natural pathway between short term and long term commitment. Lynch (2003) posits that "Only an institutionally based approach to managing these data resources, which operates in alignment with what the faculty at each individual institution are actually doing, can provide a comprehensive dissemination and preservation mechanism for the data that supports the new scholarship for the digital world. Journals will move too slowly and too unevenly to manage these resources, and disciplinary data repositories cannot be comprehensive. Institutional repositories can maintain data in addition to authored scholarly works. In this sense, the institutional repository is a complement and a supplement, rather than a substitute, for traditional scholarly publication venues."

The DSpace at Massachusetts Institute of Technology (MIT) implementation includes submissions from a number of MIT departments, including Ocean Engineering and the Laboratory for Information and Decision Systems. Each department is treated as a community and then programs can cluster under each community. It is possible to search across the communities or to select a community for searching or browsing by author or title. DSpace expects to add other communities over the next year (Tansley 2003). The resources in DSpace at MIT include preprints, technical reports, working papers, conference papers, learning objects

and e-theses, which may include audio, video, text and data sets. In addition, the Out of Print Books of MIT Press are available to MIT staff and students via this site.

As part of the NASA Goddard Space Flight Center Library's mission to preserve and provide ongoing access to information of value to Goddard project managers and researchers, the Library conducted several pilot projects in digital preservation. The focus for the Library is on internal project-related materials; the objects include videos of colloquia, seminars and internal mini-courses, intranet web sites with scientific and technical content, project documents, and images, including photographs and animations. The web site, image and video repositories have been demonstrated separately. The metadata have also been combined into a central repository so that users can search across object types.

5.4 Museums

Museums are taking an increasingly active role in digital preservation. Most museums are interested in digitization as a way to make artifacts more accessible, particularly those artifacts that are rare and fragile. In addition, digitization provides support for curation and restoration activities, for insurance and disaster recovery. While the majority of the museums do not deal with born digital objects, they provide significant digital resources for scientific investigation, valuable access points to materials that are physical and which, therefore, can "reside in only one place," and "benchmarks" for various scientific investigations and analyses, as in the case of taxonomic voucher specimens.

Museums also provide significant insight into the development of non-text digital repositories. For example, the digitization project at the National Motor Museum in the UK is part of a funded project to retrospectively document the photographic collection of the museum. The goal is to digitize the entire collection, but the current emphasis is on the 250,000 images in the working collection. The photographs are digitized as their physical storage is being re-allocated. During this re-allocation process, the 'original' or 'first generation' prints that have copies are removed from the working collection, digitized, documented and stored in a secure environmentally controlled environment. While the current effort does not include the development of a dedicated web site, the digitization methodology was designed to ensure that images created during the process can be made accessible via the web.

As an outgrowth of individual and collective work with digital objects, museums are using the digital environment to create online exhibits. These activities combine multiple media, including images, text, video and sound to support a museum's outreach and educational missions. The complexity of many of these online exhibits provides particular challenges to digital preservationists, including the need to link the digital item to its physical artifact.

5.5 National, State and Regional Archives

While archives approach preservation through different practices and approaches, they may also provide significant repositories of records related to scientific and technical endeavors. National, state and regional archives have been very active in the area of preservation technologies and practices. Their work is particularly important because it must deal with

massive quantities of information, in a wide variety of formats. Key activities are underway at the National Archives and Records Administration (US), the National Archives (UK), and the Public Record Office of Victoria (AU). While the material is generally managed by collection or class of item and an emphasis is placed on the "original order" of the e-records, the distinctions between collection and individual item are becoming increasingly blurred since access to individual items is more easily provided when the material is in digital form.

There may be particular similarities in the preservation issues of archives and the preservation issues of data centers and other scientific and technical enterprises that create massive amounts of data. Both communities must establish practices related to selection or appraisal and retention, since to keep everything may not be feasible. A recent analysis performed for the development of appraisal guidance from the National Archives and Records Administration (NARA) for US government agencies identified special issues related to scientific data. NARA is holding meetings with the scientific data community to determine the needs of this community for archival level appraisal and retention for other types of data and scientific information. Guidance for the retention of observation data from the physical sciences is provided as a special case scenario in recent appraisal guidance related to NARA's strategic initiatives (NARA 2003c).

5.6 Trusted Third Parties

Preservation may also be performed for content holders by trusted third parties. Trusted-third-parties are organizations that provide preservation services without being publishers, owners or subscribers to the materials preserved. Activities such as those of the Research Libraries Group/Commission on Preservation and Access Task Force on Archiving of Digital Information (RLG 1996) and the RLG/OCLC Working Group on Digital Archive Attributes (RLG 2002a) have helped to lay the foundation for current and evolving work on third party archiving activities. RLG and the US National Archives and Records Administration are co-creators of a task force on digital repository certification, whose resulting work is intended to go into the international standardization process through the ISO Archiving Series (RLG 2003). The trusted-third-parties highlighted below include a national library and two not-for-profit organizations.

The PubMed Central Journal Archive is an archive for life science journal literature established by the National Library of Medicine (US). It is available as a trusted third party for any qualified journal publisher (not just from the US) to deposit the electronic versions of journal articles. As of October 2003, the archive contained approximately 135 journal titles with others waiting to be included. One of the major contributions of PubMed Central has been the establishment of best practices for formats, mark-up, and e-journal selection.

The JSTOR operational archive of journal backruns and digitisation of paper journals, consists of six topical collections, including General Science and Ecology and Botany. As of July 2003, these collections included over 400,000 articles. As of August 2003, the Ecology and Botany collection included 29 titles and General Science included seven. As an extension to its digitisation services, the JSTOR's Electronic-Archiving Initiative is charged with developing the organizational and technical infrastructure necessary to ensure the long term

preservation of and access to electronic journals (JSTOR 2003). Areas of consideration include business models, governance, technical infrastructure, metadata formats, and management of supplemental information. Key decisions will be needed concerning the development of an approach that balances the needs of scholars, publishers and libraries. A pilot project is currently underway with a start-up grant from the Mellon Foundation. It involves ten publishers, including several major science publishers. Contributing publishers will submit samples during the summer and fall of 2003, and the goal is to have a prototype when the grant period concludes in March 2004.

The Internet Archive (2003) is a non-profit organization that takes periodic snapshots of the Web, and makes them available to the public. In addition, there are several large institutional customers that use the Archive as a service bureau to create snapshots of the web for them. Broad crawls of the web are done approximately every two months. Focused crawls are performed more frequently. The rules for selecting sites to archive depend on the client and are more precise for partners such as the British Government, the Israeli Government and the Library of Congress. Currently there are discussions underway with the National Archives and Records Administration in the US and the UK. National Archives (previously the Public Records Office). Agreements with other national libraries and archives are likely. The Internet Archive provides the data from its crawls as a corpus for special projects (i.e., the investigation of web surfing patterns by Xerox PARC, the 1997 snapshot of the Web at the Library of Congress, and the 1996 US Elections pages displayed by the Smithsonian).

5.7 The Role of Government

The role of government, while it varies from country to country, has focused on direct funding through national libraries, national archives, and government institutional repositories and on indirect funding of non-government initiatives and public-private partnerships. Governments have also been instrumental in funding research and establishing appropriate policies that encourage or contribute to an infrastructure for digital preservation. In many cases, e-government legislation includes establishment of archiving and preservation initiatives. Many of these activities directly involve scientific and technical information.

Early preservation research was funded through the European Union's Information Society Directorate and its focus areas are electronic publishing, digital culture and library telematics. The system for archiving the Elsevier Science journals is funded by the Dutch Ministry of Education. The Congress of the United States has appropriated $25 million in funding for the development of a strategic plan for an infrastructure for preservation of digital objects through the Library of Congress' National Digital Information Infrastructure for Preservation Program (NDIIPP 2003). Five million dollars are to be spent during the initial phase for planning and also for acquiring and preserving digital information that would otherwise vanish in the interim. The full amount of the funding is $99.8 million with $75 million available as the amount is matched by nonfederal donations, including in-kind contributions. The first call for proposals was announced in late 2003. In addition, programs like the Library of Congress' MINERVA Project have been critical in helping to determine the nature of and potential solutions to problems in web capture.

Governments also establish supportive environments through legislation and directives that require collection of digital materials or remove barriers to collecting. Many data centers, including the US Earth Resources Observation Systems (EROS), are authorized through legislation.

E-government legislation in various countries has included digital preservation components. The E-Government Act of 2002 in the US addresses issues of long term preservation (though this was significantly reduced in the final version of the bill). The creation of the E-Envoy position in the UK is indicative of the degree to which e-government is embraced in that country. There is a significant effort to move publications, transactions and communications of all types and from all levels of citizenry and government to government to an electronic environment. In Australia, the e-government policies established an infrastructure that specifies critical components for a digital preservation environment including metadata standards (Dublin Core) and persistent identifiers.

5.8 Foundations and Other Private Funding Sources

Foundations and other private funding sources have been instrumental in providing the funds needed to "jump start" activities in the area of digital preservation. Digital preservation and long term access is a public good and, therefore, the heavy investment required is hard for industry, academia and even the government to justify. Foundations have been part of many innovative partnerships in this area.

The Andrew W. Mellon Foundation over the last several years has supported a wide range of research and pilot projects through its Scholarly Communication and Research in Information Technologies Programs (Andrew W. Mellon Foundation 2003). Early projects included the development of the initial JSTOR pilot, which resulted in an operational and actively used system for the digitization of backfile journal issues, including a large number in the sciences. Mellon continues to support the effort through its funding of JSTOR's analysis of the impact of e-journals on JSTOR's activities. Following initial funding in the area of digitization of paper journals, Mellon became heavily involved in funding major projects related to the archiving and preservation of electronic journals, including projects at Harvard and Yale. While many of Mellon's activities have been irrespective of discipline, there has been significant involvement in Mellon Projects on the part of major scientific publishers, such as Elsevier Science, and major scientific research libraries such as MIT, Harvard and Yale. Mellon's more recent activities include funding an investigation into the preservation of government documents by the California Digital Library and supporting the continued development of Fedora, DSpace and LOCKSS. All these projects are discussed in more detail in subsequent sections of this report.

The Wellcome Trust, an independent research institute focused on human and animal health, has funded similar initiatives (Wellcome Trust 2004). The Wellcome Trust and the Joint Information Systems Committee (JISC) co-sponsored an investigation of web archiving by UKOLN (Day 2003). While the report focuses on the needs of the Wellcome Trust Library and JISC, it has applicability to all organizations interested in the issues and complexities of archiving the web.

A critical point for digital preservation projects is the point at which research and pilot activities move into an operational phase. Generally, support from a foundation is reduced or eliminated when the project reaches this phase. The Mellon Foundation has been particularly aware of this problem and required sustainability planning and an analysis of ongoing costs as part of its research projects. Not only has this approach recognized sustainability and cost as key issues for digital preservation, but this practical focus from the outset has resulted in better planning, appropriate expectations on the part of the stakeholder groups, and proven, long term outcomes from the investment of foundation monies.

Similarly, the Wellcome Trust funded research into another very practical issue in digital preservation, the issue of copyright, particularly when archiving and preserving web-based resources. The study co-sponsored by JISC and conducted as a companion to its more technical report, discusses copyright in the UK, EU, Australia, and the US (Charlesworth 2003).

Despite this significant support by these key foundations, Neil Beagrie of JISC (Beagrie 2003) noted that it is difficult to identify funding sources for digital preservation activities in science. He indicated that a list of foundations and remits would be a valuable tool for those trying to identify funding sources.

6.0 Preservation by Document Type

There are many document types or genres that are important in scientific communication. These include journal articles, books, theses and dissertations, conference presentations and papers, and project documentation. These document types may be presented as Web sites and they may also qualify as electronic records. These genres may include multiple format types. For example, electronic journals may require supplemental files such as spreadsheets, videos, or software.

This section discusses preservation practices by document types. More information about specific format types is included in Section 7.0, *Standards by Format Type.*

6.1 Electronic Journals

Electronic journals have been at the forefront of preservation discussions because of their critical role in scientific communication and the commercial interests involved. The practices for preserving electronic journals show an increased maturity, as evidenced by more formalized procedures such as a DTD for journals.

In 2001, the Mellon Foundation funded a study at Harvard University to investigate whether a common DTD could be developed for journals (Inera 2001). The study indicated that a common DTD could be developed but that there would be some loss in specificity, particularly in certain areas as math and chemistry. It also suggested the extension of previous work at the

National Library of Medicine's National Center for Biotechnology Information on an XML format for archiving material deposited in PubMed Central (PMC).

This previous work at PubMed Central began in 2000 with an attempt to create a common DTD across two publishers. It soon became apparent that updating this DTD every time a new publisher was added was not the optimal situation. PubMed Central decided to create a more generalized DTD for journal articles.

The Archiving and Interchange DTD Suite is based on an analysis of all the major DTDs that were being used for journal literature, regardless of the discipline. The suite is a set of XML building blocks or modules from which any number of DTDs can be created for a variety of purposes including archiving. Using the Suite, NLM created a Journal Archiving and Interchange DTD, which will replace the current PMC DTD as the foundation for the PubMed Central archive. In addition, a more restrictive Journal Publishing DTD has been released which can be used by a journal to mark up its content in XML for submission to PubMed Central. Several publishers and projects, such as JSTOR, the Public Library of Science, High Wire Press and CSIRO, are analyzing or planning to use the Journal Publishing DTD (Beck 2003).

In addition, an XML Interchange Structure Working Group was created to recommend changes and additions to the tagset. On November 1, 2003, Version 1.1 of the DTD was released. Work is beginning on other special DTD's for online books and documentation based on the suite modules (Beck 2003).

6.2 Theses and Dissertations

Many of the institutional and national library preservation efforts involve theses and dissertations, since these institutions often have responsibility for providing this genre to their respective national library for incorporation into the national bibliography.

One of the most advanced preservation projects is DiVA at the Electronic Publishing Centre of Uppsala University in Sweden. The DiVA system treats the electronic copy as the "digital master" for both electronic and print versions of the document. Local repositories at five universities create archival copies as part of the publishing process. These archival copies are provided to the Royal Library, the National Library of Sweden, as archival packages, in a system that uses a federation of remote libraries to provide full text and metadata to the national library for long term preservation purposes via e-deposit.

Local repositories such as that at The Royal Technology Library (KTH) in Sweden are working on the local repositories with the expectation of participating in the DiVA workflow when it is finalized. The goal of the effort at KTH is to create a campus archive of KTH publications, particularly dissertations, that promotes access and re-use. KTH began with abstracts for the dissertations in 1997. They receive approximately 250 dissertations per year. Preservation is an ultimate goal, so, along with the DiVA Project, they will be working to contribute the electronic publications to the National Library in Stockholm.

The National Library of Germany has developed DissOnline to provide access to the theses and dissertations of that nation. Eventually, the DissOnline collection will become part of its deposit system where long term preservation will be addressed (Steinke 2003).

Because of the nature of the authoring environment, most theses and dissertations are received in PDF, HTML, Word or TeX/LaTeX format. Many national libraries are still retaining the native format rather than transforming the original into a preservation format. In addition, many have hybrid systems where they preserve both the paper and the electronic because they are mandated to do so. The availability of these theses and dissertations to the public via the web depends on the copyright regime of the individual country.

6.3 Scientific Data Sets

Data was the earliest digital output of science to be archived. Through large data centers such as the NASA Distributed Active Archive Centers (DAACs), the data centres of the UK, and the World Data Centers, a variety of important, often non-reproducible datasets have been collected, stored, managed and made available for future reuse. These data sets range from simple numeric data streams of simple structure but large size, to large collections of still and moving images.

The Earth Resources Observation Systems (EROS) Data Center collects and preserves satellite imagery and aerial photography, cartographic and topographic data created by or for the US government and under the custody of the US Geological Survey. Currently, it has approximately 12 million objects in several general collections, including each of the Landsat missions.

Several major efforts are currently underway to improve data centers through more consistent and interoperable procedures. The NASA Space Science Data Center (NSSDC) and the NASA Life Sciences Data Center (LSDC) are moving forward in the use of the OAIS Reference Model as the conceptual basis for their systems.

Data is also increasingly being stored as a result of submission as supplementary information with journal articles in digital form. PubMed Central, BioMed Central, the American Institute of Physics, Elsevier, The American Chemical Society, the Astrophysics Data System, the International Union of Pure and Applied Physics, the International Union of Crystallographers, to name a few, are routinely accepting the submission of supplementary data. However, it isn't clear how much data will be lost because the author does not submit it or because no formal publication was created as the end result of the research.

For this reason, CODATA (The Committee on Data for Science and Technology), an international organization for the interoperability and standardization of data in the sciences for the purposes of communication, has been raising awareness of the issue of data preservation. In June 2002, the South African CODATA Committee hosted a meeting in Pretoria, South Africa, geared toward the needs of developing countries and the African Continent in particular. CODATA and ERPANET jointly sponsored a workshop on "Selection, Appraisal, and Retention of Scientific Data" in December 2003 (ERPANET 2003). Another workshop

will be held in China in 2004. In addition, CODATA and ICSTI are creating a portal to provide information resources about archiving and preserving data with an emphasis on best practices and linking people to experts. In particular, this effort is aimed at supporting developing countries by providing a network of experts and highlighting practices that can be implemented in these countries (Anderson 2003).

Similarly, the National Science Board recently convened a meeting of experts for the National Science Foundation in the US. The goal of the meeting was to discuss the role that NSF should/could play as a funding agency in the preservation and access to data of long term value that is created as the result of grant funding.

6.4 Technical Reports

Technical reports and other gray literature are key mechanisms for the dissemination of research and development results, especially in industry and government. Many government and institutional archives are focused on technical reports, since libraries may not routinely collect them.

The ANSI/NISO Standard for Technical Reports (Z39.18) is currently undergoing its five-year review. As part of this activity, the review group is considering how the standard should change to reflect the digital nature of technical report creation and publication. In addition, the group is considering as an appendix to the standard a DTD for technical reports that has been developed by Old Dominion University (Maly and Zubair 2003). While this standard and DTD do not directly address preservation and long term access, the mark-up recommended in the DTD will support automatic metadata generation and additional semantic mark-up that would disaggregate the content from the presentation of the document. These are key factors in the development of a sustainable preservation system.

6.5 Conferences, Meetings and Lectures

Significant scientific information is first presented at conferences, meetings, lectures, colloquia, etc. Many disciplines, such as biotechnology, rely heavily on this method of communication rather than the formal publications. Therefore, the ability to preserve and access this type of information into the future is important.

As part of its knowledge management activity, the NASA Goddard Space Flight Center captures the content of colloquia, lectures and courses (Hodge, et al 2003). These events are routinely webcast and then saved digitally. Older videos are also collected and digitised. The encoded files are then indexed using a video indexing program, which allows users to query the videos by keyword and find precise retrieval intervals within the video stream. The software uses advanced voice recognition techniques and a dictionary that has been enhanced by adding the NASA Thesaurus to expand the queries and locate related intervals that do not specifically include the requested term. For example, a search on "planet" will also search on the names of the individual planets because the thesaurus has these terms as narrower terms to the term "planet". Recent work has included the linking of presentation slides to the appropriate parts of the video stream.

6.6 E-Records

The extent to which government-produced scientific and technical information is treated as an electronic record depends on the practice of the particular government or institution. Of course, e-records can include any or all of the above document types. However, there are significant e-records efforts underway within the governments that will have an impact on the overall digital preservation landscape.

The Victorian Electronic Records Strategy (VERS) of Australia is one of the most explicit suites of tools, standards and best practices with regard to e-records. The system has been operational since 1999 following a proof of concept/demonstrator project. (The latest version, released in July 2003, is mandatory for Victorian Government agencies.) The standard details functional requirements, the metadata set for long term preservation, and the long term format for records, which includes XML, PDF or TIFF, and digital signatures. Documents are converted into PDF and the context metadata is stored as XML. The converted records are encapsulated, i.e., bundled together into self-describing objects. VERS addresses many office-type documents, including e-mail. However, it does not specifically address databases or non-document type records such as sound and movie files, though they can be accommodated within VERS objects. VERS is currently working on a compliance program for vendors of records management systems. In addition, the Public Record Office of Victoria is in the process of obtaining a digital repository. The contract is expected to be awarded by the end of 2003, and the repository should be completed by the end of 2004. The repository is broadly based on the OAIS Reference Model and on VERS. It will be integrated into the existing records management, repository and access mechanisms for paper records (Quenault 2003).

While the Electronic Records Archive for the US National Archives and Records Administration (NARA) is not yet operational, there has been significant progress in that direction. The ERA began several years ago with a series of pilot projects, many of which involved the San Diego Supercomputer Center and its work on the Storage Resource Broker. These pilot projects were aimed at conducting the research necessary to create specifications for the architecture needed by an operational system to manage large-scale e-records systems, including the ability to deal with collections and different layers of metadata.

In addition, a draft Requirements Document (NARA 2003a) issued as part of the draft Request for Proposal issued in August 2003 describes the system. The final Concept of Operations also released in August 2003 describes the various user scenarios for such an Archive (NARA 2003b). "... the ERA system will ingest, preserve, and provide access to electronic records of all three Branches of the US Government. ERA is envisioned as a comprehensive, systematic and dynamic means for preserving any kind of electronic record, free from dependence on specific hardware and/or software."

Meanwhile, under its Electronic Records Management initiative, NARA is working to extend the types of electronic formats that it can accept. NARA has already extended its acceptable formats to include scanned images of textual records and PDF. Three additional formats are expected in the near future; these may include web records, digital photographs or geographic

information systems (Bellardo 2003). NARA is working with Adobe in developing the PDF-A format.

Guidance is being developed for future records so that NARA will be able to accept almost any format. They are working with partner agencies on archival metadata and relevant XML schema to provide more control through mark-up, including Dublin Core elements. Transfer may take place via FTP or Digital Linear Tape, which may become a long term preservation medium (Bellardo 2003).

A NARA Appraisal Guidance document issued in October 2003 includes an appendix on the appraisal of special types of information, including environmental, health and scientific observation data in the physical sciences (NARA 2003c). Recently, NARA has created a board of scientists and publishers to discuss the specific issues related to scientific e-records.

The UK Public Records Office (recently renamed the National Archive) has also been active in the area of digital preservation (Public Record Office 2003). The Digital Archive receives selected electronic records from government departments under the management of the Records Management Department. The Digital Archive is available to onsite users from designated PCs. Advice and guidance is provided to government agencies with regard to file formats, storage media, the care and handling of removable media, graphic file formats and image compression. Future topics on which guidance will be issued include digital signatures, encryption, and checksums. In April 2003, the National Archive's Digital Preservation Department hosted an international conference on "Practical Experiences in Digital Preservation," where issues of technology, organization, and cost were discussed by a variety of national archives, including those from the US, the Netherlands, Iceland, and the UK (National Archives 2003).

These organizations and others are part of the InterPARES effort which in 2002 began the follow-on project, InterPARES II. InterPARES II broadens the number and types of archives that are included in the group, and it addresses an extended scope of e-record problems. It currently includes over 100 researchers (Eastwood 2003). It will address issues of reliability and accuracy in addition to issues of authenticity, and it will address them throughout the records' lifecycle (from creation to permanent preservation). InterPARES I was concerned primarily with authenticity and with non-current records destined to permanent preservation (InterPARES 2003).

7.0 Standards by Format Type

The best format for long term preservation remains elusive, perhaps because there is no single answer to the question. Instead it depends on the format type of the original object, the characteristics of the original that the preserving organization considers to be most important to preserve, and the expected use/re-use of the object in the future (distance education versus legal evidence). Most experts agree that the best format for preservation is that which is least proprietary while conveying significant aspects of the original.

This section outlines the status of format standards for text, images, videos, data and other products of scientific research and communication with the realization that the practices represent a range of institutions with varying needs and decision criteria.

7.1 Text

The most common formats for storing text were XML (ASCII, with or without Unicode), PDF, and TIFF. Each of these formats has its place in the preservation strategy.

For scientific and technical text, as well as other objects, ASCII is the most open format, accommodating virtually all software or browsers now and into the future. However, for some digital objects, ASCII is problematic when paired with the requirement to provide permanent access and to render the look and feel of the original. Therefore, PubMed Central, DiVA and the Humboldt University cite XML as the preferred format for preservation because it is based on ASCII, non-proprietary and well-adapted for re-purposing and interoperability. The PubMed Central Guidelines require separate SGML or XML files for the full text of each article. DiVA creates XML for all available full text and Unicode is used to preserve the extended character sets from the original.

TIFF, an image format, is used to preserve the look and feel of original text objects. The use of TIFF in text environments began with the advent of scanning and Optical Character Recognition technologies, which used the TIFF images. TIFF can be employed at various resolutions depending on the quality and flexibility of the equipment used and the requirements for future use of the archived objects. Organizations such as the National Library of Medicine and its Profiles in Science Project continue to create TIFF as a major part of their preservation activities because of the high quality resolution provided.

However, TIFF is increasingly giving way to PDF, because PDF is more readily created from existing authoring tools, is often the preferred choice for submission by authors, has viewers that are more ubiquitous, and may be more easily and reliably indexed for full text searching. Until fairly recently, PDF was not considered a viable preservation format, because of its proprietary, though openly documented nature. However, PDF appears to have gained acceptance. For some organizations this may be a pragmatic move, since it is possible for the PDF versions of the documents to be easily created by the authors before ingest or by the archive upon acquisition. Also, the ubiquitous nature of Adobe tools and PDF files has perhaps assuaged some of the concern about the proprietary nature of Adobe products.

The national archives are particularly interested in PDF. VERS believes that it is essential to preserve the appearance of the electronic record as the original creator (and user) saw it. This explains VERS preference for PDF over XML (Quenault 2003). The fact that PDF is publicly specified and published means that it will be easier to re-implement a viewer for PDF in the future. (While non-proprietary standards are preferred, VERS accepts proprietary standards provided that they meet the open publication criteria.) Similarly, the US National Archives and Records Administration in the includes PDF among its e-records submission formats.

In other cases, PDF is viewed as a beneficial but supplementary version to be submitted along with XML. In the case of PubMed Central, PDF supplements the SGML/XML format by serving as an authoritative copy against which the SGML/XML can be validated before it is included in the PubMed Central archive. PDF also provides a guide for future rendering of the material by maintaining the look and feel of the original text object. KTH keeps the native format, generally Word or TeX/LaTeX, and then creates a PDF version. However, KTH does not consider PDF to be a preservation format since it is proprietary.

The effort on the part of the National Information Standards Organization, the National Archives and Records Administration, and Adobe to create an archival version of PDF, called PDF-A also highlights the importance of PDF. As part of the agreement, Adobe will identify a PDF core that will be retained throughout the versions of PDF now and in the future. The idea is that any document can be stored in its native PDF version and also in PDF-A through the PDF interface. Therefore, if an archive requires PDF-A on the part of contributors, the contributors can simply create that format and submit it or make it available for harvesting. Adobe representatives indicated that in specific instances, a document in a future version of PDF might use functionality that is not included in the PDF-A core. In these cases, the functionality will be dropped from the PDF-A version. However, several key areas of concern have been identified – one is the ease or difficulty with which PDF-A can be incorporated in applications that use or call PDF. Also, there is concern that there should be some acknowledgement or flag in the PDF-A version that content or functionality has been dropped from the original. PDF-A is now a final draft ISO standard out for vote by January 2004 (ISO 2003).

7.2 Images

There are a variety of image formats that archives may receive including JPEG and GIF. However, the majority of the institutions interviewed who are truly doing preservation convert these formats to TIFF Group IV or V. The rationale is to preserve the best image in a format that is the most standardized and not subject to loss or compression.

The National Library of Medicine's Profiles in Science Project is a research product of NLM's Lister Hill National Center for Biomedical Communications, which is being conducted in collaboration with the History of Medicine Division at NLM. This project creates collections of important papers, videos, audios, and even e-mails from noteworthy scientists in biomedicine, particularly Nobel Laureates (National Library of Medicine 2003). First, the original document is retained, whether electronic or paper. The staff creates the highest quality TIFF possible and any browser formats are created from the TIFF. By retaining the original, the door is open for creating better access formats in the future by reprocessing the original.

PubMed Central requires original digital image files for all figures, as well as tables and equations that are constructed as images and are not encoded in the SGML or XML. PubMed Central requests lossless compression TIFFs or EPS (Encapsulated Postscript); JPEG and GIF may be sent if they are the only formats available. PubMed Central is anxious to receive the best quality image available. Similar to the Profiles in Science Project, PubMed Central converts the TIFFs to JPEG and GIF for display via the web.

The EROS Data Center receives input from a variety of imaging and mapping sources, many of which require special processing by collection. The Landsat Archive Conversion System (LACS), which will preserve Landsats 4 and 5 satellite image data, ingests a variety of multispectral scanner and thematic mapper formats from DCRSi Cassette Tapes and High Density Tapes. The transformation performed by the EROS Data Center converts these input and tape formats to a high-density computer-compatible digital tape, generates and stores imagery appropriate for browsing, and creates metadata to be added to the Archive's inventory catalog.

The International Union of Crystallography has also constructed an image file format (imgCIF) to handle the very large dynamic intensity range of scientific images. imgCIF has a natural archival function for the preservation of original image datasets of this kind. The metadata describing the data set is fully compatible with other experimental descriptions and derived data stored in CIF (crystallographic information file) format described in 7.5 below.

The specific details of image preservation can be seen in the details of the photographic digitization project at the National Motor Museum in the UK. The digitization methodology used in this project will be employed for all similar projects at the museum. The following description describes some of the technical aspects that must be considered. It also highlights the balance to be struck between cost (particularly staff time) and creating a high quality, preservation-value image. As Kenney and Rieger have questioned, ". . . to what extent is it acceptable to produce considerably fewer images at a higher cost?" (Kenney and Rieger 2000).

> All calibration and settings are based on the ICC: *Adobe RGB 1998* color workspace, and the digitization workstation monitors have been calibrated to *Adobe RGB 1998*. Color images are saved with a profile of the *ICC: Adobe RGB 1998* profile. Grayscale images are saved with a profile of the *ICC: Gray Gamma 2.2* profile. The scanner grayscale tone levels have been set up with a range of between 10 and 245, a range which provides greater than 99.5 percent accuracy in grayscale representation of the original. This means that there is no need for levels' adjustment and, therefore, no loss of pixel data. Monochrome images, regardless of the degree of color toning or degradation, are digitized as grayscale rather than color. Digitizing in color may or may not result in a more accurate representation depending on the skill of the operator. While the aim is to record the photographic media as historical objects in their own right, this has to be balanced with the amount of time (and therefore cost) required for the digitization process. Early testing on the Project established that digitizing monochrome images as color took 3-4 times longer than processing them in grayscale.
>
> The images are scanned at 300 dpi with a 'Target' pixel size of 3,600 on the longest edge. This ensures maximum print quality at A4 size. The archive version of the image is saved once as a TIFF with no compression and written to two CD-Rs. The TIFF image is used to produce two .jpg surrogates (one at 128 and the other at 512 pixels) using the 'droplet' tool in PhotoShop. This ensures that the surrogates are kept to a standard size plus or minus 20 pixels. The 128-pixel surrogates meet the requirements of web thumbnails, and the 512s are designed for intranet use either directly from the server or via the database. (Molteno 2003)

The Cornell University Library recently developed an online tutorial on digital imaging (A. Kenney, et al 2000). While the tutorial focuses on digitization of images, much of the information, particularly regarding digital preservation, is applicable for born digital images as well. The tutorial emphasizes the need to consider longevity issues early on in the process of

the image life cycle, because many of the decisions made at that point will impact the ability to preserve the image over time. The tutorial also emphasizes the need to consider organizational strategies of both a technical and administrative nature, because a successful technical solution needs to be supported by the appropriate financial and administrative commitments in order to sustain and continue to build the resulting digital asset collection.

7.3 Numeric Data

Numeric data is similar to text but it generally has more structure. Whenever possible, the preferred form for data is an ASCII delimited file or an XML tagged file. However, many datasets, particularly those stored in local laboratories and by individual researchers or research groups, are stored in proprietary database formats. Accessing and reusing the data when it is stored in these formats over the long term, especially if there has been a disruption in the migration from one version of software to another, becomes problematic. However, organizations have also noted that good documentation, particularly retention of the data dictionary (as long as it is not in a proprietary database or CASE product) is necessary as well.

A key standard related to databases is ISO 11179: Specification and Standardization of Data Elements. This standard is used by a variety of groups and disciplines including those in the environment, aerospace and healthcare industries. Many organizations began using the standard to promote interoperability and reuse of legacy databases without having to migrate them. However, the result is a standard that provides a key component for preservation – the documentation of the database structure and the definitions of the database fields.

The US, Japan, Europe, and Australia are doing significant work in this area. For example, the Environmental Data Registry (EDR) has been developed by the US Environmental Protection Agency. It provides a mechanism for reusing data within the EPA and for exchanging data between the EPA and its state and local partners. Another key partnership is between the US EPA and the European Environment Agency. The EEA is developing an open source metadata registry to the 11179 standard, based on the EDR model. The plan is to use these mechanisms to share environmental information on an international scale.

7.4 Video and Audio

Video and audio is used in the sciences to record experiments, supplement human field observations, record engineering and laboratory tests, capture knowledge and lessons learned from researchers, and to teach science and engineering in distance learning environments. However, the standards for preserving video and audio formats remain an issue as the systems grow more complex and proprietary interests dominate.

However, the commercial importance and wide opportunities for re-use of video objects have spurred activities in this area. In particular, large creators of objects such as the Corporation for Public Broadcasting, Warner Brothers, and The Walt Disney Company have been instrumental in moving best practices forward. As digital TV becomes an increasing part of the entertainment and educational landscape, all production is being done digitally and old video objects are being converted to digital, preferably high-definition.

7.5 Output from Design, Modeling and Visualization Tools

There are several types of formats of importance to various scientific disciplines that are linked to specific software and sometimes hardware requirements. These have historically been heavy graphics and data oriented systems for modeling, drawing, and creating simulations. Examples include geographic information systems (GIS), chemical structure drawing, and computer aided design and manufacturing (CAD/CAM) in various engineering disciplines. Over the last several years, the importance of these tools has expanded, and there has been increased emphasis on the need for interoperability among systems. This has spawned various industry/vendor initiated "open" activities, such as the OpenGIS Consortium, standard formats for representing chemical structures, and open CAD/CAM environments.

Geographic Information Systems have become major tools for gathering and analyzing information in various areas of science from public health to geology. These GIS systems have been developed by a variety of vendors with the outcome being a series of proprietary systems. However, in order to bridge the confines of these systems, the OpenGIS initiative is developing standards for the interoperability of GIS information.

In the area of chemical structures, a number of systems have been developed to represent molecular structures for computer and human processing. Molfiles were developed by MDL, a vendor of technology for developing molecular structures. Molfiles are also used for transferring information from one chemical system to another. The Molfile is a flat file that is coded in a special way to indicate the elements that are contained in a molecule and the necessary information about bonds between them. An alternative to the Molfile system is SMILES (Simplified Molecular Input Line Entry System), which is a linguistic structure, a language with a simple vocabulary and grammar rules. An algorithm, available in the Daylight Toolkit, allows unique SMILES structures to be created so that a unique name of the molecule is synonymous with the unique structure (James, et al 2003). Because of its uniqueness and compact structure, SMILES has been used to exchange chemical information between systems.

The International Union of Crystallography requires that supplemental data dealing with crystal structures be submitted using the CIF (crystallographic information file) format. CIF was developed to enable interoperability between equipment manufacturers (image plates, diffractometer scans, etc.), databases such as the Cambridge Crystallographic Data Centre, the Protein Data Bank, and the International Center for Diffraction Data, publishers (IUCr journals and Zeitschirft fur Kristallographie) and software applications (PLATON, NRCVAX, etc.). A number of similar formats have been developed in related branches of science such as NMR structures in macromolecules. This is a standard system that has been developed for archiving information about crystal structures (McMahon 1996, Brown and McMahon 2002)

In the engineering community, the Standard for Exchange of Product model data (STEP) is a key standard (ISO 10303). STEP is an intermediate format (a kind of lingua franca) for the exchange and sharing of information used to define a product throughout the product life cycle and throughout the supply chain from design to delivery to the end customer. It is used to share CAD (Computer Aided Design) information, product models and technical drawings in industries such as aerospace and shipbuilding (Mason 2002). In the aerospace industry, there

are federal requirements to be able to recall and redisplay the product information, particularly the drawings, for aircraft for 75 years plus the lifetime of the aircraft or almost 100 years. With major input from companies such as Boeing and AirBus, the STEP community is now looking at the possibility of using STEP as a major component of the data to be preserved over time.

A related development is the OpenOffice Project, which deals with office documents across platforms and across software (OpenOffice 2003). The component-based language and platform-neutral architecture of the StarOffice utilities appear to be ideally suited to form the basis for an open office productivity suite. The Humboldt Institute in Berlin, Germany, is investigating OpenOffice as a possible solution to the handwork and checking required to reliably mark-up electronic theses and dissertations for preservation. This effort is also related to authoring templates that support mark-up and XML creation.

In these cases, the standards described have been developed primarily for interoperability. However, there are key factors required for interoperability, such as the non-proprietary nature, the open documentation of the standards, and ease and accuracy of conversion/transformation that make these open formats important to preservation.

8.0 The Workflow

Within each archive, the standards, best practices, and technologies are organized into a workflow. The key tasks in the workflow include selection, ingestion, metadata creation, transformation, storage and dissemination.

8.1 Selection Criteria

There are still two general approaches to acquiring material for an archive – harvesting and submission. Harvesting is performed by organizations like the US National Technical Information Service, the Internet Archive, and some of the national libraries like Sweden, Norway and Australia. Others, such as DSpace at MIT, those archiving theses and dissertations, and national libraries such as KB and the British Library that have voluntary submission from publishers, use the submission approach in which the activity is initiated by the producer of the information. The approach used depends, in part, on the ability of the archive to have a relationship with the producer.

In the case of harvested materials, there are two modes of identifying materials to archive. The automatic approach crawls or spiders sites based on criteria that have been supplied to the robot or agent. The Internet Archive and the National Library of Sweden use this approach. Criteria may include the domain name, specific root URLs (servers), format (html but not jpg, for example) or dates.

The second method is hand selection in which a staff member, with the aid of Internet search tools, identifies sites of interest. This approach provides much more precision and allows for review of potentially copyrighted or undesirable material. However, it requires resources.

"PANDORA remains a selective archive with all of its inherent advantages and disadvantages. One of the big disadvantages is the labor-intensive nature of the work, even with a much better digital archiving system. We have been unable to keep pace with the growth in online publishing, especially in the government sector." (Phillips 2003)

Regardless of the method employed, a critical step in the creation of a digital archive is formalizing the selection criteria. Early projects such as Cedars (Weinberger 2000) and the *Preservation Management of Digital Materials: A Handbook* (Beagrie and Jones 2001) identified the development of well thought out and clearly stated selection criteria as an important activity in order to ensure cohesiveness of the collection and a good understanding on the part of staff and users as to what will be in the archive. As the amount of digital information available has grown, organizations are beginning to review their selection criteria.

In February 2003, a review of the collection procedures for Australian online publications indicated that PANDORA had a choice to make between collecting a broader range of publications superficially or focusing on only certain types of materials to collect and archive in-depth. The report (NLA 2003a) recommended the second option and the previously published selection guidelines were modified based on this decision (NLA 2003c). Six categories of online publications are now given priority, all of which can include scientific and technical information:

1) Commonwealth government and Australian Capital Territory government publications
2) Publications of tertiary education institutions
3) Conference proceedings
4) E-journals
5) Items referred by indexing and abstracting agencies for allocation of a persistent identifier
6) Sites in nominated subject areas on a three-year rolling basis (these are outlined in Appendix 2 of the selection guidelines) and sites documenting key issues of current social or political interest, such as election sites, Canberra bushfires, the Bali bombing (Phillips 2003)

The MINERVA Project at the Library of Congress has also developed collection guidelines (Library of Congress 2003b). These guidelines allow for various scopes of capture, but, at present, the emphasis is on providing targeted collections that are at risk or that are of particular interest to the US Congress and the American people. At present, none of these collections are scientific in nature.

Even repositories that rely on submissions must have clear guidelines as to what should be submitted. The Content Guidelines for MIT's DSpace implementation include:

- The work must be produced, submitted or sponsored by MIT faculty
- The work must be education or research oriented
- The work must be in digital form
- The work should be complete and ready for distribution

- The author/owner should be willing and able to grant MIT the right to preserve and distribute the work via DSpace
- If the work is part of a series, other works in that series should also be contributed so that DSpace can offer as full a set as possible (DSpace Federation 2003c)

PubMed Central, the Astrophysics Data System and JSTOR have agreements with the publishers that allow them to be more specific about the workflow and the standards for materials received. Following negotiations and testing, the contributed materials are finally processed. PubMed Central and JSTOR have specific criteria for selection of scholarly journals to be included in their systems. These requirements include peer review, an editorial board, and a pattern of successful publication with some frequency pattern.

Many institutional repositories that do not have arrangements or ready access to their contributors are in a situation of having to locate relevant materials for the archive and then harvesting it. This is perhaps the most difficult selection process because the domains of the enterprise, for example, the entire government, are so large. This is the current situation for the US Government Printing Office and for the National Technical Information Service. The US Government Printing Office has collected over 6000 government publications from agency web sites. The publication must be the official version of the document; it is not collected in electronic form if the GPO has already received a copy in paper or microfiche. The original is captured so that the link can be made to the archived version if and when the original is removed from the agency server. The GPO is also negotiating agreements with agencies, particularly those that do not have systems to maintain their information for the long term, to ensure receipt of their documents when the originating agency is no longer able or willing to maintain them.

8.2 Metadata Creation

In her Metadata Generation Framework, Greenberg (2003) identifies two main metadata generation processes: human metadata generation and automatic metadata generation. Supporting this bipolar framework are three types of metadata tools – metadata generators, metadata templates, and metadata editors. Metadata generators create metadata from the objects (generally text). In Greenberg's framework, generators include automatic indexing tools that are run to create indexes prior to the user search (ala, Google and Yahoo) or to dynamically create metadata such as the locator and brief title at the time of search. Metadata templates are forms or markup rules that support the human creation of metadata. Metadata editors combine the two by generating preliminary metadata, often in a template, and then presenting the human with an interface to review and edit the metadata. The more extensive the metadata (particularly for description), the more difficult it is to create metadata in a completely automated fashion.

All three approaches to metadata creation are found in the highlighted systems. The Internet Archive spiders the Web contents and the Wayback Machine pre-indexes the content prior to the search performed by the user. The DiVA Archive uses data originally entered by the document author as the basis for creation, reuse, and enhancement of all metadata. The metadata are created when the object is prepared for publishing. The metadata for DiVA is an

internal document format from which other metadata formats, including the Open Archives Initiative-Protocol for Metadata Harvesting (OAI-PMH), can be created (DiVA 2003c). At the KTH, the full text dissertation is transferred to the Library's acquisitions department via e-mail or FTP, along with the metadata that is created by the author via a form on the KTH website. The library staff checks the results. Similarly, PubMed Central has a DTD and content guidelines that publishers use to submit materials. The system at the National Library of Australia, PANDAS, and the system developed by the NASA Goddard Space Flight Center Library derive metadata from HTML metatags and by analysis of the Web site. However, the cataloger performs the final review and enhancement. In most cases, administrative and some preservation metadata elements (such as dates and file format) are created automatically by the system.

Metadata creation is particularly important for data centers where the objects to be preserved have limited searchable descriptive text. Therefore, data centers have large and extensive metadata catalogs that aid in management, preservation and reuse of the datasets. Examples of these include the National Space Science Data Center, the Earth Remote Observing System, and the Global Change Master Directory. All items are identified in a metadata catalog that utilizes the Federal Geographic Data Committee (FGDC) content standard and its follow-on ISO TC-211 for geospatial metadata. In general, these metadata catalogs have records for both the individual datasets and for collections of datasets; for example, those based on collection from a specific instrument or under a specific program. Because of the large number and size of datasets ingested by data centers, significant effort is spent automating the metadata process so that elements are populated at ingest.

Metadata is also an area of concern for e-record archives. Because of the massive amount of information to be ingested into an e-records archive and the general emphasis on collections rather than individual items, the focus on metadata among archives is slightly different. For example, the Victorian Electronic Records Strategy (VERS) expects that the digital objects will come from existing digital asset management systems that contain metadata. The metadata in these systems may range from a rich set of metadata from a formal records management or document management system to "scraps of metadata" from a file system. Identifying the metadata elements from the source systems and mapping them to the standard fields supported by the digital archive is a key part of archive management, planning with cooperating organizations, and the ingest process.

The VERS metadata is expected to serve as a "lingua franca" to which native metadata can be mapped (Quenault 2003). VERS has identified 141 possible elements to preserve and maintain the access to e-records over time. However, of these elements, only 34 are mandatory, 11 of which are automatically derived from the system, 8 are defaulted, and 2 automatically derive based on values in other elements. Of the remaining 13 elements, the contents of 11 elements are selected from pick lists and only two elements require the creator to type in the content (Sinclair 2003).

8.3 Archiving and Transformation

There are several major approaches to archiving and transformation. These choices include transforming incoming materials into a standard format conducive to archiving, retaining the original format and migrating to new formats as necessary, and "migration on request". Granger and others have concluded that a variety of preservation strategies and technologies should be available. Some simple objects may benefit from migration, while others that are more complex may require emulation (Granger 2000, Holdsworth and Wheatley 2001). The appropriate approach will also depend on the nature of the archiving organization and the needs of the targeted user community.

8.3.1 Transformation to a Preservation Format

Some organizations interviewed ingest materials in native formats. Others provide specified submission formats. Still others transform native formats into other formats that are deemed more preservation-friendly.

For text, several organizations are transforming incoming material into ASCII and XML. DiVA creates a manifestation in XML for all full text that it receives. Humboldt University in Germany creates SGML and increasingly XML for the materials that it ingests. PubMed Central stores content in XML.

PubMed Central has an extensive data flow. The journal files are received in SGML or XML with images, PDFs, and supplementary data files. The SGML is then converted to the PMC XML common format. This is primarily aimed at standardizing the tags that are received, rather than any standardization of the content of those tags. For example, one publisher may tag an author as "au" and another may use "auth". The tag is standardized to the PMC DTD tag of "author-name". However, if one publisher has the author with full names and another uses initials only, there will be no change in order to make these consistent.

Images are converted to web display formats (jpg and gif). The PMC Archive includes the source SGML/XML, the high-resolution image files, the supplementary data files, and the PDFs, if provided. The PMC Public Access Database consists of the supplementary data files and the PDFs from the Archive along with the PMC common XML files and the Web display of the images that were created above. The online page displays are created dynamically from the PMC database as the results of the queries. The table of contents pages are also created on the fly.

A lesson learned from PubMed Central is that it is not practical to work in an active archive with multiple DTDs. PMC received almost a dozen different DTDs from among its publishers. Converting to a common DTD provides tremendous efficiencies. The complexities of the DTDs are handled only once in the beginning and it is possible to write these conversions to match each input file. The uniform data format that is created can then be handled by a standard set of routines as all the other functions of the Archive, including preservation, are addressed. (Sequeira 2003)

However, not all archives view ASCII and XML as the ultimate formats. The Profiles in Science Project at the National Library of Medicine scans and/or converts electronic content to TIFF and PDF. KTH retains the native format, which is usually Word or TeX/LaTeX, but it also creates a PDF version. Other formats will be stored in native format and KTH will try to transform them as necessary.

The central preservation tenet of the Victorian Electronic Records Strategy (VERS) is that content be transformed to a long term preservation format. The standard long term preservation formats used by VERS are PDF, TIFF, and ASCII text. (Although it should be noted that VERS recommends preserving the original bit stream as well.). Unlike PubMed Central, the ideal process for VERS is that the submitting organization will perform the conversion to the preservation format. "This is for three reasons. First, the archive may not have the necessary software to read the original digital objects. Second, even if the archive has software, it may be a different product (or a different version), which will often result in the appearance of the preserved digital object changing in the migration process. Finally, the archive cannot judge the accuracy of the migration; only the creator can do this." (Quenault 2003) In addition, by taking the format, VERS is committing to support it for the long term. This may mean buying new licenses or writing new migration or viewing software. "Each format is consequently a considerable economic cost to an archive, and [VERS] takes the position that the number of formats should be strictly limited." (Quenault 2003)

8.3.2 Migration

Migration is the preservation strategy used by most data centers. EROS uses migration as its preservation strategy. EROS has traditionally tried to archive the lowest level, i.e., the least processed format level allowing the most flexibility in providing products. Some of the data sets are now served directly through the Web with browser and Java interfaces. Digital files can utilize general software that views TIFF files for some of the EROS holdings. Other digital files may require specific, third party commercial software that is not provided by EROS. Currently, EROS is migrating its largest and oldest digital satellite imagery.

DSpace at MIT also uses migration as its main strategy. MIT maintains the original and will support through migration formats classified as "supported", that is, non-proprietary. It also expects but will not guarantee to migrate files, such as MS Word, Excel, etc., that are proprietary but for which conversion tools are likely to be available because of the installed customer base for the native format.

8.3.3 Migration On-Request

In "Migration on-request," the original version of the material is retained and when necessary, conversion tools are applied to convert the original to the format required by the user (Mellor [n.d.]). This saves time and resources and accommodates the fact that users do not upgrade from one version of software or hardware to the next at the same rate. No instances of the use of this technique were found in the survey, but this concept was tested as part of the CAMiLEON Project (CAMiLEON 2002).

8.4 Storage

Storage is a key part of the infrastructure activity. In most cases, the metadata and the object are stored separately. Many projects are still at the stage where the metadata and the object are stored in Access or other databases that are proprietary in nature. This facilitates access but it is not considered to be the most appropriate for long term preservation. The DiVA Project, for example, is planning to move from a relational database management system to an XML database.

Systems such as VERS and the Electronic Records Archive of the National Archives and Records Administration, encapsulate the metadata about the digital object with the object itself, as one object. In this way, even if the indexes or the complete management system are lost, the digital objects themselves are self-describing, and the metadata can be extracted from the original objects to recreate the system.

The other aspect of storage is a technical one. It has become apparent that the ability for a storage device to hold the vast amounts of data required by some archives – for example the capturing of web sites or the large image or video files – is important. However, of equal or greater importance is the ability of the file structure to provide adequate indexing, to be easily backed up and recovered, and to provide for hierarchical and linked relationships between files (particularly between the metadata and the object). EROS stores off-line on magnetic media, near-line in storage silos, and online in disk arrays. Magnetic media migration is expected to occur about every five years.

8.5 Dissemination

The dissemination mechanism of choice across all the archives surveyed is the web. It is expected that this will continue and that demand for web dissemination of not only the metadata but the actual objects will continue to increase. This poses two potential problems. The first is the bandwidth and browser issues as the objects become larger and more complex. The data community sees a continued move toward web dissemination, but FTP, CD, and digital tape dissemination continue (Faundeen 2003a). The second problem is that the web is itself a publication mechanism and, therefore, archives are concerned not only about preserving the digital objects, but displaying them in the future (see section on Rendering).

Dissemination can also be viewed in terms of the degree to which the archive is accessible (often referred to in degrees as "dark" or "deep", "dim" or "lit"). The dissemination practices depend on the type of archive, the target and possible audiences or designated community, and the business model(s) being invoked. Current implementations at national archives and government agencies generally do not have completely open access as there may be some restrictions on the distribution of the material. Institutional repositories may or may not provide public access to the material depending on copyright/intellectual property concerns and the agreements with their contributors. In the case of national libraries, the legal deposit regimes or agreements with publishers or authors may require that access be limited to the specific library that has the deposit agreement. In some cases, an object can move from the "dark" part of the archive to the "lit" part. JSTOR calls this "the moving wall" (JSTOR 2003).

PubMed Central will also make arrangements with publishers to keep the most recent issues "dark" for a certain time period.

9.0 The Introduction of "Off-The-Shelf" Systems

Since early in the investigation of digital preservation, institutions concerned about preservation and interested in performing this function have been awaiting "off the shelf" systems or services that could be installed with limited resources but variant levels of flexibility to meet local needs. These systems are beginning to become available from a variety of organizations. Several of the highlighted systems have or are developing "turn-key" or generalized systems that can be implemented by others. These are available both commercially and as open source software.

9.1 DSpace Institutional Digital Repository System

The DSpace Institutional Digital Repository System began as a joint project of the MIT Libraries and Hewlett-Packard Company. The architecture for the system is based on a number of preceding projects including those at Cornell, CERN, OCLC, LC and OAIS. DSpace 1.1 was released in November 2003 via an open source license (available from SourceForge).

While the architecture for this system is very interesting, the most significant aspect is that it takes the institutional repository/digital library concept and incorporates the concept of preservation services. In DSpace, each bitstream is associated with one Bitstream Format, which is a consistent and unique way to refer to a particular file format. For example,, if the bitstream is encoded in JPEG, the interpretations are based on the explicit definition in the Standard ISO/IEC 10918-1. These can be more explicit than the MIME types or file suffixes.

In DSpace, each Bitstream Format has a support level, which indicates how well the hosting institution is likely to be able to preserve content in that format into the future. There are three possible levels – supported, known, and unsupported (Tansley, et al 2003). The general DSpace support levels are defined at a very high level. Therefore, each adopting institution must identify specifically what these mean for their particular environment. For example, the MIT Libraries implementation of DSpace defines the support levels as follows. "Supported" means that the format is recognized and the institution is confident that it can make the format useable in the future through whatever technique is desirable (emulation, migrations, etc.). Note that there is no attempt to dictate the preservation method. "Known" means that the format is recognized and the institution will preserve the bitstream as-is, without a complete guarantee that it will be able to render the object completely in the long term future. "Unsupported" means that the format is unrecognized by the archive, but the institution will undertake to preserve the bitstream as-is and will attempt to retrieve it.

The concept of support levels is further enhanced by DSpace's use of metadata. DSpace includes three basic kinds of metadata – descriptive metadata, administrative metadata, and structural metadata. This follows the basic Metadata Encoding and Transmission Standard

(METS) framework. The descriptive metadata in the open source version is based on the Library Application Profile for elements and qualifiers. However, an institution installing DSpace can change the element set that is used. The community and collections levels, which reside above the individual objects, have simple descriptive metadata that is a subset of the Dublin Core.

Structural metadata includes information about how to present an item or bitstream. It also includes information about how this item relates to other items, particularly those that are constituent parts of a larger item. It provides information about the "pages" and their order. Further work is anticipated on how to make DSpace understand more complex object structures.

Administrative metadata includes preservation metadata, provenance, and authorization policy information. Most of this is held within the DBMS relationship schema. Provenance information is held in Dublin Core records but in a "prose" description. This area of administrative metadata is likely to be enhanced as the OCLC/RLG PREMIS group further identifies elements for preservation metadata.

In addition to these components of DSpace that are specifically preservation oriented, the DSpace suite includes search and browse capabilities and support for the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH). This allows DSpace sites to harvest metadata from several sources and to offer services based on the metadata that is harvested.

The vision of DSpace is a federation of repositories. Following the announcement of such a federation, there were over 2,500 downloads of the open source software, as reported in the DSpace newsletter in Spring 2003. The initial federation formation includes seven partners from other research institutions in Canada, the US and the UK. The federation will explore organizational issues and advise on how DSpace should proceed.

In addition, MIT, Hewlett Packard (HP) and the World Wide Web Corsortium (W3C) are collaborating on the SIMILE Project. SIMILE is exploring the use of RDF (Resource Description Framework) and Semantic Web techniques to deal with the interoperability of heterogeneous metadata schemas and how users of DSpace repositories can find and organize information of personal relevance from among the repositories.

9.2 Digital Information Archive System

The Digital Information Archive System (DIAS) is a commercially available system, originally developed to handle the electronic deposit of electronic documents and multimedia files for the Koninklijke Bibliotheek (KB), the National Library of the Netherlands (IBM 2003a). It is based on the results of the various NEDLIB Projects led by the KB over the last several years.

The DIAS design is based on requirements for an e-deposit system which must be met through archiving and/or transformation (Steenbakkers 2001). Even the most complex electronic publications break down to three components – the bit stream, the logical format in the bit stream, and the functionality needed to decode this logical format. Each one of these

components has its own criteria/requirements for proper preservation. The bit stream will be successfully preserved if the bits are copied and the storage medium is refreshed. "If the copying is done without loss and the refreshing is timely, the authentic structure of the bit stream can be saved indefinitely." The logical format for the bit stream will become obsolete over time. The approach generally used to solve this is format migration, i.e., upgrading from one version of Word to another or converting from WordPerfect to Word. This is unlikely to always be able to be performed without loss during the conversion process. The third problem arises because an interpreter is needed to transform the bit stream into a rendered format that is visible to the human eye. This requires special software, which itself is a bit stream requiring savings. This third step is paramount to providing long term access and is considered unique to digital preservation.

Three key guidelines were identified (Steenbakkers 2001):

1) Free the publication from its original carrier or environment, which is meant for publishing and not for archiving, and store the publication in a controlled archiving environment.

2) The controlled archive environment should be constructed in compliance with the OAIS Reference Model.

3) The archiving environment – or the deposit system – should be separate from the institution's communication technology environment. It should be focused on archiving and not on searching, authentication, etc. This will make the deposit system more durable and able to be upgraded as new mass storage techniques are developed.

DIAS supports both manual and automated (batch) ingest of material. Material can be accessed either by a Web interface, if the file has a standard file type, or by a specific work environment on the Reference Workstation. The system is based on IBM's Content Manager. It requires RS/6000 and IBM/AIX or Sun/Solaris server hardware and software, PC/Windows systems, and Web clients for access. Special PC/Windows systems are required for the Reference Workstation.

The DIAS system was implemented as KB's Deposit of Netherlands Electronic Publications (DNEP) system in December 2002, making it the first system of its kind (IBM 2003b). KB's initial implementation is for e-journal publishers to deposit e-journals, but the plan is to extend this to other types of e-materials such as e-books. DNEP serves 50 employees who access the system to supply metadata and up to 100 concurrent users.

In May 2003, the KB announced that it had signed an agreement with Kluwer to archive electronic journals featured on Kluwer Online Web Site. As of May, this contained 235,000 articles from 670 journals. The collection from Kluwer is expected to grow by more than 70,000 additional articles. The KB is seeking to enter into similar agreements with other publishers. Currently, the users (members of the public) must access the system from within the library because of copyright issues.

In the current DIAS system, IBM addressed the initial ingest, transformation, storage and metadata creation. The technical issues related to long term access are being studied by IBM and are not a part of the December 2002 implementation.

In 2003, the KB started a joint project with IBM to develop the preservation subsystem of DIAS. The work began with a series of studies around key preservation issues such as authenticity, media migration management, archiving of web publication, and a proof of concept of the Universal Virtual Computer.

This subsystem will consist of a preservation manager, a preservation processor, and tool(s) for permanent access. The Preservation Manager will manage and control the long term durability of the digital objects using technical metadata. This is considered to be an essential part of the DIAS solution, since technical metadata will allow a future hardware environment to take the software bit stream and the content bit stream and provide access to the content. The problem that remains to be addressed is the obsolescence of the hardware of the rendering environment. Two major approaches are emulation and the use of a basic virtual computer. The aim is to have the turnkey system able to be generalized to other libraries and archives. Therefore, the system must be independent of either of these preservation strategies.

9.3 OCLC Digital Archive

As an outgrowth of the preservation services that OCLC has provided to its member libraries for many years, OCLC has developed the OCLC Digital Archive. It provides long term access, storage and preservation for digital materials, or "objects." The system is based on the OAIS. Records can also be ingested in batch. METS and the NISO draft standard for Technical Metadata for Still Images will be used as the structure for SIPs (Submission Information Packages) that are submitted in batch. The internal record structure is created from the METS ingest format. A METS structure can be created for output to allow for interoperability with other archives.

Currently the OCLC Digital Archive can ingest text and still images in formats such as PDF, HTML, TEXT, JPEG, BMP, GIF and TIFF. The goal is to accept more input formats in the future. This system is connected to OCLC's Connexion cataloging system, and the cataloger begins by creating a WorldCat record for the object, followed by a record that includes the preservation metadata. The preservation metadata is based on the early RLG/OCLC work in this area. These two records are linked (OCLC 2003). In principle, it follows the Metadata Encoding and Transmission Standard structure, providing for descriptive, administrative, technical and structural metadata.

The system also includes an Administration Module that allows the user to modify existing records. The Administrator can set privileges for a variety of functions so that various pieces of the metadata creation, ingest and dissemination processes can be assigned to different people with proper security. The Administration Module also allows the administrator to create collections and user groups for specific end-user access to the metadata and the content. Virus and fixity checks are run and results are reported through both the Administration and the cataloging (Connexion) modules.

9.4 PANDORA Digital Archiving System (PANDAS)

The PANDAS (PANDORA Digital Archiving System) has been operational since June 2001. (National Library of Australia 2003b) The second version was installed in August 2002. Prior to the development of its own system, PANDORA tried to buy an archiving management system. From the response to the Request for Information, it became apparent that there was no affordable system on the market that met the requirements and so NLA decided to build the system in-house.

PANDAS enabled PANDORA to increase the efficiency of capturing and maintaining the archived Australian online publications and, therefore, PANDORA's productivity. It also provides PANDORA's partners, primarily the state libraries, with more effective Web-based software for contributing to PANDORA.

The PANDAS system covers a wide range of functions needed to manage a digital archive:

- Create and maintain records for all titles considered for selection in PANDORA and record decisions made about them
- Create and maintain records for collections of titles
- Create and maintain records for publishers and indexing agencies
- Search for records using a number of options
- Transfer titles from one agency to another for consideration or processing
- Initiate archiving of individual titles
- Manage the ongoing archiving of titles with scheduled gatherings
- Manage the processing of archived titles including supplementary gathering and editing of archived files
- Log archiving problem reports with automatic notification to IT staff
- Have access to information regarding gathering of titles
- View gathering queues of titles in the process of being archived and those that have completed archiving
- Automatically create title entry pages without the need to mark-up
- Automatically update title and subject listings in PANDORA
- Log on with varying levels of access and privileges, as determined by PANDORA administrators
- Receive a range of regular reports
- Create and maintain records for users of the system
- Create and maintain records for agencies using PANDAS

The "Gather Queue" function provides the staff member with information about the titles that are in the process of being gathered, waiting to be gathered, and those completed. The user can also pause, stop or delete the results of the gathering process. Once the instance has been captured, the user reviews all the pages. The system supports the correction of various types of errors including missing pages. The user can access the working area of PANDORA via WebDav ("Web-based Distributed Authoring and Versioning, a set of HTTP extensions that

allows users to collaboratively edit and manage files on remote web servers.) In this case, the user locates the missing file, captures it and then inserts it into the set of files for that gathering.

The NLA has received a number of requests for access to the PANDAS software, since the current software options to support the creation and management of digital archives are limited. UKOLN recommended use of PANDAS for pilot web archiving projects it proposed for both Wellcome Trust and JISC (Day 2003). In response, PANDORA will soon make available an evaluation module, which will allow interested parties to have trial access to PANDAS.

9.5 Lots of Copies Keep Stuff Safe (LOCKSS)

LOCKSS (Lots of Copies Keep Stuff Safe) is an automated, decentralized preservation system designed to protect libraries against loss of access to digital materials developed by Stanford University (LOCKSS 2003). LOCKSS software, which is free and open-source, is designed to run as an "Internet appliance" or easy to use software on inexpensive hardware and to require minimal technical administration. The present beta project is testing LOCKSS security, usability, and software performance. LOCKSS has been operational for five years and will move from beta mode to production mode in March 2004.

In 2004, LOCKSS is moving toward becoming a self-sustaining alliance. "The LOCKSS Alliance will provide a small core of central support for technology, collections, and community services. In addition to a range of specific services, the Alliance will transfer knowledge, skills and responsibility for the LOCKSS Program from Stanford University" (Reich 2003).

LOCKSS development is supported by the National Science Foundation, Sun Microsystems, and the Mellon Foundation. With Mellon Foundation funding, LOCKSS is building production quality software to archive and preserve e-journals. LOCKSS has NSF funding to explore computer science issues arising in the application of peer-to-peer technology to digital preservation.

LOCKSS creates low-cost, persistent digital "caches" of authoritative versions of http-delivered e-journal content at institutions that subscribe to that content and actively choose to preserve it. LOCKSS uses the caching technology of the web to collect pages of journals as they are published, allowing libraries to take physical custody of selected electronic titles they purchase. Unlike normal caches, however, pages in these caches are never flushed. The LOCKSS server runs an enhanced web cache that collects new issues of the e-journal and continually but slowly compares its contents with other caches. Accuracy and completeness of LOCKSS caches is assured through a peer-to-peer polling system, which is both robust and secure. LOCKSS replicas cooperate to detect and repair preservation failures. If damage or corruption is detected, it can be repaired from the publisher or from other caches.

By enabling institutions to locally collect, store, archive and preserve authorized content, they are able to safeguard their community's access to that content. The LOCKSS model enforces the publisher's access control systems and, for many publishers, does no harm to their business

models. The creation of these caches, given the requirement that the caching library already have the right through subscription to obtain that content, has met with a high degree of publisher and library engagement and commitment. About 90 libraries and 50 publishers are participating in the program.

9.6 Fedora™ (Flexible Extensible Digital Object Repository Architecture)

The University of Virginia (UVa) Library has teamed with Cornell University's Digital Library Group to develop Fedora, an open-source digital repository architecture on which a variety of digital library implementations can be based (University of Virginia Library 2003). Fedora is based on the original work by Lagoze and Payette at Cornell for use in digital library environments. Similar to DSpace, Fedora is focused currently on repository development and management. However, it will eventually include preservation services.

Major features of Fedora (1.2) include three open APIs that are exposed as web services, including one for repository management; support for OAI-PMH; a flexible digital object model that allows digital objects to act as containers for datastreams of content and metadata and for disseminators (linkages to services that transform the content or perform computations); support for any MIME type; default disseminators that set behaviors, for example, for viewing the contents of a digital object; extensible disseminators that can be custom built; content versioning if the datastream (either the content or the metadata) is modified; XML ingest and export in files that conform to METS; data object storage options, including a database that enhances the performance of the system; access control and authentication based on IP addresses or IP ranges (upcoming releases will include the Shibboleth-based authentication and access policies); searching of the primary Dublin Core records for the objects as well as selected specific metadata fields; an administrator client for managing the repository; a migration utility to perform mass export and ingest of objects; and a batch utility that enables mass creation and ingest of objects.

The largest implementation to date is the University of Virginia Library's Central Digital Repository. (Fedora is not the entire Central Digital Repository, but it provides "the plumbing" (Johnston 2003)). Since 1999, a series of tests and prototypes of increasing size and complexity have been created at the Library using Fedora. With each round, improvements have been made to the software and architecture. A major boost was provided by a Mellon Foundation grant in 2001 that allowed for joint development of a production-quality system by Cornell and the University of Virginia. At that time, a larger number and variety of resources were added to the UVa system, including the *Journals of Lewis and Clark* and a large image repository. The system currently includes XML objects, text (full text and page images of e-books), and images in multiple resolutions (Payette 2003).

Fedora 1.0 was released as open source software (Mozilla Public License) May 2003. Release 1.2 was made available in December 2003 (Johnston 2003). The first phase production repository based on Fedora will be launched in 2004. However, all the functionality described in the original design proposal will not be completed until 2005.

A number of other institutions and organizations are using Fedora and others are evaluating its application (Payette 2003). Fedora is a component of the DSpace architecture. VTLS is using it as the basis for a new commercial (library system) product, and a number of US university projects are using Fedora, including Indiana, Northwestern, Rutgers, Tufts, Yale, and New York University. Other sites using/evaluating Fedora include the National Science Digital Library at Cornell, JSTOR, The British Library, the National Library of Portugal, the Thailand Office of Defense Resources, and Cornell Information Technologies, among others. Since May 2003, Fedora has had 1427 downloads from 32 countries. These represent universities, software and technology companies, defense/military, banks, national libraries and archives, publishers, research laboratories, library automation vendors, and scholarly societies.

10.0 Standards Activities

Despite the advent of operational systems and "off-the-shelf" solutions, there is an increase in standards activities related to digital preservation and permanent access. This shows a certain maturation of the field (Hodge 2002). These activities are extremely important because they increase the level of cooperation and will ultimately result in increased interoperability among archiving organizations. This will allow the burden of the large volume of digital information and the resources needed to preserve it to be shared.

Many of the standards and best practices have been put into framework documents that provide guidance without being prescriptive. This is especially important when dealing with a wide range of organizations with diverse skills and resources. *Preservation Management of Digital Materials: A Handbook* (Beagrie and Jones 2001) summarizes the findings from major projects such as NEDLIB and Cedars. "A Framework of Guidance for Building Good Digital Collections" from the Institute of Museum and Library Services is a key framework document that bridges libraries, archives, and museums (IMLS 2001).

The following sections address standards and best practice activities at this time in four major areas of metadata – descriptive, preservation metadata, technical, and structural. – realizing that there is overlap in where specific elements would be placed in these categories. (A more complete discussion of various categories can be found in Gilleland-Swetland 2000.)

10.1 Metadata

10.1.1 Descriptive Metadata

Descriptive metadata provides the basic "bibliographic" information about a digital object. Elements generally include the title, an annotation or description, the creator, etc. These elements form the basis for what a user might search to find relevant objects.

The majority of the archives use Dublin Core as the basis for their descriptive metadata. In some cases, the Dublin Core elements have been extended (or qualified) to provide more precision for the specific needs of the user community. The Victorian Electronic Records Strategy has extensive descriptive metadata (National Archives of Australia 1999) in its set of

over 100 possible elements, which is also based on the Dublin Core. The NASA Goddard Space Flight Center Library has over 50 elements in its draft Goddard Core Metadata set. Qualified Dublin Core is used as the basis for providing more detailed elements to describe project documentation of importance to Goddard's researchers and engineers (Allen 2003). Additional elements include the project name and the instrument name.

There are several systems, such as the US Government Printing Office and the Library of Congress, that use MARC (or MARC lite) metadata formats. These instances are generally based on the need to interface with library or legacy bibliographic systems.

10.1.2 Preservation Metadata

Based on the previous work of RLG and OCLC (Planning Committee of the OCLC/RLG Working Group on Preservation Metadata 2001; OCLC/RLG Working Group on Preservation Metadata 2002), OCLC has formed a follow-on group to further develop core metadata for preservation. Previous work by RLG discussed the need for preservation metadata and the difference between this type of metadata and metadata for other purposes, such as resource discovery. A follow on to the initial white paper analyzed the various preservation metadata elements identified in the course of major projects such as NEDLIB, Cedars, and the Harvard Project and attempted to reconcile them.

PREMIS (PREservation Metadata: Implementation Strategies) will address "the practical aspects of implementing preservation metadata in digital preservation systems." In May 2003, OCLC created a Working Group and an Advisory Group. Over the next year, the Working Group will develop "... a broadly applicable and implementable set of "core" preservation metadata elements and a data dictionary to support them. It also will evaluate strategies for managing preservation metadata within a digital preservation system, and for the exchange of preservation metadata between systems; establish pilot programs for testing the group's recommendations and best practices in a variety of systems settings; and explore opportunities for the cooperative creation and sharing of preservation metadata" (OCLC Research 2003). The Working Group has divided into two subgroups. The group focusing on the element set has developed a draft set which is currently being reviewed. The other group is addressing implementation issues by surveying major repositories. The Advisory Group will provide initial review and comment. PREMIS is scheduled to complete its work by June 2004.

In a similar, though narrower activity, the Defense Technical Information Center published its guidelines for preservation metadata in March 2002 (DTIC 2002). Developed in support of its prototype Defense Virtual Library, jointly developed with DARPA and CNRI, the preservation metadata spans multiple object types, including images, videos, and technical reports. The element set includes over 100 elements that DTIC believes are the primary elements needed to begin long term preservation of digital library objects. The documentation maps the elements to the OAIS RM and describes the elements using the ISO 11179 standard for data element registries. The rules for content creation are based on AACR2.

The Victorian Electronic Records Strategy has extensive preservation metadata identified for record keeping by the National Archives of Australia (1999). An unusual characteristic is that VERS includes metadata that describes VERS itself. If a future user had a VERS Encapsulated

Object with no documentation, the short textual descriptions and the techniques for constructing the encapsulated object can be extracted from the XML in order to recreate software to process the encapsulated objects and even replace the repository. In addition, a textual reference to the published standards that document the preservation format is included in each record.

Preservation metadata must also be geared to the specific format. There is significant activity related to metadata for moving images, audio, and still images. MPEG-7 is perhaps the most widely discussed standard for such metadata.

The ViDe Videoaccess Working Group was formed in 2000. It is a group of digital video and network professionals who have been involved in standards for video for many years. The group recently mapped and compared MPEG-7 and Dublin Core. MPEG-7 is very rich but too detailed for many applications. Dublin Core, on the other hand, lacks the richness often required to adequately describe these types of objects. The group is working to make the MPEG-7 standard more understandable in the context of libraries and archives, and to further identify how the Dublin Core might be used in this context. Both standards are likely to be used to describe these objects (Kniesner 2003).

The Corporation for Public Broadcasting (CPB) is developing metadata for broadcast assets (including video and audio) (White 2003). The PB Core is a result of the Public Broadcasting Metadata Initiative (PBMI), an effort to develop a metadata element set to describe the CPB assets for the purposes of sharing metadata and enhancing discovery of the assets. Following an analysis of alternative schema, the PBMI developed an application profile based on the Dublin Core, which combines elements from different standards, while applying constraints to some of the elements for specific controlled vocabularies or structured values. It includes 58 elements. Significant extensions (or qualifications) have been made to the Dublin Core elements for Title, Rights, Description and Format. The PB Core will undergo an evaluation by the members of the working group, followed by a Request for Comment from a larger group of public broadcasters, operations staff, vendors, standards organizations, and partnering institutions. Test implementations will be developed to include all aspects of public broadcasting, including radio, television, and the web.

10.1.3 Technical Metadata

Another key area of development is technical metadata. This may also be considered part of Preservation Metadata. Technical metadata documents the technology environment in which the original was produced. It may eventually be used to render, migrate, understand, or otherwise re-use the bits. The specific elements included in technical metadata will vary depending on the digital object type and format. While many metadata discussions have focused on the differences in formats -- documents versus images versus streaming media -- there is also significant need for technical metadata to store information about the original environment in which scientific data is captured. This is particularly important with scientific data that is captured from instruments or via computer technologies. For example, the nuclear research data created at CERN has several levels of technical metadata, including the detector description, alignments, calibrations, and the reconstruction parameters (Knobloch 2003).

10.1.4 Structural Metadata

As the complexity of digital objects and their relationship has increased, the need for structural metadata has grown. Structural metadata provides a framework for identifying the relationships between digital objects.

The prevalent standard under development for documenting the structure is the Metadata Encoding & Transmission Standard (METS) (Metadata Encoding and Transmission Standard 2003, Guenther and McCallum 2003) that originated following a February 2001 Digital Library Federation workshop. While METS can accommodate optional metadata for description and administration, the required part of the model is the structural metadata which documents the relationships. For example, METS structural metadata can be used to reconstruct a document made from multiple page image files, or a resource composed of different format types such as audio, video and text. METS can also be used at a higher level to identify the relationship between items in a collection or a digital library.

Because METS is a framework or a model, interoperability between METS structures or support for the creation of OAIS components, such as Archival Information Packages or Submission Information Packages, require agreement on profiles.

METS Version 1.3 is available from the Library of Congress. An editorial board has been created, and LC will act as the maintenance agency. Tools for metadata capture, transformation, and dissemination are under development. Several organizations are currently using METS, including UC Berkeley, the Library of Congress, Harvard, and the University of Virginia Library/Cornell FEDORA effort (M. Smith 2003b). While it isn't clear that any of these implementations have focused exclusively on scientific and technical digital objects, the increasing complexity of objects in the sciences makes this a standard to watch.

10.2 Permanence Ratings

The permanence rating is a specific preservation metadata element that was developed by the US National Library of Medicine. In 1999, the Library began investigating the possibility of a permanence rating system that would support the management and preservation of NLM's Web resources (Byrnes 2000, NLM 2000). Ratings were developed to indicate to users which Web documents will remain permanently available and the extent to which their content could change over time. The ratings are as follows:

Permanent: Unchanging Content
Example: Image of correspondence in NLM's Profiles in Science collection

Permanent: Stable Content
Example: a MEDLINE record

Permanent: Dynamic Content
Example: NLM's Home Page

Permanence Not Guaranteed
Examples: Conference calendars, preliminary agendas

A rating of "Permanent" means that NLM has made a commitment to keep the document permanently available. Its identifier will always provide access to the document. A rating of "Permanence Not Guaranteed" means that the identifier validity and resource availability could change.

Growing: Additional objects may be added to the resource.

Closed: Objects are no longer being added to a resource that previously was subject to growth. ***(Byrnes 2000)***

These ratings were defined in terms of the NLM environment but are being used by other organizations. The US National Agricultural Library has implemented the NLM permanence ratings. NASA Goddard Space Flight Center has included permanence ratings in their digital preservation project plan.

NLM is in the process of modifying its Web management system to accommodate permanence ratings and other additional metadata. The Library expects to have the rating system in place by Spring 2004.

10.3 Open Archival Information System Reference Model (OAIS RM)

In June 2002, the OAIS RM was officially published as ISO Standard 1472. The OAIS RM defines terms and lays out the concepts for an archive, either digital or analog. The Consultative Committee on Space Data Systems (CCSDS) originally developed the OAIS RM for the space data community. However, it was soon acknowledged as a generalized reference model. Now, terms such as "ingest" (meaning taking material into an archive) and acronyms such as "SIP" (submission information package) are commonly used in the community. The language indicates the degree to which the OAIS RM has been accepted.

Many systems have OAIS as the basis, including DIAS, OCLC's Digital Archive, and DSpace. JSTOR has found it to be a valuable framework for the discussion and development of its e-journals pilot. The DiVA Project for the preservation of Swedish theses and dissertations used OAIS as a checklist when it developed its archiving project. Now that DiVA is moving toward the long term preservation portion of its project, the OAIS is being examined in more detail.

One of the common complaints about the OAIS is that it is a reference model and not an implementation. Therefore, the Research Libraries Group has a web site on which it tracks OAIS-based systems, and provides links to schema (RLG 2002b). Included on this list are mappings and schema developed for projects such as DSpace, the e-journals project at the Harvard University Library, and the NEDLIB and Cedars projects. In addition, CCSDS and others have several follow-on activities underway, which provide more detail underneath the OAIS RM, including the development of a checklist for trusted archives and the specifications of XML Formatted Data Units (XFDU) for XML packaging of archive contents.

10.4 Producer-Archive Interface Methodology

As acknowledged earlier, the OAIS RM is not an implementation. There were many concerns raised by those who reviewed the OAIS RM and those who tried to implement the OAIS, that it did not provide enough guidance to be able to create an archive with OAIS compliance. In

fact, ICSTI members commented that they wanted more information about the "ingest" part of the model. Therefore, the CCSDS has drafted a "Producer Archive Interface Abstract Methodology Standard" (CCSDS Document 651.0-R-1) (CCSDS 2003).

Based on a detailed review of the ingest process and the interface between the producer and the archive, the Methodology provides a general framework for the producer-archive relationship. Like the OAIS RM, it does not specify an implementation. It gives a checklist for what should be considered when negotiating an agreement. Phases in the process are identified. The preliminary phrase involves the first contact, the preliminary definition of the project, and whether the project is feasible. Digital objects and the standards to be applied to these objects are identified. The number of items, security considerations, legal and contractual aspects, transfer operations, validation, and the schedule are reviewed. Based on the findings regarding these aspects of the interaction, a preliminary agreement is established. The formalization phase includes the further definition of the objects to be transferred, the identification of specific metadata, and the creation of a data dictionary and formal model. The contracts and legal aspects of the agreement are formalized. Validation routines are written and a delivery schedule is provided in detail. Change management during the life of the project is specified. Once the Submission Agreement is created, the Transfer Phase begins. The initial parts of this phase focus on extensive testing and validating of the system. Modifications are made to the system and the Submission Agreement as necessary.

A final section of the Methodology describes how it can be tailored to create a community-specific standard. Examples are provided and the phases are modified to indicate how a community can approach development of such changes. The key areas involved in achieving community consensus are the further definition of terms, and the creation of an informational model for the community. This community standard conceptually fits between the specific model that is developed between a specific producer and archive (as called for in the Methodology) and the higher generalized model provided by the draft standard itself.

When specialized, the methodology may identify particular standards and tools to be used in the negotiation and submission process. Already there is interest in the Life Sciences community to specialize this methodology for submission of data to Life Sciences archives (Sawyer 2003). If this specialization proceeds, it would become a separate standard.

10.5 Persistent Identifiers

While Persistent Identifiers are indirectly related to standards for digital preservation, they are considered to be a key infrastructure component to help ensure that digital materials can be managed and located in the future. The two most common schemes in use are the Persistent URL (PURL) and URNs, the Handle or Digital Object Identifier (DOI).

The PURL was developed by OCLC and is based on the standard HTTP, URL protocol. It uses a URL that points to a resolver server that must be maintained to redirect a broken URL to the correct one. OCLC's Digital Archive, the US Government Printing Office, and the US Department of Energy's Office of Scientific and Technical Information use the PURL.

The Handle® system is based on the URN concept, but it is not a registered URN namespace. The Handle system assigns unique handle prefixes by naming authorities that are coordinated at the global level to ensure global uniqueness of the resulting identifier. Handles also use resolver services but the "database" construct of the Handle system allows a single Handle to resolve to two or more URLs, which can support different versions, formats or locations of the same work. The Digital Object Identifier (DOI) is an implementation of the Handle. CrossRef, a system for managing persistent identifiers for reference linking among publishers, is a registration authority under the International DOI Foundation, which assigns naming authorities. Other organizations are working on becoming their own naming authorities.

DSpace provides persistent identification as a key component of an institutional repository system. It uses the Handle® System (CNRI 2003) for resolving these identifiers. Each site that runs DSpace obtains a Handle "prefix" from CNRI in order to make the identifier globally unique. The site can then use any scheme for assigning the suffix. Persistent identifiers are assigned to communities, collections, and items. Handles are not assigned to bitstreams, since, over time, the bitstream may change as it is transformed to support preservation activities and new rendering methods. The item is persistently identified and then users access the appropriate bit encoding from that citation. Assignment and resolution of persistent identifiers is also included in the electronic deposit system for the Netherlands (Steenbakkers 2002; van der Werf 1999). The Handle system has been implemented by the Defense Technical Information Center in the US as part of its Digital Virtual Library Architecture. The Stationery Office in the UK was recently added as a DOI Registration Authority.

Most recently, the International DOI Foundation announced a project to assign DOIs to scientific data sets (IDF 2003). The German National Library of Science and Technology will join the IDF for a one-year period funded by a grant from the German Research Foundation. The pilot project will be coordinated by the World Data Center for Climate at the Max-Planck Institut fur Meterologie in Hamburg. The WDCC's pilot will be extensible to other scientific data. The DOIs will be assigned to scientific data sets and then the DOI will be used to reference and cite the primary data, enhancing the ability of future researchers to locate and re-use the primary data and more closely linking the primary data to the resulting published literature.

An alternative to Handles and PURLs, the Archival Resource Key (ARK) is being used by the California Digital Library (Kunze 2003). Based on the concept that persistence is a matter of service and not inherent in the object or the particular naming syntax, the ARK specifies an actionable identifier linking to three services fundamental to the provision of credible persistence. The ARK is a special kind of URL that is divided into a Name Mapping Authority, which is temporary and may be changed or dropped over time, and the persistent Name Assigning Authority Number and Name. The two portions are separated by the ARK label "ark:". The Name Mapping Authority may change from one service provider to another, but the Name Assigning Authority Number and Name do not change. Services are specified that allow an object to be found based on its Name Assigning Authority Number and Name regardless of the Name Mapping Authority. The ARK is supported by services to deliver the user to an object, to deliver the user to the object's metadata, and to deliver the user to a statement of commitment. The latter is a faceted scheme that describes how long the identifier

(the association between the name string and the object) will be valid, how long the object will be available, and how changeable its content may be. The California Digital Library has assigned ARKs to over 150,000 ingested objects, and it is working on the development of the three supporting services described above.

National libraries may opt to use a URN (Uniform Resource Name) scheme based on the National Bibliography Number, since it is an existing scheme for unique identification and it ties easily to the national bibliographies. The DiVA Project at Uppsala University in Sweden uses a URN-based identifier with the National Bibliography Number as the unique identifier (Muller 2003). This provides a convenient mechanism for ensuring interoperability with the National Bibliography produced by submitting Archival Information Packages to the National Library. The German National Library's planned E-Deposit system also uses a URN as the persistent identifier (Germany Persistent ID 2003), based on initial work done under the EPICUR and Carmen projects among several major universities and libraries in Germany. The scheme also uses the ISBN and the National Bibliography Number.

Identifiers unique to the specific system can also be used. The US PubMed Central and the Astrophysics Data System (ADS) do not use globally recognized persistent identifier schemes. However, the identifiers are persistent within these systems. In the case of the ADS, the community is so cohesive around the ADS as a resource that most documents and other objects of interest to the community are stored within the system and there are few references to outside objects to which ADS identifiers have not been assigned. The NLA has implemented a local identifier scheme, so all publications in the PANDORA Archive and their component parts can now be identified and cited using a persistent identifier. The identifier for each title is cited on its title entry page and secondary services are beginning to incorporate the identifier in their bibliographic records. PANDORA is partnering with eight secondary services to archive their indexed resources and to provide a persistent identifier for the citation, including a small number of medical and technical resources. Researchers are interested in using the persistent identifier to cite articles or parts of articles in their works. Support for automatic creation of the persistent identifier is incorporated in the PANDAS software. The local scheme could be converted to a global system at a later date.

10.6 XML DTDs and Schema

The impact that XML is having in the areas of content/document management, information sharing, and cross-platform development is significant for the digital preservation community. Import and export of XML has become more common, and the proliferation of XML schema and DTDs can be seen. Many organizations now have the capability to produce XML easily through commercial products. As this trend continues, the production of a low level bitstream for preservation purposes will increase. PubMed Central, BioMed Central and the DiVA Archive of theses and dissertations are all XML-based.

In the DiVA Archive, the descriptive and administrative metadata are stored in XML that conforms to the DiVA Document Format. The document itself is stored in PDF and, whenever possible, in XML. It is possible for different manifestations of the same document (XML, PDF, etc.) to be created. The metadata are locked and stored in the folder of the corresponding

manifestations. Each archival package contains a single manifestation and consists of an XML file conforming to the DiVA Document Format (DiVA 2003a). The Document Format contains the metadata and as much of the full text content in DocBook as can be created from the source format (MS Word, Star Office, Open Office, Tex/LaTex), the DiVA Document Format specification which is an XML schema, file such as multimedia files that are linked from the DiVA Document Format file, presentation layer files such as style sheets, a PDF file containing the full text, and checksums for all the files. In the future, the DiVA staff hope that PDF can be dropped in favor of XML, if they can guarantee that all the data will remain authentic and easily readable from the XML files without any additional plug-ins for mark-up languages.

The development of community schema will also improve the ability to render the documents into a form that can take advantage of future technologies. The use of XML embraces the concept of keeping the bitstream and the presentation separate, because the part that will change most rapidly is the technology for presentation (Lynch 2002). The key then becomes the metadata, which intercedes between the bitstream and the presentation technology.

In a related effort, the Consultative Committee for Space Data Systems, developer of the OAIS Reference Model, is developing a packaging methodology based on XML called XML Formatted Data Units (XFDU). As a follow-on to the OAIS, the XFDU provides an XML schema for the wrapping of the various "packages", i.e., the Archival Information Package and the Submission Information Package, identified in the OAIS. Funding is pending for continued development of this approach (Sawyer 2003).

11.0 New Issues and the Research Agenda

As prototypes and pilot projects have moved toward more operational environments, there is a renewed interest in research. One such analysis was sponsored by the US National Science Foundation and Library of Congress. A workshop on research challenges in digital archiving and long term preservation was held in 2002. It brought together government program managers, archivists, computer and information scientists, and digital library experts to discuss the issues and shape recommendations for a national research agenda. The discussions and the final report focused on four main themes: technical architectures for archival repositories; attributes of archival collections, digital archiving tools and technologies; and organizational, economic and policy issues (Hedstrom 2003).

Another major international activity related to a research agenda in digital preservation was recently completed by the US National Science Foundation and the European Union under the Fifth Framework Programme by the Network for Excellence in Digital Libraries (DELOS) (Hedstrom and Ross 2003). The report identifies research challenges and opportunities that are common across government, private, university, and cultural heritage institutions. Generally, the research agenda is divided into Emerging Research Domains, Re-engineering Preservation Processes, and Preservation Systems and Technologies. Of the many specific research areas identified under these three categories, the Working Group identified three specific areas that are likely to have the greatest impact. These are the development of self-contextualizing

objects, metadata and the evolution of ontologies, and mechanisms for preserving complex and dynamic objects.

While the NSF/LC and NSF/DELOS reports focus on the broader research agenda, specific issues and possible research areas related to scientific and technical information have been identified based on the interviews and analysis conducted for this study. There are overlaps and it is hoped that these research areas will be addressed by NDIIPP, NSF, DELOS and others with funding initiatives.

11.1 Authenticity

Authenticity is a key issue, particularly for electronic records systems, national archives, corporate archives, and high-risk areas such as health-related data. It is of particular concern to government agencies that are viewed as trusted sources. Authenticity is a security property that has not been discussed in great detail (Gladney and Bennett 2003). In addition to discussions about what authenticity really means, particularly from the user's point of view, there have been discussions about the technologies that can be used to support authenticity of digital information into the future. These include watermarks and other security measures as well as public key infrastructures and digital signatures. In many cases, these technologies will need to be incorporated into archiving systems if they are to be of value for certain constituencies. These technologies have a part to play in recording the provenance of the preservation life cycle.

Authenticity is of particular concern to government archives. The Victorian Electronic Records Strategy has studied the use of digital signatures to ensure the long term authenticity of digital objects (Waugh 2002). The VERS study determined that it is possible to use the archive itself to avoid having to ensure the long term survival of the certificate infrastructure. The Government Printing Office in the US is in the process of implementing a digital signature system, which it hopes will give users assurance that the documents they have in hand are official government documents.

11.2 Rendering Objects for Permanent Access

A major research agenda item for all archives is the lack of best practices concerning the provision of permanent access. Many archives are successfully storing the information, but there is no guarantee, short of saving old hardware, operating systems and software, to ensure that the digital information can be viewed in perpetuity. The joint NSF/EU agenda includes this item.

An area of investigation in JSTOR's Electronic-Archiving Initiative is to determine what users are expecting from the archiving of electronic journals. Are they expecting to see a replica of the original online e-journal, as they now do, with a digitised version of the paper journal? Or, are users primarily interested in the content regardless of whether it is precisely rendered as originally presented online? What are the "visual cues of trust" that need to be considered (Fenton 2003)?

11.3 Saving the Dynamic Web

A particularly difficult aspect of dealing with archiving Web resources is the increased use of dynamic Web content (Arms 2001, Kahle 2003). This content may be coming from databases, from content management systems, via active server pages, etc. In these cases, the "view" or the rendering of what is seen changes depending on the activity, usually a search or other request made by the user. Replicating the web pages in this instance is not only a matter of capturing the HTML, but also retaining the background databases and the software that intermediates and then presents the content requested by the user. In terms of archiving for legal purposes, it would also be necessary for the user or some system to keep track of the users' request, since it is more of a "dialog" than a one-way presentation of information. Some suggestions on how to deal with the deep web were provided by William Arms (2001), but they work best when partnerships have been developed with the web site owners.

The inaccessibility of publications structured as databases is an ongoing issue for the PANDORA Archive (Phillips 2003). These types of publications are not currently included in the Archive. The dynamic web, the deep web, and other technical issues such as non-standard browser plug-ins were highlighted as issues in the Wellcome Trust and JISC-sponsored investigation of web archiving (Day 2003). The NASA Goddard Space Flight Center Library has also noted problems in archiving Web pages that are generated from content management systems or via portals, as well as mouse-over and intermittent or dynamic animations (Ormes and Hodge 2003). The dynamic nature of e-records is part of the research being conducted in the InterPARES II Project, which began in 2002 and will conclude in 2006. "It will focus on records produced in new digital environments, experiential, dynamic, and interactive..." (InterPARES 2003). Research projects are underway at the California Digital Library, Stanford University, and within the WebFountain Project of the IBM Almaden Research Laboratory (IBM Almaden Research Laboratory 2003).

The National Library of Australia, the Bibliotheque Nationale, the Library of Congress and other national libraries have formed the International Internet Preservation Consortium's Deep Web Working Group. This effort, just getting underway in the Fall of 2003, will be working to determine the requirements and technical solutions for preserving content that is in the Deep Web (Massanes 2003).

11.4 Appraising and Retaining Scientific Data

There is increased impetus on the part of organizations involved with the creation, management, and exchange of scientific data to address issues related to the preservation of this data. Some of the key issues surrounding scientific data are the size of many of these files, the need to retain and understand the structure of the original data, and the need to have the data useable in a computer, rather than a human readable environment. A key question is the metadata that is needed for discovery, preservation, and reuse, and how to create this metadata in a cost effective, efficient and accurate manner. An early report (Uhlir 1995) raised the major issues, which have not changed significantly over time.

However, the increased use of data as primary research in informatics-based science; the potentials to exchange, manipulate and collaborate using this data on the Grid; and concerns about the lack of attention to these issues have caused major groups such as CODATA, the US National Science Board, and the World Data Centers to become increasingly involved in discussions about appraising and preserving scientific data.

A key sub-area of investigation is the degree to which the appraisal and retention criteria differ from discipline to discipline and even by sub-disciplines. This was highlighted in a recent workshop sponsored by ERPANET and CODATA (Ross and Uhlir 2003). The various disciplinary case studies highlighted different definitions for raw versus processed data. For example, the nuclear physics research at CERN is conducted using a series of tunnel experiments. The original data flow is tremendous but CERN's system redacts the data and selects events that are then considered to be the raw data. Events are reconstructed and simulations are used to "process backwards" to understand the influences that are found. The reconstructed events become event summary data which are the real meaningful data for this type of physics (Knobloch 2003). The workshop identified a need for tools to support the incorporation of archiving principles as early in the data creation process as possible; a need to raise awareness about the need for data preservation, particularly among managers and funding institutions; the need for a better understanding of how scientific disciplines use and reuse data; and general guidance for appraisal, retention and preservation that could be tailored to the needs of particular sub-disciplines or research groups.

Many of the same issues were raised in a white paper for a recent workshop sponsored by the National Science Board on behalf of the National Science Foundation. The goal of the workshop was to address what level of support the NSF might provide for collecting, organizing and preserving data, how the increased importance of data in e-science might impact the cyberinfrastructure needed to conduct research in the future, and how this might impact federal agencies and grantees. Technology, intellectual property rights, and national and international policy issues were discussed.

11.5 Preserving Government Information

Perhaps because of the emphasis on e-Government, significant concerns have been raised about the preservation of digital government information, particularly by national archives and national libraries. As e-Government initiatives are gaining momentum in Europe, Canada, the US and Australia, in particular, there are more publications than the libraries can identify, capture, catalog and preserve with their own resources.

In Australia, for example, the online version is now the primary format for Commonwealth government publications and they are being produced in numbers that far exceed PANDORA's capacity to archive given current procedures and infrastructure (Phillips 2003). There is an urgent need to find a way of increasing PANDORA's ability to harvest and capture government information, and they are investigating ways of automating some of the identification, selection, description and archiving processes. To this end, in 2003, the NLA began a pilot project to work with seven government agencies to develop workflows and procedures for ingesting metadata about the agencies' publications into the Library's National

Bibliographic Database. The goal is to develop a model or a small number of models that government agencies can use to contribute metadata, in order to reduce the Library's selection and cataloging effort. In most cases, this will involve taking in metadata in a form other than MARC and then converting to MARC. Once this metadata is in the Library's database, it will be extracted in a form that can be fed into PANDAS in batch mode. PANDAS will have to be modified to accept, register and process batched data automatically or semi-automatically. In 2004, the pilot project will continue with the addition of more agencies.

In addition, there are special efforts underway at certain libraries to ensure that government information is more likely to be transferred for long term preservation. The legal deposit provisions in the Australian Copyright Act 1968 still do not cover electronic publications at the Commonwealth level. For the past 18 months, the Library, together with ScreenSound Australia, has been working with the relevant government departments to push for the needed amendments. The need to obtain permissions from individual publishers continues to be a time-consuming aspect of PANDORA'S work (Phillips 2003). Therefore, PANDORA has entered into an agreement with the Commonwealth Copyright Administration in Australia, which is actively negotiating with Commonwealth agencies for blanket permission to archive all publications on their domains. To date, 28 agencies for a total of 70 domains have given their permission. A number of these agencies have scientific content including the Defence Science and Technology Organisation, The Australian Greenhouse Office, and Biotechnology Australia. These arrangements will enhance PANDORA's ability to archive larger quantities of government information.

The US Congress has mandated that the Government Printing Office and the Federal Depository Library Program move to increasingly electronic submissions and dissemination (Government Printing Office 1998). This requirement came as the GPO was faced with a changing environment in which agencies were able to publish their materials on the Web instead of through the traditional print process. GPO traditionally obtained its materials by riding the print orders for agency publications, adding enough copies to supply them to the Federal Depository Libraries as required. To handle these changes, a new, more electronic flow was required. Therefore, GPO has developed a system for harvesting, cataloging and preserving government publications.

The creation of its own system is in addition to work with the OCLC Digital Archive and other organizations to provide supplemental and compatible archives of government publications (Barnum 2003). GPO is making arrangements with libraries that are members of its Federal Depository Library Program to take control of information from specific agencies. The National Library of Medicine and the National Agricultural Library have also made arrangements with the National Archives and with the Government Printing Office to ensure preservation of their databases over time.

Stanford University has received a $50,000 planning grant from the National Science Foundation to explore the potential applicability of the LOCKSS technology to preservation of US Federal government information (named LOCKSS-DOCS). LOCKSS and the team of eight content partners from the government documents community are exploring technical, economic, social and legal viability of various LOCKSS architecture models for the GPO

Federal Depository Library Program. It is hoped that LOCKSS-DOCS will be further developed into a system that can be used by other federal government agencies (LOCKSS-DOCS 2002).

11.6 Archiving the Archive

As noted by the checklist for the determination of a reliable archive, it is important to investigate whether the archive itself has sufficient preservation policies. Replication and reliable security, backup, and recovery procedures are cited as indications of archive trustworthiness (RLG 2002a). The benefit of redundancy of an object across multiple archives has been discussed.

The Internet Archive recently made arrangements for a complete mirror of its site (including the hardware and software) in the new Library of Alexandria. Kahle (2003) points out the importance of ensuring that archives are duplicated under other political conditions, in disparate geographic areas, and under different cultural regimes, etc., as a means of guarding against physical damage as well as neglect.

Within the DiVA Project in Sweden, the cooperation and interoperability between the local university repository and the Swedish National Library guarantees future access to the material, even if the local DiVA archive closes. A resolver (using the persistent identifier) will redirect the traffic to the copy at the national library.

PubMed Central has addressed the issue of archiving the archive. Sequeira (2003) indicates that PMC would like to have other non-profit archives maintain copies of the PMC under the same copyright, ownership, and free access conditions with which the PMC is provided. Even though the PMC has backup copies on DVD and follows the normal off-site storage and disaster recovery procedures, duplication of the archive (particularly through a lit archive) will help to ensure greater possibility of recovery. Some European organizations have expressed an interest in serving in this role, but the technical infrastructure for keeping the archives synchronized is not yet in place.

"Losing the metadata that describes the content is a catastrophic failure for an archive as it would normally mean the effective loss of the digital objects. It seems to be often forgotten that the archival management system is complex and has a far shorter life than the digital objects it holds" (Quenault 2003). It is for this reason that archives, such as the Victorian Electronic Records Strategy, encapsulate the metadata with the object as described in Section 8.4. In addition to routine backup and recovery, self-describing objects provide another means of securing the archive.

Similarly, PubMed Central is aiming to make each article file self-defining. This includes copyright and some control elements, such as dates, volume and issue. Version control is needed for multiple submissions.

11.7 Interoperable Archives

In addition to sharing best practices and the cost of developing preservation systems, organizations are considering the benefits of sharing the preserved content through interoperable archives. The definition and degree of interoperability may vary. In some cases, the emphasis is on providing backup and redundancy in the case of disaster, while in other cases the emphasis is on providing cross-archive access.

The DiVA Project shares both content and access. DiVA's full text and metadata flow between distributed local repositories at universities and the national library (Muller 2003a, Muller 2003b). The DiVA-created Archival Information Packages (Archival Information Packages from the OAIS RM) for theses and dissertations are submitted to the National Library to satisfy legal deposit requirements. In addition, the DiVA Portal provides a single point of access for searching the various repositories (DiVA 2003b). The network members also share in the development of related tools and services.

Similarly, the DSpace Federation is interested in establishing a network of repositories that might provide a wide array of content and services across repositories at academic libraries and institutions. Based on the DSpace Institutional Repository software, and, in particular, the Open Archives Initiative protocol for harvesting and searching repository metadata, DSpace Federation members would be able to offer access across the repositories of the Federation as well as specialized services that could be used by members of the Federation.

However, critical to the development of interoperable archives is a clearer definition of what constitutes "interoperability." In addition, there are policy, security, and technical issues that must be included in such a definition.

11.8 Partnerships

Partnerships have always been important in the digital preservation community. From the very beginning it was apparent that no one organization – whether library, government or academic – could adequately archive, preserve and continue to provide access to the digital material, even with stringent selection criteria.

This is the hallmark of the National Digital Information Infrastructure for Preservation Program (NDIIPP) at the US Library of Congress in that it calls for an infrastructure, and, from the outset, has included participation from a wide variety of traditional and non-traditional organizations, including those from the entertainment industry. The importance of partnerships to the NDIIPP infrastructure is particularly apparent in the recent call for proposals. The call for proposals quoted Laura Campbell, Associate Librarian for Strategic Initiatives who is leading the NDIIPP effort, as saying "The Library of Congress looks forward to collaborating with many partners in this task, as we work together to preserve America's digital heritage." A major goal of the first set of projects is to develop a network of NDIIPP partners with defined roles and responsibilities to support the long term collection and preservation of born digital content (Library of Congress 2003c). The enabling legislation for NDIIPP requires LC to collaborate with various stakeholders within the federal government as well as outside.

The DSpace Federation Project (2003) involves eight universities (as of 30 December 2003) that have implemented the DSpace system. The goal of the project, funded by a series of Mellon grants, is to share experiences and to determine what they can do most effectively together in terms of enhancing the software, identifying funding to sustain the system and its development, and additional value-added services that can be provided. One value-added service that has been identified is that of "journals" created on-the-fly by searching across the repositories (Branchofsky 2003).

The importance of partnerships is highlighted in Day's report on web archiving (Day 2003). He encourages not only the Wellcome Trust and JISC to partner, but to work with the Digital Preservation Coalition in the UK and the British Library to further these efforts. He notes that nationally and internationally, the web does not know geographic boundaries and, therefore, partnerships are needed to guard against gaps in the archiving process, to reduce redundancy, to share costs, and to develop tools and best practices.

In order to further these partnerships and to make interoperable and federated archives a reality, additional work is needed on standards and best practices. Key components will be efforts such as the Open Archives Initiative-Protocol for Metadata Harvesting (OAI-PMH), technologies for cross-database searching, and metadata registries to support interoperability such as those proposed by the CORES initiative (CORES 2003, Heery 2003).

11.9 Costs and Sustainability

Almost from the initial discussion of digital preservation, there has been concern about the costs and sustainability of digital archives. Discussions about costs and cost models can be found in Hendley (1998), Sanett (2002), Russell and Weinberger (2000), and Beagrie and Jones (2001). Sanett (2003) even provides a framework for evaluating the cost models themselves.

Lavoie (2003) views economics as a key component to the research agenda for digital preservation. He equates economics to incentives, and so determining the incentives for stakeholders in the digital preservation process is key to ensuring that these activities will take place. Incentives are impacted by the particular structure for digital preservation, i.e., the organization of the various stakeholder roles. In addition, there are several characteristics of digital information that reduce incentives or make it difficult to remedy the lack of incentives. These include the fact that the owners of digital materials are often different from those who would benefit most from the preservation of the materials and the ease with which multiple copies can be made, making the preserved copy non-unique. Furthermore, the variety in the demand for long term preservation services from low-end to high-end will increase the cost of providing follow-on services and, potentially reduce the revenues from such services.

In the meantime, organizations are trying to fund the building of collections, many through the provision of value-added services related to preservation itself. For example, OCLC is providing a system called ContentDM for the management of digital library collections; this software can be leased or used on the OCLC server for a fee. OCLC also has developed a

consortium called the Digital Co-op. By joining the Co-op, members are eligible for discounts on a variety of services, including the licensing of software, training classes, and digitization support services. Originally, the plan was to charge a fee for membership in the Co-op, but this was dropped at the May 2003 meeting.

MIT developed a business plan to transform the DSpace research project into a sustainable technology platform and service. This was funded by a grant from the Mellon Foundation in April 2000. The basis for this operational system is the DSpace Digital Depository system, which is detailed in the section on Off-the-Shelf Systems. Internally, MIT is providing value-added services, such as metadata creation. This will remove a barrier to submission of content to the repository and also ensure more complete and high quality metadata. The charge for this service will be part of the sustainability model.

Those organizations that provide third-party archiving services have various pricing models. OCLC Digital Archive bases its pricing structure on whether the Web document archiving is done individually or in batch. The Web document archiving consists of an annual subscription fee, plus monthly storage costs. The storage fees are based on the number of gigabytes of content that are stored in the archive. The batch ingest will be fee-based rather than subscription, but the specific costs have not yet been formalized. The same monthly storage fees apply for batch. Digital and Preservation Cooperative members receive a discount on subscription fees.

JSTOR's archive of digitised print journals is based on agreements with publishers and tiered participation payments from libraries and other institutions for access. The pricing model includes a base archive fee, which is intended to sustain the archive and tiered pricing by library type and number of collections accessed. Part of the Electronic-Archiving Initiative is to determine if this is a viable business model for e-journals. To support this effort, JSTOR has launched a study to examine whether non-subscription expenditures for journals are higher or lower in electronic form than in print form. Non-subscription expenditures include collection development, cataloguing, storage, ongoing access, etc. Long term cost implications are being analyzed using a life-cycle methodology (Fenton & Schonfeld 2003).

The cost for data migration by scientific data centers is extremely high. For example, the Earth Resources Observation Systems (EROS) Data Center estimates the cost for each migration at $1 million (Faundeen 2003a). The ongoing cost for maintenance and acquisition of the data is approximately $3 million per year.

Cost factors into many of the decisions noted in the previous sections of this report, including the selection of preservation strategies, the metadata standards and creation procedures, and interface design. VERS notes that because an archive is committing to support forever each long term preservation format (or at least until they commit to migrating every object to another format), costs for licensing and writing new migration and viewing software must be considered, resulting in the position of VERS management that the number of formats should be strictly limited (Quenault 2003).

The inclusion of several commercial technology firms in partnerships with academic and government organizations (HP, IBM, etc.) raises the question of "what is in it for them?" Spedding (2003) reports that these companies are learning from their R&D process and that, eventually, the results will be incorporated into commercial archiving products aimed at academic and business environments, such as records management and digital asset management systems. Ultimately, the more traditional preservation markets will benefit from the reduced prices that such commercialization will undoubtedly bring.

There has been significant research into the cost of preservation that should provide basic information needed for organizations to make decisions about whether to maintain their own systems or look to trusted third parties. This research will also help repository providers to determine what services to provide, to what clients, and with what level of sustainable return. Chapman (2003), in a comparison of archiving costs between OCLC's Digital Archive and the Harvard Repository, notes that there is no single factor that determines the cost of preservation (such as the number of bits). Instead, it is a number of factors including the format (ASCII being the cheapest to preserve), the size of some objects, and the services provided by the repository (acquisition versus submission, simple ingest versus transformation, limited or full access services over the long term). This study of paper versus digital archives with essentially the same business models includes comparisons of specific scenarios such as text, images, and audio. While Chapman frames the study as being very narrowly focused and constrained, it also provides some key information that can be further analyzed by individual archives and object owners as they make decisions about operational systems.

While several organizations have determined that operational systems are needed despite the limited cost data and the lack of firm business models, there are organizations that are still evaluating the situation. Dumouchel (2003) voices the thoughts of many others when he says that his organization, The Canadian Institute for Scientific and Technical Information (CISTI), is evaluating its involvement in this critical activity for Canada. CISTI has potential partnerships lined up with publishers, but "the amount of resources required to undertake initiatives in this field looks somewhat daunting; we are carrying out more in-depth reviews before committing such resources."

12.0 Findings and Trends

12.1 Systems solutions are being developed by a variety of stakeholders and partnerships

The advent of off-the-shelf solutions shows advancing maturity in the area of digital preservation. The library model with shared cataloging tools and service providers is apparent. The six operational systems that are available as open source or "off-the-shelf", the OCLC Digital Archive, DSpace, LOCKSS (Lots of Copies Keep Stuff Safe), Fedora™, PANDAS, and the Digital Information Archive System (DIAS) from IBM, come from different types of organizations – a library service provider, a university repository, a large academic research library paired with a provider of publishing services, a university repository teamed with another university's digital library research group, a national library system, and a national library working with a commercial company. These partnerships show the need for interactions among a variety of stakeholders.

12.2 The Open Archival Information System (OAIS) Reference Model has been widely adopted

The OAIS Reference Model, which became an ISO standard in June 2003, has been adopted widely. All types of archives use the OAIS terminology and conceptual model. However, it is not as prevalent in the scientific data community for which it was initiated, partly because these organizations already had systems, customers, producers, and processes of a legacy nature. Efforts are underway among some data archives to minimally ingest Submission Information Packages (SIPs) and to produce Dissemination Information Packages (DIPs) in order to respond to the spirit of the standard. As systems are redesigned and the need for interoperability increases, it is likely that the OAIS Model will become more prevalent as the conceptual basis for scientific archives.

12.3 Organizations are focused on capturing and acquiring digital information, rather than preservation or permanent access

Even if they use the term archive or have preservation in their mission, the initial goal is to get a critical mass of material; to promote a culture of deposit, submission, harvesting, and sharing; and to provide access to the currently collected materials. While many of the institutional repository activities are committed to long term preservation and access, the technical and metadata aspects required are not yet well incorporated into their systems.

12.4 Efforts for digital depository legislation are gaining momentum

There are significant activities on the part of national libraries and other stakeholder groups with regard to changing existing laws or adding new laws that would require deposit of digital materials. This has gained significant momentum over the last several years, and, most recently, the UK and New Zealand have passed such legislation. Digital deposit legislation may be more accepted now that there have been major pilot projects involving national libraries and large commercial publishers. In addition, voluntary arrangements are already in place, so the legislation more closely reflects current practice rather than leading it.

12.5 Migration remains the preservation strategy of choice; it is still too soon for most archives to have undergone a significant technological change

Other than the large data archives, which have existed for many years, archives have not yet faced large-scale technological changes. This means that migration remains the strategy for most of the materials of interest to libraries, archives, and publishers. The prevalence of migration, particularly from one version of software to another, also indicates the prevalence of commercially available products, such as Microsoft Office and Adobe products, in the scientific environment. While concerns were expressed about outdated software, hardware, and media, these issues are not the current focus as the institutions grapple with collecting and ingesting the flood of current archival content.

12.6 There are increased standards-related activities

There are standards-related activities underway in the areas of producer-archive interaction, permanence ratings, persistent identifiers as critical components of digital preservation systems, preservation metadata, and preservation formats (e.g., PDF-A for text). These activities are likely to produce significant results because they are codifying many of the best practices that have been identified over the last several years of pilot projects.

12.7 Open standards developed for interoperability hold promise as the basis for preservation formats

While the main rationale for development of open standards is interoperability among software environments, these standards may be applicable for long term archiving. Open formats such as those for geographic information systems (OpenGIS), product design and manufacturing (STEP), open office documents (OpenOffice), and chemical structures (Molfiles and SMILES) are working toward hardware and software independence. The potential for using these formats for preservation should be investigated further.

12.8 Key technical issues remain

There are several key technical areas requiring future research that have been identified in recent studies funded by the US National Science Foundation. Additional research is needed into the automatic generation of metadata, through self-describing objects, or the provision of archiving mechanisms in authoring tools. Registries, perhaps of a global nature, are needed to maintain authoritative, computer-actionable information about metadata tag sets, reference information for formats, and hardware/software behaviors. Research into the archiving and preservation of dynamic, non-HTML, and database-driven Web content is a major research activity for several groups. Other technical issues include creating interoperable archives and best practices for archiving and preserving the archive itself.

12.9 Partnerships are increasingly important

Over the last several years, there has been an increasing realization that partnerships are the only way to ensure that digital information will be preserved. In addition to ensuring some measure of comprehensiveness over the wide spectrum of scientific information in digital form, partnerships have the benefit of providing some measure of redundancy, sustainability, and sharing of the cost for preservation which is likely to exceed the revenues that can be made on the reuse of any particular object. A workable infrastructure will result from a multi-pronged approach involving publishers, libraries, archives, institutions, and trusted third parties, with appropriate support from governments, other funding sources, users and creators during the life cycle of the material to be preserved.

12.10 Key social, political, and economic issues remain, including the need to develop a "will to preserve and provide permanent access" within the scientific and technical community and society in general

There are several outstanding social and political issues that require further discussions by the various stakeholder groups involved in preserving scientific and technical information. For example, the social, political and legal aspects of creating federated archives and working partnerships that cross stakeholder groups and object types (data, publications, multimedia, etc.) must be resolved. The archiving and preservation of government information poses special challenges in this regard. Sustainable business models that will survive for the long term also remain elusive. Collecting information about the cost of digital archiving and preservation proved to be as difficult as in the first report, with most of the respondents unable or unwilling to provide cost information. However, several major organizations (OCLC, DSpace, National Library of Australia) are trying value-added services and licensing of software to other organizations as ways of offsetting the cost of preservation activities.

Overriding these social, political, and economic issues is the need to develop within the scientific and technical community and society in general a culture that encourages the "will to preserve." Waters (2002) argues that the archiving and preservation of information must be perceived as a "public good." A major initiative at the Cornell University Library focuses on training, including a new online course on digital preservation management that emphasizes the implementation of practical short-term strategies (Cornell, University Library 2003). Similarly, education of scientists, either in academic institutions or during work-based training, could raise the awareness of the importance of preservation for the public good and for the good of the scientific community.

13.0 Recommended Next Steps

The work on digital preservation is continuing apace with significant developments in off-the-shelf, generalized systems, legal deposit legislation, partnerships and federations, and standards activities. However, much remains to be done. The following sections suggest how CENDI and ICSTI, independently, jointly, or in concert with other groups, could help to move the digital preservation agenda forward.

ICSTI:

1) Continue to work with the Committee on Data (CODATA), the International Council for Science (ICSU), individual scientific unions, institutional repositories, and university management on issues related to the archiving and preservation of data and its relationship to publications. A key component of this effort should be identifying the similarities and differences between preserving data (of various types) that results from scientific research and the textual documentation such as journal articles and technical reports. Another key area of investigation is the identification of similarities and differences between preserving data in various disciplines. It will be important to determine what standards can be shared and what must be different.

2) Analyze the impact of Open Access (including author self-archiving), institutional repositories, and e-Science initiatives on digital preservation and permanent access, and identify a framework in which all these initiatives can be successfully achieved.

3) Investigate the usefulness of interoperability standards, such as OpenGIS, OpenOffice and STEP, for long term preservation formats.

4) Promote the "will to preserve and provide permanent access" as well as best practices by encouraging the incorporation of preservation concepts in science education and work-based training. This might be done in collaboration with science education organizations such as the American Association for the Advancement of Science, learned societies, and academia.

5) Produce a list of foundations that are interested in supporting digital preservation in science in order to help those looking for funding. This could be done by polling members of ICSTI and those involved in digital preservation at national libraries and academic research libraries.

CENDI:

1) Work with the US Government Printing Office (GPO), the National Archives and Records Administration (NARA), the LOCKSS-DOCS Project, the Library of Congress (particularly the NDIIPP), and others to develop effective and sustainable preservation guidelines for government scientific and technical information in the context of the R&D component of the Federal Enterprise Architecture, e-Government, the federal research process, and the environments of the federal science agencies.

2) Host a follow-on workshop to the previous workshops sponsored by the National Science Board and the National Archives and Records Administration to continue the discussions about the selection, retention, organization and on-going preservation of scientific data produced as a result of government funding.

3) Support the development of technical and social solutions to the federation of archives, which are likely to be needed to address the preservation of government information. This would involve the development of core metadata standards for different digital objects, the implementation of high level Open Archival Information System Reference Model functions, and producer-archive interaction checklists specific to the federal science environment.

4) Continue involvement with standards efforts; specifically, review the Journal Publishing DTD and the DTD for Technical Reports, and determine how these efforts might be addressed by the agencies and as part of the Federal Enterprise Architecture.

5) Support the development of technical solutions for archiving the dynamic and deep Web.

14.0 References

Allen, R. (2003). "Metadata for Project Resources: Development of the Goddard Core." Presented at DC-2003, Seattle WA, 28 September – 2 October 2003.

American Institute of Physics. (2003) "Publishing Services." [Online]. Available: http://www.aip.org/publishing/services/cs_archleg.html [29 April 2004]

Anderson, W. (2003). "Introduction to ERPANET/CODATA" workshop. Presented at the ERPANET/CODATA International Archiving Workshop on the Selection, Appraisal and Retention of Digital Scientific Data held in Lisbon, Portugal, 15-17 December 2003.

Andrew W. Mellon Foundation. (2003). [Online]. Available: www.mellon.org [29 April 2004]

Arms, W. (2001). "Web Preservation Project: Final Report." 3 September 2001. [Online]. Available: http://www.loc.gov/minerva/webpresf.pdf [29 April 2004]

Barnum, G. (2003). Personal communication.

Beagrie, N. and Jones, M. (2001). *Preservation Management of Digital Materials: a Handbook.* [Online]. Available: http://www.dpconline.org/graphics/handbook/index.html [29 April 2004]

Beck, J. (2003). "PubMed Central & the NLM DTDs." Presented at the ASIS&T DASER Summit held in Cambridge, MA, 21-23 November 2003. [Online]. Available: http://www.asis.org/Chapters/neasis/daser/Jeff_Beck_presentation.ppt [29 April 2004]

Bellardo, L. (2003). "Revolutionizing E-Records: Agency NARA Partnerships." Presented at the February 2003 CENDI Meeting. [Online]. Available: http://www.dtic.mil/cendi/minutes/pa_0203.html#nara [29 April 2004]

Bethesda Statement on Open Access Publishing. (2003). [Online]. Available: http://www.earlham.edu/~peters/fos/bethesda.htm [29 April 2004]

BioMed Central. (2003). "National Library of the Netherlands and BioMed Central Agree to Open Access Archive": Press release, 27 September 2003. [Online]. Available: http://www.biomedcentral.com/info/about/pr-releases?pr=20030917 [29 April 2004]

Branschofsky, M. (2003, July). Personal communication.

British Library Press & Public Relations. (2003). "Historic Change in Legal Deposit Law Saves Electronic Publications for Future Generations – Bill to Extend Legal Deposit to UK Non-print Materials Receives Royal Assent." British Library Press Release, 31 October 2003. [Online]. Available: http://www.bl.uk/cgi-bin/press.cgi?story=1382 [29 April 2004]

Brown, I. D. & B. McMahon. (2002). "CIF: the Computer Language of Crystallography." *Acta Crystallogr. Section B*, 58(3), 317-324.

Budapest Open Access Initiative. (2002). [Online]. Available: http://www.soros.org/openaccess/read.shtml [29 April 2004]

Butterworth, I. (2003, September). Personal communication.

Byrnes, M. (2000). "Assigning Permanence Levels to NLM's Electronic Publications." Presented at *Information Infrastructures for Digital Preservation: A One Day Workshop*, 6 December 2000, York, England. [Online]. Available: http://www.rlg.org/events/pres-2000/infopapers.html/byrnes.html [29 April 2004]

Byrnes, M. (2003, September). Personal communication.

CAMiLEON: Creative Archiving at Michigan & Leeds: Emulating the Old on the New. (2001). [Online]. Available: http://www.si.umich.edu/CAMILEON/ [29 April 2004]

Carroll, B. & G. Hodge. (1999). "Digital Electronic Archiving: The State of the Art, the State of the Practice." [Online].

Available: http://www.icsti.org/Dig_Archiving_Report_1999.pdf [29 April 2004]
Cedars: CURL Exemplars in Digital Archives. [Online]. Available: http://www.leeds.ac.uk/cedars/ [29 April 2004]
Chapman, S. (2003). "Counting the Costs of Digital Preservation: Is Repository Storage Affordable?" [Online]. Available: http://jodi.ecs.soton.ac.uk/Articles/v04/i02/Chapman/chapman-final.pdf [29 April 2004]
Charlesworth, A. (2003). "Legal Issues Relating to the Archiving of Internet Resources in the UK, EU, US and Australia: A Study Undertaken for the JISC and Wellcome Trust." [Online]. Available: http://library.wellcome.ac.uk/assets/WTL039230.pdf [29 April 2004]
CODATA. (2003). "International Symposium on Open Access and Public Domain in Digital Data & Information for Science." Organized by ICSU, UNESCO, the US National Academies, CODATA and ICSTI. Held in Paris, France, 10-11 March 2003. [Online]. Available: http://www.codata.org/archives/2003/03march/index.html [29 April 2004]
Committee on an Information Technology Strategy for the Library of Congress, Computer Sciences and Telecommunications Board, National Research Council. (2001). "LC21: A Digital Strategy for the Library of Congress." National Academy Press: Washington DC. [Online]. Available: http://books.nap.edu/books/0309071445/html/index.html [29 April 2004]
CCSDS (Consultative Committee for Space Data Systems). (2002). "Producer-Archive Interface Methodology Abstract Standard." CCSDS-651.0-R-1. Red Book. December 2002. [Online]. Available: http://ssdoo.gsfc.nasa.gov/nost/isoas/CCSDS-651.0-R-1-draft.pdf [29 April 2004]
Consultative Committee for Space Data Systems. (2001). "Reference Model for an Open Archival Information System (OAIS)." Red Book CCSDS 650.0-R-2, June 2001. [Online]. Available: http://ssdoo.gsfc.nasa.gov/nost/isoas/ref_model.html [29 April 2004]
CORES.(2003) "CORES: Summary of 2003 Activities (Final Report)." [Online]. Available: http://www.cores-eu.net/final-report/ [29 April 2004]
Cornell University Library. (2003). "Digital Preservation Management: Implementing Short-term Strategies for Long-term Problems." [Online tutorial]. Available: http://www.library.cornell.edu/iris/tutorial/dpm/ [29 April 2004]
Cox, J. & L. Cox. (2003). "Scholarly Publishing Practice: the ALPSP Report on Academic Journal Publishers' Policies and Practices in Online Publishing." (Executive Summary). [Online]. Available: http://www.alpsp.org/news/sppsummary0603.pdf [29 April 2004]
Crowe, Raym. (2002) "The Case for Institutional Repositories: A SPARC Position Paper." [Online]. Available: http://www.arl.org/sparc/IR/ir.html [29 April 2004]
Dack, D. (2001). "Persistent Identification Systems." [Online]. Available: http://www.nla.gov.au/initiatives/persistence/PIcontents.html [29 April 2004]
Day, M. (2003). "Collecting and Preserving the World Wide Web: A Feasibility Study Undertaken for the JISC and Wellcome Trust." [Online]. Available: http://library.wellcome.ac.uk/assets/WTL039229.pdf [29 April 2004]
Defense Technical Information Center. (2002) "Technical Metadata for the Long term Management of Digital Materials: Preliminary Guidelines." March 2002. [Online]. Available: http://dvl.dtic.mil/metadata_guidelines/TechMetadata_26Mar02_1400.pdf. [27 January 2004]

Dissertations Online. [Online]. Available: http://www.dissonline.de/index_e.htm [29 April 2004]

DiVA Project. (2003a). "DiVA Document Format." [Online]. Available: http://publications.uu.se/schema/1.0/diva.xsd [29 April 2004]

DiVA Project. (2003b). [Online]. Available: http://www.diva-portal.se/ [29 April 2004]

DiVA Project. (2003c). "Metadata Workflow Based on Reuse of Original Data." [Online] Available: http://publications.uu.se/etd2003/papers/MetadataWorkflow.pdf [29 April 2004]

DSpace Federation. (2003a). [Online]. Available: http://www.dspace.org/ [27 January 2004]

DSpace Federation. (2003b). "The DSpace Federation Project." [Online]. Available: http://dspace.org/federation/project.html [27 January 2004]

DSpace Federation. (2003c). "Content Guidelines for DSpace at MIT." [Online]. Available: http://libraries.mit.edu/dspace-mit/mit/policies/content.html [27 January 2004]

Eastwood, T. (2003). "Overview of Selection, Appraisal and Retention of Scientific Data Across Disciplines." Presented at the ERPANET/CODATA International Archiving Workshop on the Selection, Appraisal and Retention of Digital Scientific Data held in Lisbon, Portugal, 15-17 December 2003.

Electronic Resource Preservation and Access NETwork: ERPANET. [Online]. Available: http://www.erpanet.org [29 April 2004]

ERPANET. (2003). "ERPANET/CODATA International Archiving Workshop on the Selection, Appraisal and Retention of Digital Scientific Data." Lisbon, Portugal, 15-17 December 2003.

Faundeen, J. (2003a). Personal communication.

Faundeen, J. (2003b). "US Land Remote Sensing Archive." Presented at the ERPANET/CODATA International Archiving Workshop on the Selection, Appraisal and Retention of Digital Scientific Data held in Lisbon, Portugal, 15-17 December 2003.

Feenstra, B. (2000). "Standards for Implementation of a DSEP." *NEDLIB Report Series*; #4, Koninklijke Bibliotheek: Den Haag. [Online]. Available: http://www.kb.nl/coop/nedlib/results/NEDLIBstandards.pdf [29 April 2004]

Fenton, E. (2003). Personal communication.

Fenton, E. & R. Schonfeld. (2003). "Digital Preservation Library Periodicals Expenses: Variance between Non-Subscription Costs for Print and electronic Formats on a Life-Cycle Basis." Presented at the Fall 2003 CNI Task Force Meeting, Portland, OR, 8-9 December 2003. [Online]. Available: http://www.cni.org/tfms/2003b.fall/abstracts/PB-digital-fenton.html [29 April 2004]

Gilleland-Swetland, A. (2000). "Setting the Stage." In Introduction to Metadata: Pathways to Digital Information, M. Baca (ed.) [Online]. Available: http://www.getty.edu/research/institute/standards/intrometadata/2_articles/index.html [29 April 2004]

Gladney, H. M. & J. L. Bennett. (2003). "What Do You Mean by Authentic?: What's the Real McCoy?" *D-Lib Magazine*, July/August 2003, Vol. 9, No. 7/8. [Online]. Available: http://www.dlib.org/dlib/july03/gladney/07gladney.html [29 April 2004]

Granger, S. (2000). "Emulation as a Digital Preservation Strategy." *D-Lib Magazine*, 6(10). [Online]. Available: http://www.dlib.org/dlib/october00/granger/10granger.html [29 April 2004]

Greenberg, J. (2003). "Metadata Generation: Processes, People and Tools." *Bulletin of the American Society for Information Science & Technology*. December/January 2003. p. 16-19.

Guenther, R. & S. McCallum. (2003). "New Metadata Standards for Digital Resources: MODS and METS." *Bulletin of the American Society for Information Science & Technology*, December/January 2003. p. 12-15.

Hedstrom, M. (2003). *It's About Time: Research Challenges in Digital Archiving and Long-term Preservation*. Final Report of a Workshop on Research Challenges in Digital Archiving and Long-term Preservation, 12-13 April 2002. Sponsored by the National Science Foundation and the Library of Congress. [Online]. Available: http://www.si.umich.edu/digarch/NSF%200915031.pdf [29 April 2004]

Hedstrom, M. & S. Ross. (2003). *Invest to Save: Report and Recommendations of the NSF-DELOS Working Group on Digital Archiving and Preservation*. [Online]. Available: http://delos-noe.iei.pi.cnr.it/activities/internationalforum/Joint-WGs/digitalarchiving/Digitalarchiving.pdf [29 April 2004]

Hendley, T. (1998). *Comparison of Methods & Costs of Digital Preservation*. British Library Research and Innovation Report # 106. [Online]. Available: http://www.ukoln.ac.uk/services/elib/papers/tavistock/hendley/hendley.html [29 April 2004]

Heery, R. et al (2003). "Metadata Schema Registries in the Partially Semantic Web: The CORES Experience." in 2003 Dublin Core Conference: Supporting Communities of Discourse and Practice – Metadata Research & Application, Seattle, WA, 28 September – 2 October 2003. p. 11-18. [Online]. Available: http://www.siderean.com/dc2003/102_Paper29.pdf [29 April 2004]

Hey, T. (2001). "E-Science, Archives and the Grid." Presented at Digital Curation: digital archives, libraries and e-science Seminar, 19 October 2001. York, UK. Sponsored by the Digital Preservation Coalition and the British National Space Centre. [Online]. Available: http://www.dpconline.org/graphics/events/presentations/pdf/tonyhey.pdf [29 April 2004]

Hodge, G. (2000). "Digital Archiving: Bringing Stakeholders and Issues Together: A Report on the ICSTI/ICSU Press Workshop on Digital Archiving." *ICSTI Forum* 33. [Online]. Available: http://www.icsti.org/forum/33/#Hodge [29 April 2004]

Hodge, G. M. (2000). "Best Practices in Digital Archiving: A Life Cycle Approach." *D-Lib Magazine*. Vol. 6, No. 1. Jan. 2000. [Online]. Available: http://www.dlib.org/dlib/january00/01hodge.html [29 April 2004]

Hodge, G. (2002). "Managing S&T Data: Preservation in the Broader Context." Presented at the CODATA/NRF Digital Archiving Workshop, 20-21 May 2002, Pretoria, South Africa.

Hodge, G., J. Ormes & P. Healey. (2003). "Using the NASA Thesaurus to Support the Indexing of Streaming Media." Presented at the Networked Knowledge Organization Systems Workshop, "Building a Meaningful Web: From Traditional Knowledge Organization Systems to New Semantic Tools." May 31, 2003, Houston TX. [Online]. Available: http://www.acm.org/sigir/forum/2003F/jcdl03_soergel.pdf [29 April 2004]

Holdsworth, D. and P. Wheatley. (2001). "Emulation, Preservation and Abstraction." *RLG DigiNews*, 5 (4), Feature #2. [Online]. Available: http://www.rlg.org/preserv/diginews/diginews5-4.html#feature2 [29 April 2004]

Hunter, K. (2002). "Yale-Elsevier Mellon Project." [Online]. Available: http://www.niso.org/presentations/hunter-ppt_01_22_02/index.htm [29 April 2004]

Hunter, K. (2003). "Elsevier and the Royal Library of the Netherlands." Presentation at the American Medical Publishers Association Annual Meeting, Philadelphia, PA, 4 March 2003. Webcast available: http://www.videocast.nih.gov/PastEvents.asp?c=1&s=41 [27 January 2004]

International Federation of Library Associations and Institutions and the International Publishers Association. (2002). "Preserving the Memory of the World in Perpetuity: a joint statement on the archiving and preserving of digital information." June 2002. [Online]. Available: http://www.ifla.org/V/press/ifla-ipa02.htm [29 April 2004]

IBM. (2003a). "Digital Information Archiving System." [Online]. Available: http://www-5.ibm.com/nl/dias/ [29 April 2004]

IBM. (2003b). "Royal Dutch Library Preserves Culture with Content Manager and DB2." [Online]. Available: http://www-5.ibm.com/nl/dias/resource/rdl.pdf [29 April 2004]

IBM. (2003c). "IBM/KB Long term Preservation Study." [Online]. Available: http://www-5.ibm.com/nl/dias/preservation.html [29 April 2004]

IDF. (2003). "Project Announced to Develop DOIs for Scientific Data: German National Library of Science and Technology Joins IDF." Press release. 15 September 2003. [Online]. Available: http://www.doi.org/news/TIBNews.html [29 April 2004]

Inera Inc. (2001). "E-journal Archive DTD Feasibility Study." Prepared for the Harvard University Library, Office of Information Systems E-Journal Archiving Project. p. 62-63. [Online]. Available: http://www.diglib.org/preserve/hadtdfs.pdf [29 April 2004]

Institute for Museum and Library Services. (2001). "A Framework of Guidance for Building Good Digital Collections." 6 November 2001. [Online]. Available: http://www.imls.gov/pubs/forumframework.htm [29 April 2004]

Internet Archive. (2001). "Internet Archive: Building an 'Internet Library'". [Online]. Available: http://www.archive.org [29 April 2004]

International Standards Organization. (2003). "Document Management – Long term Electronic Preservation – Use of PDF (PDF/A)." ISO Working Draft. September 2003. [Online]. Available: http://www.aiim.org/documents/standards/ISO_19005_(E).pdf [29 April 2004]

International Union of Crystallography Committee on Electronic Publishing, Archiving and Dissemination of Information. (2001). "Archive Policy of the IUCr." [Online]. Available: http://journals.iucr.org/services/archivingpolicy.html [29 April 2004]

InterPARES: International Research on Permanent Authentic Records in Electronic Systems. (2002). [Online]. Available: http://www.interpares.org [29 April 2004]

James, C., et al. (2003). "Daylight Theory Manual. Chapter 3: SMILES – A Simplified Chemical Language." [Online]. Available: http://www.daylight.com/dayhtml/doc/theory/theory.smiles.html [29 April 2004]

James, H. et al. (2003). "Feasibility and Requirements Study on Preservation of E-prints." Report Commissioned by the Joint Information Systems Committee (JISC). 29 October 2003. [Online]. Available: http://www.jisc.ac.uk/uploaded_documents/e-prints_report_final.pdf [29 April 2004]

Johnston, L. (2003). "Fedora™ and Repository Implementation at UVa." Presented at the DASER Summit, Cambridge, MA, 21-23 November 2003. [Online]. Available: http://www.lib.virginia.edu/digital/resndev/fedora_at_uva_DASER_files/frame.htm [29 April 2004]

JSTOR. (2002). "JSTOR: The Scholarly Journal Archive." [Online]. Available: http://www.jstor.org [29 April 2004]

JSTOR. (2003). "The Challenge of Digital Preservation and JSTOR's Electronic-Archiving Initiative." [Online]. Available: http://www.jstor.org/about/earchive.html [29 April 2004]

Kenney, A. R. and O. Y. Rieger. (2000). *Moving Theory into Practice: Digital Imaging for Libraries and Archives*. Mountain View, California: Research Libraries Group.

Kenney, A. R., et al. (2000). "Moving Theory into Practice: Digital Imaging Tutorial." [Online]. Available: http://www.library.cornell.edu/preservation/tutorial/preface.html [29 April 2004]

Kneisner, D. (2003). "MPEG-7 and Dublin Core: Mapping Between Them." Presented at the American Society for Information Science & Technology Annual Conference, Long Beach, CA, 19-22 October 2003.

Knobloch, J. (2003). "Disciplinary Case Study 1: Physical Sciences – European Organisation for Nuclear Research (CERN)." Presented at the ERPANET/CODATA International Archiving Workshop on the Selection, Appraisal and Retention of Digital Scientific Data held in Lisbon, Portugal, 15-17 December 2003.

Kresh, D. (2003). "Harnessing the Web: The MINERVA Program at the Library of Congress." Presented at the Joint RLG/JISC Symposium, 25 March 2003. [Online]. Available: http://www.loc.gov/minerva/presentations/diane_rlg.ppt [29 April 2004]

Kunze, J. (2003). "Towards Electronic Persistence Using ARK Identifiers." Proceedings of the 3rd ECDL Workshop on Web Archives, August 2003. [Online]. Available: http://bibnum.bnf.fr/ecdl/2003/proceedings.php?f=kunze [29 April 2004]

Kyong-Ho, L., O. Slattery, R. Lu, X. Tang and V. McCrary. (2002). "The State of the Art and Practice in Digital Preservation." Journal of Research of the National Institute of Standards and Technology. Vol. 107, No. 1, January-February 2002, p. 93-106. [Online] Available: http://nvl.nist.gov/pub/nistpubs/jres/107/1/j71lee.pdf [29 April 2004]

Lariviere, J. (2000). "Guidelines for Legal Deposit Legislation." UNESCO: Paris, 2000. [Online]. Available: http://www.ifla.org/VII/s1/gnl/legaldep1.htm [29 April 2004]

Lavoie, B. (2003). "The Incentives to Preserve Digital Materials: Roles, Scenarios and Economic Decision-Making." OCLC: Columbus, OH, April 2003. [Online]. Available: www.oclc.org/research/projects/digipres/incentives-dp.pdf [29 April 2004].

Library of Congress. (2003a). "Preserving Our Digital Heritage: Plan for the National Digital Information Infrastructure and Preservation Program." Washington, D.C. [Online]. Available: http://www.digitalpreservation.gov/index.php?nav=3&subnav=1 [29 April 2004]

Library of Congress. (2003b). "Collection Policy Statement: Web Site Capture and Archiving." April 2003. [Online]. Available: http://lcweb.loc.gov/acq/devpol/webarchive.html [29 April 2004]

Library of Congress. (2003c). "Call for Proposals: Program Announcement to Support Building a Network of Partners."August 2003. [Online]. Available: http://www.digitalpreservation.gov/index.php?nav=4 [29 April 2004]

LOCKSS. (2003). "LOCKSS: Lots of Copies Keep Stuff Safe." [Online]. Available: http://lockss.stanford.edu/index.html [29 April 2004]

LOCKSS-DOCS, (2002). "Full Project Proposal." [Online]. Available: http://lockss-docs.stanford.edu/lockssproposal.html [29 April 2004]

Lorie, R. (2001). "A Project on Preservation of Digital Data." *RLG DigiNews*, 5(3), Feature # 2. [Online]. Available: http://www.rlg.org/preserv/diginews/diginews5-3.html#1 [29 April 2004]

Lyman, P. (2002). "Archiving the World Wide Web," in *Building a National Strategy for Preservation: Issues in Digital Media Archiving*. Washington, D.C.: Council on Library and Information Resources. [Online]. Available: http://www.clir.org/pubs/reports/pub106/web.html [29 April 2004]

Lynch, C. (2000). "Authenticity and Integrity in the Digital Environment: An Exploratory Analysis of the Central Role of Trust," in *Authenticity in a Digital Environment*. Washington, D.C.: Council on Library and Information Resources. C.T. Cullen, editor. [Online]. Available: http://www.clir.org/pubs/reports/pub92/lynch.html [29 April 2004]

Lynch, C. (2003). "Institutional Repositories: Essential Infrastructure for Scholarship in the Digital Age." *ARL Bimonthly Report,* No. 226, February 2003. [Online]. Available: http://www.arl.org/newsltr/226/ir.html [29 April 2004]

Mahon, B. (2002). "Summary Report. ICSTI/CODATA/ICSU Seminar on Preserving the Record of Science." *Information Services & Use*, 22 (2-3), p. 51-56.

Maly, K. & A. Zubair. (2003). "XML for Technical Reports." Presentation at the CENDI Workshop on *XML in Scientific and Technical Information Management: The Basics*, 30 April 2003. [Online]. Available: http://www.dtic.mil/cendi/presentations/xml_zubair_maly_4_30_03.ppt [29 April 2004]

Mason, H. (2002). "ISO 10303 – STEP: A Key Standard for the Global Market." *ISO Bulletin,* January 2002. [Online]. Available: http://www.iso.ch/iso/en/commcentre/isobulletin/articles/2002/pdf/step02-04.pdf [29 April 2004]

Massanes, J. (2003). Personal communication.

McMahon, B. (1996). "Electronic Publishing in Crystallography," *IFLA Journal*, 22(3), p. 199-205.

Mellor, P. et al. (n.d.) "Migration on Request, a Practical Technique for Preservation." [Online]. Available: http://www.si.umich.edu/CAMILEON/reports/migreq.pdf [29 April 2004]

Messerschmitt, D. (2003). "Opportunities for Libraries in the NSF Cyberinfrastructure Program." *ARL Bimonthly Report # 229*. August 2003. [Online]. Available: http://www.arl.org/newsltr/229/cyber.html [29 April 2004]

Metadata Encoding and Transmission Standard (METS). (2003). [Onli ne]. Available: http://www.loc.gov/standards/mets/ [29 April 2004]

Morgan, C. (2000). "Metadata and Deposit Protocols." Presentation at the *BIC Seminar* 5 July 2000. [Online]. Available: http://bic.org.uk/Cliff%20Morgan [27 January 2004]

Muir, A. (2003). "Copyright and Licensing for Digital Preservation." *Update*. June 2003. [Online]. Available: http://www.cilip.org.uk/update/issues/jun03/article2june.html [29 April 2004]

Muller, E., et al (2003a). "Archiving Workflow Between a Local Repository and the National Archive: Experiences from the DiVA Project." Paper presented at the Web Archives Workshop, European Conference on Digital Libraries, 2003. [Online]. Available: http://publications.uu.se/epcentre/conferences/ecdl2003/archiving_ECDL_2003.pdf [29 April 2004]

Muller, E., et al (2003b). "The DiVA Project: Development of an Electronic Publishing System." *D-lib Magazine*. November 2003. Vol. 9 No. 11. [Online]. Available: http://www.dlib.org/dlib/november03/muller/11muller.html [29 April 2004]

NARA. (2003a). "Electronic Records Requirements Document (RD)." Draft July 31, 2003. Prepared by Integrated Computer Engineering. [Online]. Available: http://www.archives.gov/electronic_records_archives/pdf/requirements.pdf [29 April 2004]

NARA. (2003b). "Electronic Records Archives Concept of Operations (CONOPS). August 25, 2003. Prepared by Integrated Computer Engineering. [Online]. Available: http://www.archives.gov/electronic_records_archives/pdf/concept_of_operations.pdf [29 April 2004]

NARA. (2003c). "Strategic Directions: Appraisal Policy. Appendix 2 – Special Considerations for Selected Types of Records." [Online]. Available: http://www.archives.gov/records_management/initiatives/appraisal.html#appendix_2 [29 April 2004]

NDIIPP. (2003). "Digital Preservation." [Online]. Available: http://www.digitalpreservation.gov/ [29 April 2004]

National Archives (UK). (2003). "Practical Experiences in Digital Preservation, 2003: Conference Report." Kew, 2-4 April 2003. [Online]. Available: http://www.pro.gov.uk/about/preservation/digital/conference/report.htm [29 April 2004]

National Archives of Australia. (1999). "Recordkeeping Metadata Standard for Commonwealth Agencies." Version 1.0. May 1999. [Online]. Available: http://www.naa.gov.au/recordkeeping/control/rkms/rkms_pt1_2.pdf [29 April 2004]

National Library of Australia. (2003a). "Collecting Australian Online Publications." [Online]. Available: http://pandora.nla.gov.au/BSC49.doc [29 April 2004]

National Library of Australia. (2003b). PANDAS Manual. [Online]. Available: http://pandora.nla.gov.au/manual/pandas/index.html [29 April 2004]

National Library of Australia. (2003c). Online Australia Publications: Selection Guidelines for Archiving and Preservation by the National Library of Australia. [Online]. Available: http://pandora.nla.gov.au/selectionguidelines.html [29 April 2004]

National Library of Canada, Electronic Collections Coordinating Group. (1998). "Networked Electronic Publications Policy and Guidelines." [Online]. Available: http://www.nlc-bnc.ca/9/8/index-e.html [29 April 2004]

National Library of Medicine. (2000). "Phase II Report of the Working Group on Permanence of NLM Electronic Publications." Revised October 2000. [Online]. Available: http://www.nlm.nih.gov/pubs/reports/permanence.pdf [29 April 2004]

National Library of Medicine. (2003). "Profiles in Science." [Online]. Available: http://profiles.nlm.nih.gov/ [29 April 2004]

National Science Board. (2002). "Science and Engineering Infrastructure for the 21st Century." (NSB 02-190). Draft dated 4 December 2002. Washington, D.C.: National Science Board. [Online]. Available: www.nsf.gov/nsb/documents/2002/nsb02190/nsb02190.doc [29 April 2004]

National Science Foundation. (2003). "Revolutionizing Science and Engineering through Cyberinfrastructure: Report of the National Science Foundation Blue-Ribbon Advisory Panel on Cyberinfrastructure." Washington, D.C.: National Science Foundation. [Online]. Available: http://www.communitytechnology.org/nsf_ci_report/ [29 April 2004]

OCLC Office of Research. (2003). "Web Characterization." [Online]. Available: http://wcp.oclc.org/ [29 April 2004]
OCLC Digital Archive. (2003). [Online]. Available: http://www.oclc.org/digitalpreservation/ [29 April 2004]
OCLC. (2002). "OCLC Digital Preservation Resources, Digital & Preservation Co-op." [Online]. Available: http://www.oclc.org/digitalpreservation/about/co-op/ [29 April 2004]
OCLC Research. (2003). "PREservation Metadata Working Group II: Implementation Strategies." [Online]. Available: http://www.oclc.org/research/pmwg/ [29 April 2004]
OCLC/RLG Working Group on Preservation Metadata. (2002). "Preservation Metadata and the OAIS Information Model: A Metadata Framework to Support the Preservation of digital Objects." [Online]. Available: http://www.oclc.org/research/projects/pmwg/pm_framework.pdf [29 April 2004]
OpenOffice.org Project. (2003). [Online]. Available: http://www.openoffice.org/white_papers/OOo_project/introduction.html#office_prod [29 April 2004]
PADI. (2003). "Legal Deposit." [Online]. Available: http://www.nla.gov.au/padi/topics/67.html [29 April 2004]
PADI: Preserving Access to Digital Information. (1999). [Online]. Available: http://www.nla.gov.au/padi/ [29 April 2004]
PANDORA. [Online]. Available: http://pandora.nla.gov.au/index.html [27 January 2004].
Payette, S. "The Fedora Project." Presented at the DLF Forum, 17 November 2003. [Online]. Available: http://www.fedora.info/presentations/DLF-Nov2003.ppt [29 April 2004]
Phillips, M. (2003). Personal communication.
Pinfield, S. & H. James. (2003). "The Digital Preservation of E-Prints." *D-Lib Magazine*, Vol. 9, No. 9, September 2003. [Online]. Available: http://www.dlib.org/dlib/september03/pinfield/09pinfield.html [29 April 2004]
Planning Committee of the OCLC/RLG Working Group on Preservation Metadata. (2001). "Preservation Metadata for Digital Objects: A Review of the State of the Art." [Online]. Available: http://www.oclc.org/research/projects/pmwg/presmeta_wp.pdf [29 April 2004]
Pothen, P. (2001). "Digital Curation: Digital Archives, Libraries and E-Science: A Report on an Invitational Seminar." 19 October 2001. York, UK. Sponsored by the Digital Preservation Coalition and the British National Space Centre. [Online]. Available: http://www.dpconline.org/graphics/events/richtext/digital-seminarrepdg.html
[29 April 2004]
Public Library of Science. (2003). [Online]. Available: http://www.publiclibraryofscience.org/ [29 April 2004]
PubMed Central: an Archive of Life Science Journals. (2002). [Online]. Available: http://www.pubmedcentral.gov/ [29 April 2004]
Quenault, H. (2003). Personal communication.
Reich, V. (2003). Personal communication.
Reekie, P. (2003). Personal communication. March 2003.
Research Libraries Group. (2001). "Attributes of a Trusted Digital Repository for Digital Materials: Meeting the Needs for Research Resources." August 2001. [Online]. Available: http://www.rlg.org/longterm/attributes01.pdf [29 April 2004]

Research Libraries Group. (2002a). "Trusted Digital Repositories: Attributes and Responsibilities." [Online]. Available: http://www.rlg.org/longterm/repositories.pdf [29 April 2004]

Research Libraries Group. (2002b)."Open Archival Information System (OAIS) Resources." October 2002. [Online]. Available: http://www.rlg.org/longterm/oais.html [29 April 2004].

Research Libraries Group. (2003). "Task Force on Digital Repository Certification." [Online]. Available: http://www.rlg.org/longterm/certification.html [29 April 2004]

Ross, S. and P. Uhlir. (2003). "International Workshop on the Selection, Appraisal and Retention of Digital Scientific Data: Background Paper." [Online]. Available: http://www.erpanet.org/www/products/lisbon/Lisbon_Codata_ERPANET_backgrounddoc2.pdf [29 April 2004]

Russell, K. & E. Weinberger. (2000). "Cost Elements of Digital Preservation." (Draft). [Online]. Available: http://www.leeds.ac.uk/cedars/documents/CIW01r.html [29 April 2004]

Russon, D. (1999). "Access to Information Now and in the Future." Paper presented at the *World Conference on Science*, Budapest, Hungary, 27 June, 1999. [Online]. Available: http://www.icsti.org/russon-budapest.html [29 April 2004]

Sanett, S. (2002). "Toward Developing a Framework of Cost Elements for Preserving Authentic Electronic Records into Perpetuity." *College & Research Libraries* 63(5), September 2002, p. 388-404.

Sanett, S. (2003). "The Cost to Preserve Authentic Electronic Records in Perpetuity: Comparing Costs across Cost Models and Cost Frameworks." *RLG DigiNews*. Vol. 7, No. 4, 15 August 2003. [Online]. Available: http://www.rlg.org/preserv/diginews/v7_n4_feature2.html [29 April 2004]

Sawyer, D. (2003). Personal communication.

Sequiera, E. (2003). "PubMed Central--Three Years Old and Growing Stronger." *ARL Bimonthly Report*, #228, June 2003. [Online]. Available: http://www.arl.org/newsltr/228/pubmed.html [29 April 2004]

Seville, C. and Weinberger, E. (2000). "Intellectual Property Rights Lessons from the CEDARS Project for Digital Preservation." (Draft) [Online]. Available: http://www.leeds.ac.uk/cedars/colman/CIW03.pdf [29 April 2004]

Sinclair, K. "The VERS Standards." [Online]. Available: From the VERS Toolkit site http://vers.imagineering.com.au/erecord_library/library.htm#voapstandard [29 April 2004]

Smith, A. (2001). "Long Term Archiving of Digital Documents in Physics." Report of the Meeting sponsored by the Working Group on Communication in Physics, International Union of Pure and Applied Physics held in Lyon, France, 5-6 November 2001. [Online]. Available: http://publish.aps.org/IUPAP/ltaddp_report.html [29 April 2004]

Smith, A. (2003). "New-Model Scholarship: How Will It Survive?" CLIR Reports – Publication 114, Washington DC: Council for Library and Information Resources. March 2003. [Online]. Available: http://www.clir.org/pubs/abstract/pub114abst.html [29 April 2004]

Smith, A. (2000). "Authenticity in Perspective," in *Authenticity in a Digital Environment*, Council on Library and Information Resources Report pub92, 2000. [Online]. Available: http://www.clir.org/pubs/reports/pub92/smith.html [29 April 2004]

Smith, M., et al (2003a). "DSpace: An Open Source Dynamic Digital Repository." *D-Lib Magazine*, Vol. 9, No. 1. January 2003. [Online]. Available: http://www.dlib.org/dlib/january03/smith/01smith.html [29 April 2004]

Smith, M. (2003b). "METS: Metadata Encoding & Transmission Standard." Presented at the ASIS&T DASER Summit held Cambridge, MA, November 21-23. [Online .Available: http://www.asis.org/Chapters/neasis/daser/MacKenzie_Smith_presentation.ppt [29 April 2004]

Spedding, V. (2003). "Data Preservation: Great Data, But Will It Last?" Research Information. Spring 2003. [Online]. Available: http://www.researchinformation.info/rispring03data.html [29 April 2004]

Steenbakkers, J. F. (2001). "Setting up a Deposit System for Electronic Publications." NEDLIB Report Series 5. [Online]. Available: http://www.kb.nl/coop/nedlib/results/NEDLIBguidelines.pdf [29 April 2004]

Steenbakkers, J. F.(2002). "Preserving Electronic Publications." *Information Services & Use*. Vol. 22. p. 89-96.

Steinke, T. (2003). Personal communication.

Suber, P. (2003). "How Should We Define Open Access?" *SPARC Open Access Newsletter*, Issue #64. [Online]. Available: https://mx2.arl.org/Lists/SPARC-OANews/Message/96.html [29 April 2004]

Tansley, R., et al. (2003). "The Dspace Institutional Digital Repository System: Current Functionality" Proceedings of the 2003 Joint Conference on Digital Libraries, Rice Univ. Houston, TX 27-31 May 2003. Los Alamitos, CA: IEEE. p. 87-97.

Thibodeau, K. (2002). "Overview of Technological Approaches to Digital Preservation and Challenges in Coming Year: What Does It Mean to Preserve Digital Objects?" in *The State of Digital Preservation: An International Perspective*. Washington, D.C.: Council on Library and Information Resources, April 2002. [Online]. Available: http://www.clir.org/pubs/reports/pub107/thibodeau.html [29 April 2004]

Uhlir, P. (1995). *Preserving Scientific Data on Our Physical Universe*. Washington DC: National Academy Press, 1995. [Online]. Available: http://www.nap.edu/books/030905186X/html/R1.html [29 April 2004]

US Government Printing Office. (1998). "Managing the FDLP Electronic Collection: A Policy and Planning Document." [Online]. Available: http://www.access.gpo.gov/su_docs/fdlp/pubs/ecplan.html [29 April 2004]

University of Virginia Library. (2003). "UVA Library Central Digital Repository." [Online]. Available: http://www.lib.virginia.edu/digital/resndev/repository.html [29 April 2004]

Van der Werf, T. (1999). "Identification, Location and Versioning of Web Resources. URI Discussion Paper." *DONOR Report* (1999).

Van de Werf, T. (2000). *The Deposit System for Electronic Publications: A Process Model*. NEDLIB Report Series; # 6, Koninklijke Bibliotheek: Den Haag, [Online]. Available: http://www.kb.nl/coop/nedlib/results/DSEPprocessmodel.pdf [29 April 2004]

Van Nuys, C. (2003). "The Paradigma Project." *RLG DigiNews*, April 15, 2003, v. 7, #2. [Online]. Available: http://www.rlg.ac.uk/preserv/diginews/diginews7-2.html#2one [29 April 2004]

Vasquez, R. & R. Hammen. (2003). "Persistent Identifier." [Online]. Available: http://www.persistent-identifier.de/?lang=en [29 April 2004].

Victorian Electronic Records Strategy. (2003). [Online]. Available: http://www.prov.vic.gov.au/vers/welcome.htm [29 April 2004]

Waters, D. (2002). "Good Archives Make Good Scholars: Reflections on Recent Steps Toward the Archiving of Digital Information," in *The State of Digital Preservation: An International Perspective*. CLIR, July 2002. [Online]. Available: http://www.clir.org/pubs/abstract/pub107abst.html [29 April 2004]

Waters, D. and J. Garrett. (1996). *Preserving Digital Information: Report of the Task Force on Archiving of Digital Information.* Research Libraries Group. [Online]. Available: http://www.rlg.org/ArchTF/ [29 April 2004]

Waugh, A. (2002). "On the Use of Digital Signatures in the Preservation of Electronic Objects." Presented at the DLM-Forum 2002, Barcelona, Spain, May 2002. p.155. [Online]. Available: http://europa.eu.int/historical_archives/dlm_forum/doc/dlm-proceed2002.pdf [29 April 2004]

Weinberger, E. (2000). "Toward Collection Management Guidance." (Draft) [Online]. Available: http://www.leeds.ac.uk/cedars/colman/CIW02r.html [29 April 2004]

Wellcome Trust. (2004). [Online]. Available: http://www.wellcome.ac.uk/ [29 April 2004]

White, A. et al (2003). "PB Core – the Public Broadcasting Metadata Initiative: Progress Report." in *2003 Dublin Core Conference: Supporting Communities of Discourse and Practice – Metadata Research & Application*, Seattle, WA, 28 September – 2 October, 2003, p. 213-222. [Online]. Available: http://www.siderean.com/dc2003/603_paper81.pdf [29 April 2004]

15.0 Appendix I: Follow Up Discussion Questions

1. Purpose

1.1 What is the purpose of the archive/preservation function?

1.2 Is it a lit archive, partly lit or dark archive? (i.e., ongoing access and re-use or is its main goal to preserve?)

1.3 Is this the only copy archived – are you also archiving paper or microfiche? If so, why? Is another institution also archiving the same item (as far as you know)? If so, why?

1.4 Do you have a diagram of the archive components (technology, metadata, and/or process) that you could share?

2. What is archived?

2.1 What are the raw objects that are being archived?

2.2 Is the whole original object being archived?

2.3 Does the original include hyperlinks to other material? Are these links archived? Is the content of the link archived? If so, what are the criteria for capturing these links?

2.4 How many items does the archive contain?

2.5 Does the archive contain individual items or "collections" of items?

3. Format

3.1 What is the object's native format(s) (Word, PPT, JPEG, etc.)?

3.2 Is the native format retained in the archive or transformed? Both?

3.3 If transformed, to what format?

3.4 Using what software?

3.5 How will this preservation format be handled in the future as software changes?

4. Accessibility

4.1 Is the archive accessible on a routine basis? (if dark archive, bypass these questions)

4.2 Through what interface/software?

4.3 To whom? When? At what cost?

4.4 What are the terms and conditions for use?

4.5 Can the users download the objects to personal files?

4.6 If the material archived requires special software for re-use, how do you plan to maintain accessibility as the software changes?

5. Reference Model

5.1 Is the archive built on the Open Archival Information System Reference Model? If so, is it loosely based, modified slightly or does it adhere strictly to the model?

5.2 Did you use any other systems as the model for your archive?

6. Process

6.1 Who is the creator of the native object?

6.2 What is the information flow between the creator and the archive?

6.3 What is the flow between the archive and future users (the access)?

6.4 Who is involved in the process of archiving and preservation?

7. Metadata

7.1 Is metadata assigned to the object? Before preservation and/or after?

7.2 What metadata format/standard is used? Is it one or more than one? Was it extended or profiled for your system's requirements?

7.3 How is the metadata used in access? To support preservation? To manage the archive?

7.4 Is the metadata assigned to the whole object or to parts of it?

7.5 Is the metadata maintained external to the object, embedded with the object, or both?

7.6 What location identifier that is used (URL, DOI, PURL, etc.)?

7.7 Does the metadata include a content standard (i.e., cataloging rules), including authority lists or controlled terminologies?

8. Technology

8.1 What storage technologies are being used?

8.2 What is the plan for technology migration? How often do you expect this will be necessary?

8.3 Has the archive been through a technology migration cycle?

9. Costs

9.1 What was the start-up cost?

9.2 What are the projected ongoing annual costs for maintaining the operational archive?

10. Policies

10.1 Did you have to establish any particular policies to make the archive work?

10.2 Did you have to address intellectual property concerns and if so how?

10.3 Does this effort reflect specific national or organizational policies, for example, requirements for depositing into an institutional archive or a national library depository?

11. Future Plans

11.1 What are the future plans for your archive?

11.2 Are you prototyping or piloting any enhancements?

11.3 What are the major challenges you are facing with your operational archive? Do you have a specific research agenda?