Preservation of and Permanent Access to Electronic Information Resources

Gail Hodge
Information International Associates, Inc.
312 Walnut Place, Havertown, Pennsylvania 19083
USA

gailhodge@aol.com

ABSTRACT

The rapid growth in the creation and dissemination of electronic information has emphasized the digital environment's speed and ease of dissemination with little regard for its long-term preservation and access. To some extent, electronic libraries, that is those libraries that are moving toward provision of materials in electronic form, have been swept up in this attitude as well. Electronic information includes a variety of object types such as electronic journals, e-books, databases, data sets, reference works, and web sites, which are born digital or which have their primary version in digital form.

But, electronic information is fragile in ways that traditional paper-based information is not. Electronic information is more easily corrupted or altered, intentionally or unintentionally, without the ability to recognize that the corruption has occurred. Digital storage media have unknown life spans. Some formats, such as multimedia, are so closely linked to the software and hardware technologies that they cannot be used outside these proprietary environments. Aggravating this situation is the fact that the time between creation and preservation is shrinking, because technological advances are occurring so quickly.

The Open Archival Information System (OAIS) Reference Model provides a framework for discussing the key areas that impact on digital preservation – the creation of the electronic information, the acquisition of and policies surrounding the archiving of resources, preservation formats, preservation planning including issues of migration versus emulation, and long-term access to the archive's contents.

Many projects, worldwide, have contributed to the growing collection of best practices and standards. The numerous stakeholder groups involved in preservation of electronic resources, including creators (authors), publishers, librarians and archivists, and third-party service providers, are working more closely to build a cohesive and sustainable response to the issues. An issue of continuing stakeholder interest is the economic model(s) that will provide ongoing support to electronic preservation.

Despite the remaining issues, local institutions managing electronic libraries can become involved. They are encouraged to monitor developments and projects in the field, to raise awareness of the need for preservation within their institutions, to consider preservation and long-term access issues when negotiating licenses for electronic resources, and to look for opportunities to begin small projects at the local level.

1.0 BACKGROUND

Major activities have been underway in digital archiving and preservation since the early 1990s. In an effort to quickly focus on the current state of the practice and research, this section provides a definition of key terms that will be used throughout the paper and introduces important projects that are used as examples.

Paper presented at the RTO IMC Lecture Series on "Electronic Information Management", held in Sofia, Bulgaria, 8-10 September 2004, and published in RTO-EN-IMC-002.

1.1 Definition of Terms

Key terms used throughout this paper are defined below. In some cases, these definitions are for consistency within the presentation and are not indicative of general consensus within the community.

Born digital – materials that are created in bits and bytes rather than being digitized from paper or other analog medium

Digital archiving – storing the digital information (e.g., creating an institutional repository or digital archive)

Digital preservation – keeping the bits and bytes safe and unaltered for a long period of time

Digitization – converting materials in non-digital form (analog) such as paper, to digital form

Emulation – running old products by recreating the environment of the old hardware and software without actually using the old hardware and software

Migration – moving a digital product from one version of a program, operating system or hardware environment to another over time

Permanent or Long-term access – the ability to use a preserved object long after its initial preservation

Recapturing – copying the content from the original resource again in order to ensure that changes made to the resources are incorporated in the archival version

Refreshing – moving a digital object to a new instance of the same storage medium, retaining the same operating system and hardware environment

1.2 Outline of Major Projects and Operational Systems

There are several major projects in digital preservation which can serve as examples. (Good sources for information about such projects include the PADI (Preserving Access to Digital Information) Web site from the National Library of Australia (NLA 2003), the joint Web site from PADI and the Digital Preservation Coalition (DPC/PADI 2004), and RLG's *DigiNews* newsletter.) This section briefly describes these major research projects and operational systems, since they are used in the following sections of the paper.

CAMiLEON, (Creative Archiving at Michigan and Leeds: Emulating the Old on the New) a joint project of the University of Michigan and the University of Leeds, conducted analysis and testing to determine if emulation is a viable technical strategy for preservation (CAMiLEON 2001).

Cedars (CURL Exemplars in Digital Archiving) was sponsored by the Joint Information Systems Committee in the UK. It was established to determine the feasibility of distributed digital archives. The first implementation included the three institutions in the Consortium of University Research Libraries. In order to prove scalability, Cedars incorporated several other test sites. This project was completed in 2002 with a series of guideline documents (Cedars n.d.).

DIAS (Digital Information Archive System) was developed by IBM for the Dutch National Library. It is based on the OAIS Reference Model and outcomes of the NEDLIB project of the European Union which was completed in early 2001 (NEDLIB 2001). It is considered to be the first commercially available operational archiving system for electronic journals. It is particularly geared toward the needs of national libraries with

legal deposit responsibilities. The current implementation deals primarily with the deposit and tracking aspects. Major issues such as preservation and long-term access are still being investigated.

Digital Preservation Coalition (DPC) is the umbrella organization for the UK preservation efforts. It has incorporated many of the organizations and lessons learned from projects such as Cedars. Most recently it has been instrumental in promoting the incorporation of non-print materials in legal deposit in the UK (Digital Preservation Coalition 2002).

DSpace is a suite of tools for developing an institutional repository to archive various digital objects. There are several implementations, most notably at the Massachusetts Institute of Technology Libraries. The current focus is on archiving, but there is a planned component for preservation and for the creation of federated repositories.

ERPANET (Electronic Resources Preservation and Access Network) is funded by the European Commission to provide a knowledge base and advice to all sectors on issues of archiving and preservation of electronic resources. ERPANET is best known for its workshops, including those related to archiving various types of digital objects and those related to electronic records (ERPANET 2004).

EVA is a project of the National Library of Finland at the University of Helsinki, which uses a series of automatic tools including robots, harvesters, and metadata creation tools to support its goal of capturing electronic network publications of Finland (Lounamaa and Salonharju 1999).

InterPARES (International Research on Permanent Authentic Records in Electronic Systems) is a global project among archiving institutions, including regional consortia for Asia and Europe. The project's goals are to develop best practices related to the creation, preservation and long-term access to *authentic* electronic records. In its second phase, InterPARES2 is investigating issues such as multimedia and dynamic content preservation (InterPARES n.d.).

Kulturaw3 is a project of the Royal Library of Sweden to capture the cultural heritage that is being published via the Internet (Royal Library n.d.).

JSTOR, originally funded by the Andrew J. Mellon Foundation, is now a non-profit organization that archives back issues of journals for publishers by digitizing them. It is just beginning to deal with current journal issues that are in electronic form (JSTOR 2004).

LOCKSS (Lots of Copies Keep Stuff Safe) is a project of the Stanford University Library, its publishing arm, HighWire Press, and several other libraries to develop a system for redundant archives. Its major contribution is an infrastructure for keeping redundant archives synchronized. There is a special project called LOCKSS-DOCS, which is focused on the preservation of U.S. government documents (LOCKSS n.d.).

NDIIPP (National Digital Information Infrastructure and Preservation Program) is a program at the Library of Congress to develop an infrastructure for digital preservation within the U.S. It is viewed as a joint partnership with content owners ranging from traditional publishers to multimedia providers. In cooperation with the National Science Foundation, NDIIPP has developed a research agenda. In its current phase, it is developing partnerships across the stakeholders and promoting solutions to the research agenda through a series of grant awards.

OCLC Digital Archive is a service of OCLC that grew out of its electronic journals' project. In this service OCLC acts as a trusted third party archive receiving deposits of electronic journals into its repository.

It provides several levels of access (continuous or just in case) and controls access rights so that a library can access only the issues equating to the period for which it had a license (OCLC 2004a).

PANDORA (Preserving and Accessing Networked DOcumentary Resources of Australia), a project of the National Library of Australia, captures the Web-based cultural heritage of Australia. It involves capturing content, creating metadata, and making arrangements with rights holders. A federated approach includes the libraries in all the Australian states. (PANDORA n.d.) PANDAS is the operational system that supports PANDORA.

PREservation Metadata: Implementation Strategies (PREMIS) is a joint project of OCLC, the Research Libraries Group (RLG), and others, to establish a standard metadata element set for preservation and to identify best practices related to implementation (OCLC 2004b).

2.0 A FRAMEWORK FOR ARCHIVING AND PRESERVATION

It is valuable to discuss archiving and preservation within a framework. The framework used in this paper is provided by a reference model, which is used extensively throughout the digital preservation community. The Open Archival Information System Reference Model (CCSDS 2001) provides high level data and functional models and a consistent terminology for discussing preservation. The reference model was originally developed by the Consultative Committee on Space Data Systems (CCSDS) to support the archiving of data among the major space agencies. However, it has become the de facto standard for the development of digital archives. It is used by most major projects including those in Australia, the United Kingdom, the Netherlands, and the United States. The OAIS Reference Model became a formal ISO standard in June 2002.

In its simplest form the OAIS looks like this (Fig. 1):

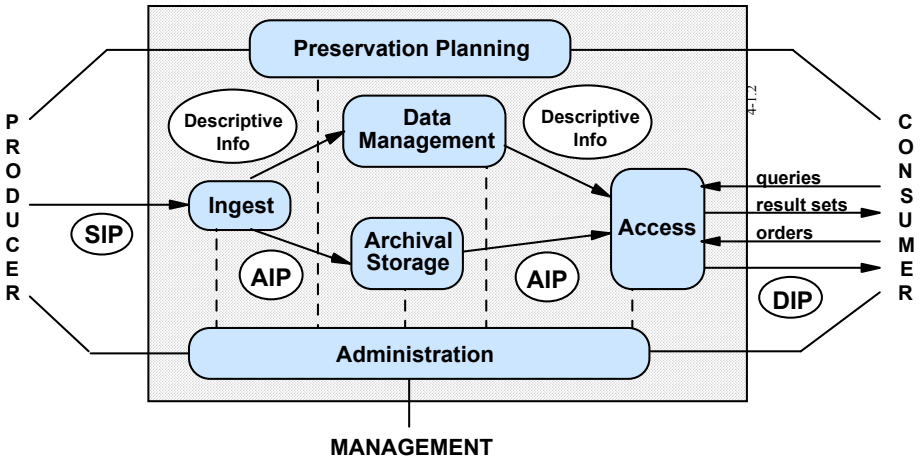

Figure 1: Open Archival Information System.

Source: Consultative Committee on Space Data Systems (used with permission)

SIP – Submission Information Packet (what is submitted or acquired from the producer)
AIP – Archival Information Packet (the object that is archived)
DIP – Dissemination Information Packet (the object that is distributed based on access requests)
Descriptive Info – Metadata

2.1 Production and Creation of Electronic Information

Preservation and permanent access begin outside the purview of the archive with the producer or the creator of the electronic resource. This is where long-term archiving and preservation must begin. Information that is born digital may be lost if the producer is unaware of the importance of preservation. Practices used when electronic information is produced will impact the ease with which the information can be digitally archived and preserved.

Several key practices are emerging involving the producers of electronic information. First, the archiving and preservation process is made more efficient when attention is paid to issues of consistency, format, standardization and metadata description before the material is considered for archiving. By limiting the format and layout of certain types of resources, archiving is made easier. This is, of course, easier for a small institution or a single company to enforce than for a national archive or library. In the latter cases, they are faced with a wide variety of formats that must be ingested, managed and preserved.

In the case of more formally published materials, such as electronic journals, efforts are underway to determine standards that will facilitate archiving, long-term preservation and permanent access. The Andrew J. Mellon Foundation has funded a study of the electronic journal mark-up practices of several publishers. The study concluded that a single SGML document type definition (DTD) or XML schema can be developed to support the archiving of electronic journals from different subject disciplines and from different publishers with some loss of special features (Inera, Inc. 2001). Such standardization is considered key to efficient archiving and preservation of electronic journals by third-party vendors. The DTD developed by PubMed Central for deposit of biomedical journals is being considered as a generalizable model for all journals. 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 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 the case of less formally published material such as web sites, the creator may be involved in assessing the long-term value of the information. In lieu of other assessment factors, the creator's estimate of the long-term value of the information may be a good indication of the value that will be placed on it by members of its designated community or audience in the future. The Preservation Office at the National Library of Medicine has implemented a "permanence rating system" (Byrnes 2000). The rating is based on three factors: integrity, persistent location, and constancy of content. These factors have been combined into a scheme that can be applied to any electronic resource. At the present time, the ratings are being applied to NLM's internal Web sites, and guidelines have been developed to assist creators in assigning the ratings to their sites. This information will be used to manage the ongoing preservation activities and to alert users about a Web site's long-term stability.

Another aspect of the creator's involvement in preservation is the creation of metadata. The best practice is for metadata to be created prior to incorporation into the archive, i.e., at the producer stage. However, most of the metadata continues to be created "by hand" and after-the-fact. Unfortunately, metadata creation is not sufficiently incorporated into the tools for the creation of most objects to rely on the creation process alone. However, as standards groups and vendors move to incorporate XML and other architectures into software products, such as word processors, the creation of metadata should become easier and more automatic.

2.2 Ingest: Acquisition and Collection Development

The first function to be performed by the archive is acquisition and collection development. This is the stage in which the created object is "incorporated" physically or virtually into the archive. In the terminology of the reference model, this is called "Ingest". There are two main aspects to the acquisition of electronic information for archiving – collection policies and gathering procedures.

2.2.1 Collection Policies

Just as in the paper environment, there is more material that could be archived than there are resources with which to accomplish it. Guidelines are needed to tailor the collection policies to the needs of a particular organization and to establish the boundaries in a situation where the responsibility for archiving among the stakeholders is still unregulated. The collection policies answer questions such as what should be archived, what is the extent of a digital object, should the links that point from the object to be archived to other objects also be archived, and how often should the content of an archived site be recaptured?

2.2.1.1 Selecting What to Archive

In the network environment where any individual can be a publisher, the publishing process does not always provide the screening and selection at the manuscript stage on which traditional archiving practices have relied. Therefore, libraries are left with a larger burden of selection responsibility to ensure that publications of lasting cultural and research value are preserved (National Library of Canada 1998).

The scope of NLA's PANDORA (Preserving and Accessing Networked DOcumentary Resources of Australia) Project is to preserve Australian Internet publishing. The NLA has formulated guidelines for the *Selection of Online Australian Publications Intended for Preservation by the National Library of Australia* (NLA n.d.). Scholarly publications of national significance and those of current and long term research value are archived comprehensively. Other items are archived on a selective basis "to provide a broad cultural snapshot of how Australians are using the Internet to disseminate information, express opinions, lobby, and publish their creative work." The National Library of Canada has written similar guidelines (National Library of Canada 1998). The broadest guidelines for Collection Management are provided in a document from the Cedars Project (Weinberger 2000). The most comprehensive analysis of such guidelines is in the *Digital Preservation Handbook*, which is based on the combined lessons learned of all the major projects (Beagrie and Jones 2001).

Even the Internet Archive (Internet Archive n.d.), which considers the capture of the entire contents of the Internet as its mandate, has established limitations. The sites selected do not include those that are "off-limits," because they are behind firewalls, require passwords to access, are hidden within Web-accessible databases, or require payment.

The major lesson from efforts to develop selection guidelines is the importance of creating such a document in order to set the scope, develop a common understanding, and inform the users now and in the future what they can expect from the archive.

2.2.1.2 Determining Extent

Once the site has been selected for inclusion, it is necessary to address the issue of extent. What is the extent or the boundary of a particular digital work, especially when capturing a complex Web site? Is it a "home page" and all the pages underneath it, or are the units to be archived (and cataloged) at a more specific level?

The PANDORA (NLA/PANDORA) project in Australia evaluates both the higher and lower site pages to determine which pages form a cohesive unit for purposes of preservation, cataloging, and long-term access. While preference is given to breaking down large sites into components, the final decisions about extent depend upon which pages cluster together to form a stand-alone unit that conveys valuable information. Each individual component must meet PANDORA's initial selection guidelines.

2.2.1.3 Archiving Links

The extensive use of links in electronic publications raises the question of whether these links and their contents should be archived along with the original site. The answer to this question by any particular project will depend on the purpose of the archiving, the anticipated stability of the links, and the degree to which they contribute to the overall information value of the site.

Most organizations archive the URLs (Uniform Resource Locators) or other identifiers for the links and not the content of the linked pages, citing problems with the instability of links. Some projects have established variants on this approach. For example, PANDORA's decision to archive the content of linked objects is based on its selection guidelines; the content of the linked site is captured only if it meets the same selection criteria as other sites. The National Library of Canada captures the text of a linked object as long as it is on the same server as the object that is being archived, because these intra-server links have proven to be more stable than external links. The American Institute of Physics (AIP) points to the content of a linked reference if it is an item in AIP's archive of publications or supplemental material.

Elsevier cites a technology-related problem as the main reason it does not archive links (Hunter 2002). Elsevier's links are created on the fly, so there is no URL or live page to capture. Similar problems exist when trying to capture pages that are active server pages or those that are created out of a database, portal system, or content management system.

The American Astronomical Society (AAS) has perhaps the most comprehensive approach to the archiving of links. The AAS maintains all links to documents and supporting materials based on collaboration among the various astronomical societies, researchers, universities and government agencies involved in this specific domain. Each organization archives its own publications, retaining all links and access to the full text of all other links. In the future, similar levels of cooperation may be achieved in other subject domains or by publisher collaborations such as CrossRef.

2.2.1.4 Recapturing the Archived Contents

In cases where the site selected for archiving is updated periodically, recapturing the object is necessary. This would be the case for an electronic journal that publishes each article online as it becomes available or for a preprint service that allows the author to modify the content of the preprint as it proceeds through the review process.

When making decisions about recapturing the content of an archived site, a balance must be struck between the completeness and currency of the archive and the burden on the system resources. PANDORA allocates a gathering schedule to each "publication" in its automatic harvesting program. The options include on/off, weekly, monthly, quarterly, half-yearly, every nine months, or annually. The selection is dependent on the degree of change expected and the overall stability of the site. When making decisions about recapturing the content, the EVA Project (Lounamaa and Salonharju 1999) at the University of Helsinki considers the burden on its system resources and the burden of its robots on the sites from which the content would be recaptured.

The National Library of Medicine's Permanence Rating System (see Section 2.1.1) also provides information to support limited recapturing.

2.2.2 Gathering Procedures

There are two general ways in which the archive acquires material. The producer can submit the material to be archived, or the archive can gather the material proactively.

In the first method, the best practices identified in the earlier section on creation become extremely important. Even within an organization, where the producer and the archive are almost one and the same organization, attention to standardization and limitations on the number of formats will have a significant impact on the ease with which submissions can be processed.

In the second approach, the archive may or may not have a formal relationship with the creator or the producer. In this gathering approach, the information to be archived is hand-selected or harvested automatically. In the case of the NLA, sites are identified, reviewed, hand-selected, and monitored for their persistence before being captured for the archive.

In contrast, the Royal Library, the National Library of Sweden, automatically acquires material by running a robot to capture sites for its Kulturaw3 project (Royal Library n.d.). The harvester automatically captures sites from the .se country domain and from foreign sites with material about Sweden, such as travel information or translations of Swedish literature. While the acquisition is automatic, priority is given to periodicals, static documents, and HTML pages. Conferences, usenet groups, ftp archives, and databases are considered lower priority.

2.3 Data Management: Metadata for Preservation

Metadata is needed to preserve the object and for users in the future to find and access it. Metadata supports organization, preservation and long-term access. This section deals with metadata for preservation. Other issues surrounding metadata for description and discovery were covered in the previous lecture on Metadata for Electronic Information Resources.

Archiving and preservation require special metadata elements to track the lineage of a digital object (where it came from and how it has changed over time), to detail its physical characteristics, and to document its behavior in order to reproduce it on future technologies. Each of the major preservation projects – Cedars, PANDORA, NEDLIB, the Harvard Library Project, etc., had its own set of metadata that it considered important for preservation. In 2001, the Research Libraries Group and OCLC reviewed the various sets of preservation metadata and concluded that there was sufficient similarity among the elements that a core set of metadata for preservation could be identified (OCLC/RLG 2001).

In 2001-2002, the Preservation Metadata Working Group developed a draft set of over 20 elements and numerous sub-elements for metadata preservation in the framework of the OAIS Reference Model (OCLC/RLG 2002). OCLC is already using the set as the basis for its Digital Archive and for the work that has been done with the U.S. Government Printing Office. In order to gain consensus on this set and to provide operational and implementation guidance, a follow on group, PREMIS, the PREservation Metadata: Implementation Strategies working group was formed (OCLC 2004b). Two subgroups have been formed, one on the elements and the other on implementation. The implementation group recently concluded a survey of operational preservation systems. The draft element set for preservation metadata and the results of the implementation survey are expected to be published in late 2004.

2.4 Archival Storage: Formats for Preservation

A major issue for the archiving community is which format(s) should be used for archival storage. Should the electronic resource be transformed into a format more conducive to archiving? Is the complexity of an interactive journal necessary or should it be simplified? Should the organization create a dark archive of archival copies in one format and a light archive for dissemination, which might be in a different format? Should consideration be given to the re-use of information and its enhancement or representation in more advanced access technologies in the future? Should the goal be complete replication of the electronic resource or should preservation provide a copy that is "just good enough"? (For example, Cedars has identified the concept of "significant properties," which are properties that are absolutely required in order for a user in the future to get the information value from the resource (Russell 2000).)

Of course the answers to these questions differ by resource type, and there is little standardization at this point. Most electronic journals, reference books, or reports use TIFF image files, PDF, or HTML. TIFF is the most prevalent for those organizations that are involved with conversion of paper issues of journals. For example, JSTOR, a non-profit organization that supports both storage of current journal issues in electronic format and conversion of back issues, processes everything from paper into TIFF and then scans the TIFF image. The OCR, because it cannot achieve 100% accuracy, is used only for searching; the TIFF image is the actual delivery format that the user sees. However, this does not allow embedded references to be active hyperlinks.

SGML (Standard Generalized Mark-up Language) is used by many large publishers after years of converting publication systems from proprietary formats to SGML. The American Astronomical Society (AAS) has a richly encoded SGML format that is used as the archival format from which numerous other formats, including HTML and PDF, are made (Boyce 1997).

For purely electronic documents, Adobe's PDF (Portable Document Format) is the most prevalent format. This provides a replica of the Postscript format of the document, but relies upon proprietary encoding technologies. PDF is used both for formal publications and grey literature. While PDF is increasingly accepted, concerns remain for long-term preservation and it may not be accepted as a legal depository format, because it is a proprietary format. Therefore, Adobe, the Association for Information and Image Management (AIIM) and several other organizations have developed a draft standard for archival PDF, called PDF-A. This provides a file specification for a minimal set of PDF features and functions that will continue to be migrated from one version of PDF to another. The draft is currently in the ISO process.

Preserving the "look and feel" is difficult in the text environment, but it is even more difficult in the multimedia environment, where there is a tightly coupled interplay between software, hardware and content. The University of California at San Diego (UCSD) has developed a model for object-based archiving that allows various levels and types of metadata with separate storage of the multimedia components in systems that are best suited to the component's data type. The UCSD work is funded by the U.S. National Archives and Records Administration and the U.S. Patent and Trademark Office.

2.5 Preservation Planning: Migration and Emulation

Preservation planning is the bridge between the decisions made about archival storage of the bits and bytes and issues of future access and user needs. There is no common agreement on the definition of long-term preservation, but some have defined it as being long enough to be concerned about changes in technology and changes in the user community. This may be as short as 2-10 years.

Two strategies for preservation are migration and emulation. Migration means copying the object to be archived and moving it to newer hardware and software as the technology changes. Migration is, of course, a more viable option if the organization is dealing with well-established commercial software such as Oracle or Microsoft Word. However, even in these cases migration is not guaranteed to work for all data types, and it becomes particularly unreliable if the information product has used sophisticated software features. Unfortunately, this level of standardization and ease of migration is not as readily available among technologies used in fields of study where specialized systems and instruments are used.

Emulation, a strategy that replicates the behavior of old hardware and software on new hardware and software, is being considered as an alternative to migration. There are several types of emulation. Encapsulation would store information about the behavior of the hardware/software with the object. For example, a MS Word 2000 document would be labeled as such and then metadata information would be stored with the object to indicate how to reconstruct the document at the engineering – bits and bytes – level. An alternative to encapsulating the behavior with every instance is to create an emulation registry that uniquely identifies the hardware and software environments and provides information on how to recreate the environment. Each instance would point to the registry (Rothenberg 1999; 2000). Taking emulation a step further is the idea of creating a virtual machine – a new machine that based on the information in the registry could replicate the behavior of the hardware/software of the past (Lorie 2001).

While the best practice for the foreseeable future continues to be migration, machine emulation has been tested with some success by the CAMiLEON Project, a joint project between the University of Michigan and the University of Leeds. However, Granger concludes 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; see also Holdsworth and Wheatley 2001).

2.6 Access

The life cycle functions discussed so far are performed for the purpose of ensuring continuous access to the material in the archive. Successful practices must consider changes to access mechanisms, as well as rights management and security requirements over the long term.

2.6.1 Access Mechanisms

The way in which access is viewed depends on the purpose of the archive, the audiences it will serve and the anticipated needs of those audiences over the long term. For example, national and institutional archives must be concerned with the ability to provide long-term access to the electronic information in a way that virtually replicates the look and behavior of the object today. This is a requirement because of the legal functions served by these archives of record.

Other organizations are interested in how they might actually improve access to current information in the future. A major reason for storing the information related to the U.S. National Library of Medicine's Profiles of Science materials in TIFF and other standardized forms, such as tagged ASCII, is so that the information can be re-purposed or enhanced. Even in its development stage, the project was able to improve the quality of the video clips by converting them to High Definition Video. The belief is that there will always be newer and better technologies, and a goal of the archive is to be able to take advantage of these advances in the future.

2.6.2 Rights Management and Security Requirements

One of the most difficult access issues for digital archiving involves rights management. What rights does the archive have? What rights do various user groups have? What rights has the owner retained? How will the access mechanism interact with the archive's metadata to ensure that these rights are managed properly? How will access rights be updated as the copyright status or security level of the material changes? Numerous groups, including the IEEE, ContentGuard, and MPEG, are developing digital rights management standards including expression languages to support interoperability in e-commerce transactions.

3.0 EMERGING STAKEHOLDER ROLES

A number of stakeholders can be identified including creators/authors, publishers, libraries, archives, Internet service providers, secondary publishers, aggregators, and, of course, users (Haynes, et al. 1997; Carroll & Hodge 1999; Hodge 2000; Hodge & Frangakis 2004). The roles these various stakeholders will play in the archiving process described above remains unclear, but there are several types of electronic information for which some patterns of responsibility are emerging.

In the early stages of the digital age, most electronic journal publishers considered the creation of an electronic archive to be the same as the internal production system. However, many publishers have since come to realize that archiving and production are not one and the same function. In some cases, they are quite antithetical.

The current environment shows a growing understanding of the need for archiving and long-term preservation among the major electronic journal publishers. This may not be the situation with smaller learned society publishers, but that may be more an issue of economics than of desire. The major electronic journal publishers such as Elsevier, Nature and Blackwell have committed to preservation using national libraries or trusted third parties.

Librarians and archivists, particularly those at national libraries, were early advocates of digital preservation issues. Many national libraries spearheaded initiatives and research projects without additional funds and without legislative mandates to cover digital deposit. In most cases, these projects have been instrumental in advancing the research and implementation of operational systems. For example, the Dutch National Library serves as the archive for Elsevier journals.

In addition to new roles for publishers and librarians/archivists, trusted third party archives are emerging. These third parties, such as the OCLC Digital Archive (2004a), JSTOR (2004), PubMed Central (2004), and BioMedCentral see archiving/preservation as an additional business/service opportunity.

A significant outgrowth of the OAIS Reference Model has been RLG's development of attributes of an OAIS-compliant archive (RLG 2001). This check-list can serve to assure a library, publisher or other organization that a particular third-party archive meets minimal requirements for import/export and basic functionality related to the other aspects of an archive.

Another significant development in the emergence of clearer stakeholder responsibilities, particularly for commercially published materials, is a January 2002 announcement on digital preservation by the International Federation of Library Associations and Institutions (IFLA) and the International Publishers Association (IPA). The draft presented for discussion highlights the importance of "born digital" materials and suggests that the appropriate place for preservation of last resort is with the national libraries. It is hoped that

additional legislative/policy efforts and funding for cooperative initiatives will result from this statement and from the inclusion of digital preservation on the agendas of these two major international stakeholder organizations.

4.0 SYSTEMS DEVELOPMENT

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.

4.1 DSpace Institutional Digital Repository System

The DSpace Institutional Digital Repository System began as a joint project of the MIT Libraries and Hewlett-Packard Co. 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). The MIT Libraries' implementation of DSpace defines various levels of support for different input formats. For example, "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, migration, 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.

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.

4.2 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. 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. 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.

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 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 the either of these preservation strategies.

4.3 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. 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 2004a). In principle, it follows the Metadata Encoding and Transmission Standard (METS) 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 also 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.

4.4 PANDORA Digital Archiving System (PANDAS)

The PANDAS (PANDORA Digital Archiving System) has been operational since June 2001 (NLA 2003). 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 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.

4.5 Lots of Copies Keep Stuff Safe (LOCKSS)

LOCKSS (Lots of Copies Keep Stuff Safe) is an automated, decentralized preservation system developed by Stanford University to protect libraries against loss of access to digital materials (LOCKSS n.d.). LOCKSS development is supported by the National Science Foundation, Sun Microsystems, and the Mellon Foundation. LOCKSS software, which is free and open-source, is designed to run as an "Internet appliance" on inexpensive hardware and to require minimal technical administration. LOCKSS has been operational at Stanford for five years and the production version of the software was released in April 2004.

LOCKSS creates low-cost, persistent digital "caches" of authoritative versions of http-delivered e-journal content at institutions that subscribe to that content. LOCKSS uses the caching technology of the web to collect pages of journals as they are published, but unlike normal caches, the cached pages are never flushed. The LOCKSS server runs an enhanced web cache that collects new issues of the e-journal and continually compares its contents with other caches via a peer-to-peer polling system. If damage or corruption is detected in an institution's cache it can be repaired from the publisher or from another cache. LOCKSS safeguards the institution's access to the content while enforcing the publisher's access control systems and, the LOCKSS model generally does not harm the publisher's business model since it is based on the original subscription to the e-journal.

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" (Hodge & Frangakis 2004).

4.6 Fedora™ (Flexible Extensible Digital Object Repository Architecture)

The University of Virginia 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). Similar to DSpace, Fedora is focused currently on repository development and management. However, it will eventually include preservation services.

Fedora 1.0 was released as open source software (Mozilla Public License) in 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. The largest implementation of Fedora is at the University of Virginia Library's Central Digital Repository. A 2001 Mellon Foundation grant allowed for joint development of a production-quality system by Cornell and the University of Virginia. The system currently includes XML objects, text (full text and page images of e-books) and images in multiple resolution (Payette 2003). A number of other institutions and organizations are using or evaluating Fedora, including The British Library, the National Library of Portugal and the Thailand Office of Defense Resources. Fedora is a component of the DSpace architecture.

5.0 TRENDS AND ISSUES

The trend in archiving and preservation has moved from theoretical discussions to pragmatic projects and operational systems. There are more initiatives focused on the realistic details of metadata, selection criteria, technologies and systems for archiving. While the need to raise awareness has not completely disappeared, more time is being spent on partnership development, testing and implementation.

The focus of research and development has shifted to "filling in the gaps." The National Science Foundation (NSF) and the Library of Congress have announced research grants in areas such as extremely large data sets and long-term access to complex multimedia objects. The International Internet Preservation Consortium's Deep Web Working Group is investigating the capture of the dynamic web.

In addition to the trend toward pragmatic initiatives, cooperation has increased among projects and across stakeholder groups. OCLC, the UK's Digital Preservation Coalition (JISC) and RLG have been instrumental in identifying, supporting and advancing key areas of cooperation. As a real sign of maturity, the work is being "divided up". While some projects are developing operational systems, others are working in the background to achieve consensus on standards among/between projects. Unlike many standards activities in the past that have developed from local and regional practices, the work related to digital preservation is starting with the goal of international consensus.

Another key issue for electronic libraries is intellectual property rights. A recent study shows that Canada, Denmark, New Zealand, Norway, South Africa, and the United Kingdom have enacted legislation or have a legislative process in place that covers some form of digital publications (Hodge & Frangakis 2004).

Despite these positive trends, key issues remain. The cost of archiving and the lack of established business models that will sustain long-term preservation may prove to be significant stumbling blocks in the advancement of the cause of preservation. However, even these issues are being addressed in a pragmatic fashion. OCLC, Stanford University Libraries/HighWire Press, JSTOR, and major publishers such as Elsevier are actively dealing with questions of cost and how and who will pay for the archiving. Projects such as the archive of Elsevier material at Yale Library (also funded by the Mellon Foundation) (Hunter 2002) identified practices that can accommodate the access needs of libraries and users while meeting the economic requirements of producers. The development of value-added services that can help to subsidize basic archiving and preservation activities is being considered.

6.0 LOCAL INSTITUTIONAL RESPONSES

Many of the projects highlighted in this paper are national, regional or even global in scale. However, what can a local institution do to ensure the preservation of electronic resources?

First, it is important to be aware of what is going on in this field. What are the outcomes of the major projects? How are standards being developed?

There are several sources for this information. The major projects have extensive web sites, and many like Cedars and NEDLIB have produced numerous publications, which are available from the web sites. Secondly, the PADI site at the NLA (PADI, n.d.) and the joint DPC/PADI *What's New in Digital Preservation* site (Digital Preservation Coalition 2004) are major portals to digital preservation information. The Electronic Resources Preservation and Access Network (ERPANET 2004) provides workshops and reports on these workshops. Newsletters such as *RLG DigiNews* from the Research Libraries Group are excellent sources of up-to-date information about projects. The CODATA Working Group on Digital Data Preservation and the International Council for Scientific and Technical Information are developing a portal for resources related to the preservation of digital data sets.

The local librarian should take every opportunity to raise awareness about the importance of digital preservation at his or her institution. When possible, be proactive in seeking funds to start small projects for preserving digital materials. A concrete way to raise awareness is to ensure that archiving and preservation are

considered when negotiating licenses for electronic resources, such as electronic journals and databases. With many national regimes for deposit of digital materials lagging behind the practical uses of these materials, it is important to address these archiving issues in license agreements. Equally, it is important to try to establish a balance between the rights of the rights holders and those of the library and users.

The major lesson is to think globally but to act locally – scaling the findings of the major global activities to the local needs.

7.0 CONCLUSIONS

A review of the cutting-edge projects shows the beginning of a body of best practices for digital archiving. The early adopters in the area of digital archiving are providing lessons that can be adopted by others in the stakeholder communities. Through the collaborative efforts of the various stakeholder groups – creators, librarians, archivists, funding sources, and publishers – a new tradition of stewardship will be developed to ensure the preservation and continued access to our intellectual heritage.

8.0 REFERENCES

Beagrie, N. and Jones, M. (2001). *Preservation management of digital materials: a handbook*. London: The British Library.

Beck, J. (2003). PubMed Central and the NLM DTDs. *Presented at the ASIS&T DASER Summit, November 12-23, 2003, Cambridge, MA.* [Online]. Available: http://www.asis.org/Chapters/neasis/daser/Jeff_Beck_presentation.ppt [5 July 2004].

Boyce, P. (1997, November). Costs, archiving, and the publishing process in electronic STM journals. *Against the Grain*, *9*(5): 86. [Online]. Available: http://www.aas.org/~pboyce/epubs/atg98a-2.html [21 April 2004].

Byrnes, M. (2000). Assigning permanence levels to NLM's electronic publications. *Presented at Information Infrastructures for Digital Preservation: A One Day Workshop, Dec. 6, 2000, York, England.* [Online]. Available: http://www.rlg.org/events/pres-2000/infopapers.html/byrnes.html [21 April 2004].

CAMiLEON: Creating creative archiving at Michigan & Leeds: Emulating the old on the new. (2001). [Online]. Available: http://www.si.umich.edu/CAMILEON/ [21 April 2004].

CCSDS (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 [21 April 2004].

Cedars: CURL Exemplars in Digital Archives. (n.d.). [Online]. Available: http://www.leeds.ac.uk/cedars/ [21 April 2004].

Carroll, B. and Hodge, G. (1999). Digital electronic archiving: The state of the art, the state of the practice. [Online]. Available: http://www.icsti.org/conferences.html [21 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 [21 June 2004].

Digital Preservation Coalition. (2002). [Online]. Available: http://www.dpconline.org/graphics/index.html [21 April 2004].

Digital Preservation Coalition. (2004) DPC/PADI What's new in digital preservation. [Online]. Available: http://www.dpconline.org/graphics/whatsnew/ [21 June 2004].

ERPANET: Electronic Resource Preservation and Access NETwork. (2004). [Online]. Available: http://www.erpanet.org [21 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 [21 April 2004].

Haynes, D., Streatfield, D., Jowett, T. and Blake, M. (1997). Responsibility for digital archiving and long term access to digital data. JISC/NPO Studies on Preservation of Electronic Materials. [Online]. Available: http://www.ukoln.ac.uk/services/papers/bl/jisc-npo67/digital-preservation.html [4 June 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 [21 April 2004].

Hodge, G. and Frangakis, E. (2004). Digital preservation and permanent access to scientific information: The state of the practice. A joint report by the International Council for Scientific and Technical Information and CENDI. [Online]. Available with free registration: http://www.icsti.org/icsti_reports.html [21 April 2004].

Holdsworth, D. and Wheatley, P. (2001). Emulation, preservation and abstraction. *RLG DigiNews*, 5 (4), Feature #2. [Online]. Available: http://www.rlg.org/preserv/diginews/diginews5-4.html#feature2 [21 April 2004].

Hunter, K. (2002). Yale-Elsevier Mellon Project. [Online]. Availabile: http://www.niso.org/presentations/hunter-ppt_01_22_02/index.htm [21 April 2004].

IBM. (2003a). Digital Information Archiving System. [Online]. Available: http://www-5.ibm.com/nl/dias/ [21 June 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 [21 June 2004].

Inera Inc., (2001). E-journal archive DTD feasibility study. Prepared for the Harvard University Library, Office of Information Systems E-Journal Archiving Project. Pg. 62-63. [Online]. Available: http://www.diglib.org/preserve/hadtdfs.pdf [21 April 2004].

Internet Archive. (n.d.). [Online]. Available: http://www.archive.org [21 April 2004].

InterPARES: International Research on Permanent Authentic Records in Electronic Systems. (n.d.). [Online]. Available: http://www.interpares.org [21 April 2004].

JSTOR: The Scholarly Journal Archive. (2004). [Online]. Available: http://www.jstor.org [21 June 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 [21 June 2004].

LOCKSS. (n.d.) [Online]. Available: http://lockss.stanford.edu/index.html [21 June 2004].

Lorie, R. (2001, June). 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 [21 April 2004].

Lounamaa, K. and Salonharju, I. (1999, January). EVA-the acquisition and archiving of electronic network publications in Finland. *Tietolinja News, 1*. [Online]. Available: http://www.lib.helsinki.fi/tietolinja/0199/evaart.html [21 April 2004].

NEDLIB: Networked European Deposit Library. (2001). [Online]. Available: http://www.konbib.nl/nedlib [21 April 2004].

NLA (n.d.). Selection of online Australian publications intended for preservation by the National Library of Australia. [Online]. Available: http://pandora.nla.gov.au/selectionguidelines.html [21 April 2004].

NLA. (2003). PANDAS Manual. [Online]. Available: http://pandora.nla.gov.au/manual/pandas/index.html [21 June 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 [21 April 2004].

OCLC. (2004a). OCLC Digital Archive. [Online]. Available: http://www.oclc.org/digitalarchive/default.htm [21 June 2004].

OCLC. (2004b). PREMIS (PREservation Metadata: Implementation Strategies. [Online]. Available: http://www.oclc.org/research/projects/pmwg/ [21 June 2004].

OCLC/RLG Working Group on Preservation Metadata. (2002). Preservation metadata and the OAIS information model: A framework to support the preservation of digital objects. [Online]. Available: http://www.oclc.org/research/projects/pmwg/pm_framework.pdf [21 June 2004].

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/pmwg/presmeta_wp.pdf [21 June 2004].

PADI: Preserving Access to Digital Information. (n.d.). [Online]. Available: http://www.nla.gov.au/padi/ [21 April 2004].

PANDORA. (n.d.) [Online]. Available: http://pandora.nla.gov/au/index.html [21 April 2004].

Payette, S. (2003). The Fedora Project. Presented at the DLF Forum, 17 November 2003. [Online]. Available: http://www.fedora.info/presentations/DLF-Nov2003.ppt [21 June 2004].

PubMed Central: a Free Archive of Life Science Journals. (2004). [Online]. Available: http://www.pubmedcentral.nih.gov/ [21 April 2004].

RLG. (2001). Attributes of a trusted digital repository for digital materials: Meeting the needs for research resources. [Online]. Available: http://www.rlg.org/longterm/attributes01.pdf [21 April 2004].

Rothenberg, J. (1999, January). Avoiding technological quicksand: Finding a viable technical foundation for digital preservation. Report to CLIR. [Online]. Available: http://www.clir.org/pubs/reports/rothenberg/contents.html [21 April 2004].

Rothenberg, J. (2000, April). An experiment in using emulation to preserve digital publications. NEDLIB Report Series; 1. [Online]. Available: http://www.kb.nl/coop/nedlib/results/NEDLIBemulation.pdf [21 April 2004].

Royal Library. National Library of Sweden. (n.d.) Kulturaw3 – Heritage Project: Long term preservation of published electronic documents. [Online]. Available: http://www.kb.se/ENG/kbstart.htm [21 April 2004].

Russell, K. (2000). Digital preservation and the Cedars Project experience. Presented at Preservation 2000: An International Conference on the Preservation and Long-Term Accessibility of Digital Materials, York, England, December 7-8, 2000. [Online]. Available: http://www.rlg.org/events/pres-2000/russell.html [21 April 2004].

University of Virginia Library. (2003). UVA Library Central Digital Repository. [Online]. Available: http://www.lib.virginia.edu/digital/resndev/repository.html [21 June 2004].

Weinberger, E. (2000). Toward collection management guidance. (Draft) [Online]. Available: http://www.leeds.ac.uk/cedars/colman/CIW02r.html [21 April 2004].