Preserving Electronic Resources to Warrant Public Access

by

Olivia M.A. Madison
Dean of Library Services
Iowa State University

In the twenty-first century, the first place people should think of, when they need information is the library. Libraries are creative partners in the learning and discovery processes within our universities, communities, and countries. Librarians select, organize, present, and preserve information resources for present and future scholars. In doing so, they are the active stewards of rich collections of knowledge that anticipate user needs and respect a diverse community of ideas.

Access to library collections through electronic means is now a standard and an increasingly important benchmark in evaluating library collections. Today most individual libraries are responsible for building and maintaining two libraries – one is physical and the other virtual – and they represent the fusing of the traditional and the electronic. The same careful planning and resources we place into our physical structures and collections are needed to design, build and maintain our electronic architectures. The virtual brick and mortar of these electronic libraries include their scholarly content and technological infrastructures. Perhaps George Soete stated the task facing us in maintaining these new libraries most succinctly when he wrote, "when we confront the problem of preserving digital information, we confront the very essence of what it will mean to be a library in the 21st century.[1]

Our electronic collections have evolved rapidly over the past three decades from mediated searches in remote indexing services conducted by librarians to individual web access to a rich array of full-text resources housed on complicated networks of commercial and local servers. Particularly with the evolution of full-text digital collections, the challenges to preserve these fragile materials are enormous due to complexities of ownership and technology. For the most part, what we increasingly physically now own for electronic materials in our libraries are CD-ROMs of specialized resources, local databases of unique local materials, and locally digitized materials that fall outside of copyright restrictions. We depend increasingly on commercial content providers, whom we hope will be trusted repositories of scholarly resources and reference tools – repositories that should provide perpetual access to and the necessary preservation of the collections we have purchased through contractual licensing agreements.

In this paper I will discuss briefly the current environment relating to issues of access to and preservation of electronic resources and will follow that with examples of standards under development, examples of some developing and existing preservation models, and examples of recent institutional archiving projects. Through these examples I hope to provide you with a glimpse of the challenging complexities of preserving electronic resources and to underscore that there are no simple and straightforward approaches. Rather we must take advantage of a complex set of circumstances, a wide-variety of partners, and a rapidly growing body of technological advances and new models.

Access and Preservation

Providing access and preserving library materials have long been solitary and essential responsibilities of individual libraries. While there have been cooperative and collabora-

[1] George J. Soete, *Issues and Innovations in Preserving Digital Information*. Transforming Libraries no.l5,
SPEC Kit 228 (Washington: Council on Library Resources, 1996), 7.

tive projects and initiatives, these functions have been long recognized in institutional isolation. During the last decade of the twentieth century these traditional responsibilities changed dramatically. Libraries no longer physically hold some of their most valuable and most used resources, rather they purchase access to materials through often complicated licenses and jointly through complex consortial purchasing arrangements. The importance of a trusted repository (content provider or individual library) has new meaning in this new digital context. We largely purchase access to the scholarly publication record, and we expect the respective content providers (and often the repositories) to ensure perpetual access to that record. There are legal documents that now delineate our respective roles in these new legal arrangements. As a result, we are no longer secure in our assumptions about what we own. The lack of credible and affordable technological systems that preserve the electronic scholarly record have led both the content provider and the purchaser to publish and hold print versions of electronic resources as preservation copies. In fact, many libraries are beginning to count electronic titles as their purchased titles and their corresponding print copies (that may represent an added purchase cost) as the “preservation” copies. Often in these legal arrangements, the fall back for the discontinuance of access to an individual title (whether by cessation of a subscription or closure of the repository) is a CD-Rom. While on one hand, it does represent the content, it does not provide itself broad remote access with potential active links to other live web electronic resources.

Regardless of type of digital repository we represent (whether a university library, an archive, a national bibliographic agency, a not-for-profit agency, a scholarly society or commercial entity), we all face the same preservation challenges – among them being:

Fragile format – Digital materials are especially vulnerable to loss and destruction because they are stored on fragile magnetic and optical media that deteriorates rapidly and can fail suddenly.
Technology obsolescence – Digital materials become unreadable and inaccessible if the playback devices necessary to retrieve information from the media become obsolete or if the software that translates digital information is no longer available.
Legal questions surrounding copying and access – Libraries, archives, and other cultural institutions have limited and uncertain rights to copy digital information for preservation or backup purposes, to reformat information so that it remains accessible by current technology, and to provide public access.[2]

A troubling question facing all of us in our decentralized and increasingly cost-centered environment is whether or not repositories (private, public or commercial) will be willing to bear the continuing costs of refreshing data and upgrading the structure of content files, software, and hardware according to current standards. Just as individual libraries might decide to withdraw older, little used books and serials, publishers may decide to do the same with their electronic serial back files and older monographs because of minimal use coupled with technology costs.

Libraries are now doing business with a global commercial environment rife with business plans geared to expected profits for stockholders; mergers and hostile takeovers resulting in clear monopolistic and non-competitive pricing and access practices; publishing houses going into bankruptcy with the potential loss of continued access to their electronic holdings; numerous entrepreneurial dot.com companies that have ventured into the commercial sector of information providers with risky capital infrastructures; and not-for-profit ventures that begin with grant support but are not economically scalable as that unstable support ends. What does this mean for the mission of libraries to provide continuing access to electronic resources that may or may not be duplicated by preservation print copies?

[2] *LC21 : A Digital Strategy for the Library of Congress.* (Washington, D.C.: National Academy Press, 2000), 106-107.

Moreover, even in terms of what traditionally would be seen as trusted access to government information, that access can change dramatically overnight – as has been seen in the United States since September 11, 2001. Various government agencies and departments have begun to remove what might be considered sensitive or now classified information from governmental Internet sites. In some cases federal and state government agencies are placing their sites behind firewalls with restricted access or are shutting down entire sites. For example, the U.S. Nuclear Regulatory Commission has shut down its web site, which included information about nuclear plant designs. The U.S. Department of Transportation's Office of Pipeline Safety recently restricted access to its National Pipeline Mapping System, which provides locations of natural gas pipelines and where drinking water might be at risk. Even Google, a major Internet search engine, voluntarily has removed copies of sites that its staff believes might pose security threats.

Standards for Digital Preservation

With growing cultural and practical dependence on the electronic medium for access by a wide variety of users across the full spectrum of our communities, there is finally a clear momentum towards creating the necessary standards and policies to meet this need. Also, digitization offers a new format for actual preservation reformatting of books, manuscripts, and other materials to provide access in order to safeguard original materials or to provide for its total replacement due to the loss of original artifacts. The resulting standards and technologies should result in a wide array of tools to meet appropriate needs for access and preservation.

At present there are numerous local, regional, national, and international pilot projects and initiatives underway. Descriptions or updates are provided for three standards that are designed to demonstrate the broad range of standards needed to meet the complexities of digital preservation.

Attributes of a Trusted Digital Repository and OAIS

In March 2000, the Research Libraries Group (RLG) and OCLC began to explore and propose attributes of a digital repository for research organizations that might be seen as a certification program for digital archives or repositories – this was in response to an earlier 1996 RLG report on archiving of digital information.[3] The result of the study is a draft RLG-OCLC report proposing the attributes of a trusted digital repository (such as a library, a national library, or an archive) and a request for public comment. The report defines long-term preservation as having two essential functions: long-term maintenance of a bytestream and access to its contents over time through changing technology.[4] The report assumes that to ensure long-term preservation of digital research and scholarship a deep and decentralized infrastructure will be necessary and the scholarly community will demand an overall climate of trust of the systems, protocols, principles, and organizations that support these infrastructures.

The report proposes the following definition:

A reliable digital repository is one whose mission is to provide long-term access to managed digital resources; that accepts responsibility for the long-term maintenance of digital resources on behalf of its depositors and for the benefit of current and future users; that designs its system(s) in accordance with commonly accepted conventions and standards to ensure the ongoing management, access, and security of materials depos-

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

[4] *Attributes of a Trusted Digital Repository: Meeting the Needs of Research Resources – AN RLG-OCLC Report.* Draft for Public Comment. (Mountain View, CA: RLG, August 2001), l.

ited within it; that establishes methodologies for system evaluation that meet community expectations of trustworthiness; that can be depended upon to carry out its long-term responsibilities to depositors and users openly and explicitly; and whose policies, practices, and performances can be audited and measured.[5]

Then the report discusses potential processes of certification for digital archives that would create an overall climate of trust in repositories' capabilities to preserve digital information. It identifies two models: the audit model and the standards model. The audit model frequently is used within depository systems, often involving governmental agencies or legislative bodies that create guidelines for those individual and national libraries that receive and provide access to their resources. The standards model, on the other hand, is one that most libraries are accustomed to use. Generally libraries follow accepted standards such as ISO interlibrary lending, national and international cataloguing standards, guidelines for the creation of metadata, etc. Peer institutions or agencies then "certify" the product or service by their acceptance and/or use of them.[6]

The report calls for a deep infrastructure that is capable of supporting a distributed digital archive system and suggests consideration be given to the Open Archival Information Systems (OAIS) Reference Model, which was developed by the Consultative Committee for Space Data Systems. This model can and has been applied to both traditional and depository repositories and is being used across disciplines with a range of technical expertise. The OAIS Reference Model provides a unifying set of concepts for an OAIS archive. It calls for an organizational team of people and technological systems that have assumed responsibilities for preserving information and making it available to its community.

OAIS has been successfully used within the NEDLIB Project for developing its deposit system for electronic publications, by the British Library's Digital Storage Project, and in development work at the Library of Congress, the National Archives, Harvard University, Stanford University, RLG and OCLC. The National Library of Australia has used it as a generic model to validate the functions and relationships in its PANDORA Archive, which is discussed later in this paper. Its success relates to its function as a reference model and not a model for system design.[7]

The report delineates the responsibilities of a trusted digital repository (largely based on the OAIS model), which includes:

Negotiates for and accepts appropriate information from information producers and rights holders
Obtains sufficient control of the information provided to support long-term preservation
Follows documented policies and procedures that ensure the information is preserved against all reasonable contingencies and enables the information to be disseminated as authenticated copies of the original
Makes the preserved information available to the designated community

[5] *Ibidem*, 12.
[6] *Ibidem*, 14-15.
[7] *Ibidem*, 23-24.

ICOLC Statement of Current Perspective and Preferred Practices

The International Coalition of Library Consortia (ICOLC) is an informal organization that began meeting in 1997. It includes sixty library consortia in the United States, the United Kingdom, the Netherlands, Canada, Germany, Israel, and Australia thereby representing worldwide 5,000 member libraries. Primarily the Coalition serves higher education institutions by facilitating discussion among its members of issues of common interest. ICOLC conducts meetings to keep its members informed about new electronic information resources and pricing practices of electronic providers and vendors; to provide opportunities to meet with the information provider community and discuss their products; and to engage in a dialog with Coalition members about issues of mutual concern.

In March 1998, ICOLC issued its "Statement of Current Perspective and Preferred Practices for the Selection and Purchase of Electronic Information."[8] The Statement establishes for the first time an international perspective on consortial licensing and purchasing of electronic information by libraries. The document addresses current and future electronic information environment issues such as the increasing expectations of library users in a stable funding environment, fair use, archiving of information, pricing strategies, and electronic information delivery metrics. The preferred practices section covers contract negotiations, pricing, data access and archiving, system platforms, licensing terms, information content and its management, and user authentication.

In terms of archiving, the statement calls for providers to provide "a perpetual license when the consortium purchases the content. Consortia and their member libraries should be allowed to take reasonable steps to archive content that they purchase or lease (to make backup copies)."[9] Furthermore, "when an information provider gives access to data from its Web site (rather than through local mounting of data), the provider should guarantee perpetual availability of the content. This availability need not obligate the provider to realtime access. For example, it may be possible to provide the consortium with copies of data files in an appropriate format, escrowing of data files, or other appropriate means."[10]

Dublin Core Metadata

The Dublin Core Metadata Initiative began in 1995 with an invitational workshop sponsored by OCLC in Dublin, Ohio. Those present began development of a standard core set of bibliographic/access descriptors for electronic files that would typically be embedded in any given file by which the file might be searched for identification and potential retrieval purposes. Literally it consists of searchable and descriptive data describing data that might be embedded in a file or is external to a file. The data can be created by the "author" of the electronic file or by an external creator. Metadata is often seen as another term for a traditional cataloging record and provides a mechanism for straightforward searching across complex textual data files. The Dublin Core elements are title, subject, description, source, language, relation, coverage, creator, publisher, contributor, rights, date, type, format, and identifier. These elements are closely aligned to the elements defined in the IFLA's functional requirements for bibliographic records. Why I bring this particular metadata scheme to your attention is that in early October 2001NISO (the National Information Standards Organization) announced that the Dublin Core element set was approved as an ANSI standard (Z39.85-2001) – thereby giving it a status of the national standard in the United States with international implications.[11]

[8] International Coalition of Library Consortia (ICOLC), *Statement of Current Perspectives and Preferred Practices for the Selection and Purchase of Electronic Information*. Available at: http://www.library.yale.edu/consortia/statement.html

[9] *Ibidem*.

[10] *Ibidem*.

[11] Marilyn Geller E-mail dated 5 October 2001 (10:00:54-0400) that announces the approval of the Dublin Core Metadata Element Set Approved by ANSI.

Examples of Developing and Existing Models

At present there is no single working preservation model and set of technologies that work for all repositories. In fact, given the broad range of missions of and collections in our repositories, it would be naïve to assume a simplistic approach. Rather, as mentioned earlier, we need a broad set of standard models to meet our access and preservation objectives. The two examples provided demonstrate the range of necessary models and represent a number of developing, emerging, and evolving models. One example represents a collaborative and voluntary initiative to preserve access to the content integrity of scholarly journals and the other represents one content provider's rather unique approach towards ensuring perpetual use of electronic materials.

Ensuring Continuing Access through LOCKSS™

LOCKSS represents an exciting collaborative method for ensuring access to e-journals that reflects the traditional decentralized method of institutional responsibility to preserve individual collections. The name is an acronym for "Lots of Copies Keep Stuff Safe™" and is an initiative begun by Stanford University with support from the United States National Science Foundation, the Mellon Foundation, and Sun Microsystems. Beta testing began in April 2001 with a worldwide set of library and publisher participants and with a production version expected in 2002.

The concept behind the LOCKSS system is based on simple rules. Acquire lots of copies. Scatter them around the world so that it is easy to find some of them and hard to find all of them. Lend or copy your copies when other libraries need them. And collaborate only with competent and trusted libraries. The technology requires only basic personal computers at each participating library with basic computing expertise.[12]

The project includes forty-five beta test libraries throughout five continents, and fifty-five publishers have endorsed the beta test. Up-to-date project status is available at: http://lockss.standard.edu/projectstatus.htm. The system is designed to preserve access to published content available through the Internet. It builds upon the traditional concept that individual libraries, through the purchase and preservation of their collections, collaboratively assist in ensuring that there is more than one electronic copy for a set of identified scholarly journals.

The goal of the LOCKSS project is to enable libraries to take custody of the material to which they subscribe--in the same way they do for paper--and to preserve it permanently. In any search, users first access publishers' electronic journal copies. For a variety of reasons, if this copy is not available, the user is directed to a copy located at a LOCKSS site. Through a straightforward sophisticated polling system, the LOCKSS system permanently caches copies of online content – with enough copies to assure continuous access around the world. This will ensure that links and searches by authorized individuals continue to locate published material even if it is no longer available from the publisher. And when a copy of an online journal is lost or damaged, the LOCKSS system will identify this and replace it.[13]

While LOCKSS does not represent a traditional preservation process, for scientists, librarians, and publishers who are concerned that the digital scientific record might disappear, despite all careful managed licenses and best intentions, it does have the capabilities to meet a wide variety of user expectations – such as:

[12] "LOCKSS: Protecting & Preserving Web Documents." In: Sun Microsystem Laboratories, *Ten Years of Impact*. Available at http://research.sun.com/features/tenyears/LOCKSS.httml

[13] *Ibidem*.

Providing future generations of scientists with access to all current literature for research, teaching, and learning.

Ensuring that current and future librarians have an inexpensive, robust mechanism that they control to provide their communities with long-term access to essential literature.

Providing current and future publishers with an assurance that their journals' editorial values and brands will be available only to authorized and authenticated readers. [14]

[14] *Ibidem.*

JSTOR

JSTOR represents a unique model in the electronic archiving environment whose purpose is to provide a retrospective archive of published scholarly journals and, in the future, those journals published only in electronic form. The JSTOR mission is to create a trusted repository of back issues of published scholarly journals. Licensing contracts are signed with participating publishers that permit JSTOR to create digital versions of their journals and give JSTOR "perpetual rights" to these electronic versions. If an individual library chooses to cancel a JSTOR title, JSTOR will provide it with a CD-ROM version of the purchased materials. JSTOR also has an agreement with the Center for Research Libraries (CRL) in Chicago that the center will maintain viable preservation copies of the print journals.

Initially JSTOR began its services as a way to offset binding and storage costs through electronic access. This remains an important benefit, but it also represents a way to ensure Internet access to which users have become accustomed. It appears that research libraries primarily purchase JSTOR titles for the benefits of providing convenient distributed access to journal backfiles. With each year, given agreements with publishers, additional back files of journals are added to this electronic archive. JSTOR currently scans and converts paper issues to electronic forms with a typical "currency" of three to five years in age.

In simple terms, a library continues its ongoing subscription for current journal issues and purchases their electronic retrospective counterparts through the JSTOR electronic file repository. The library then has the option of keeping the current issues in paper form as long as it desires and may never chose to bind or retain those issues duplicated by JSTOR electronic access.

For example, the JSTOR Arts & Sciences collection has 127 journals, which represents approximately 770 volumes that would require approximately 1,200 linear feet if shelved in its paper form. JSTOR estimates a savings for the complete collection for open stack storage to be $125,000, not including ongoing annual costs.[15] Libraries have not been quick to withdraw the duplicated print collections, however that is changing. At Iowa State University, given the preservation safeguards that JSTOR is providing, we are beginning to withdraw selected duplicated print issues due to space constraints. We have involved faculty in our decisions and have asked them to evaluate the viability of the scanned images – particularly for mathematical formulas and graphics – before deciding to withdraw duplicated print copies.

Recent Archiving Projects

While major digital library projects by national libraries are becoming commonplace, they often are characterized by varying approaches. This is largely due to different national definitions of the national record, available technical and staffing resources, and potential partnerships. At the 2000 IFLA General Conferences, several speakers described selected national projects. Two projects are described below, and they were selected because of their differing approaches. In addition, described below is a unique private archiving project for the entire Internet, which demonstrates how one person can make a difference for us all.

PANDORA: Australia's National Collection of Selected Online Publications

The PANDORA Project represents a focused approach by a national library to preserve its electronic national record. In 1996, the National Library began to build an archive of selected online publications that were born digital. In outlining this initiative, it devel-

[15] Kevin M. Guthrie, "Archiving in the Digital Age: There's a Will, But Is There a Way?" Scheduled for publication in *EDUCAUSE Review* in the November/December 2001 issue.

oped a set of selection guidelines and a set of business principles to guide the project and define its objectives. The business model is available at http://pandora.nla.gov.au/pandora/bpm.html. The project's electronic unit is responsible for managing the online publications, which in part includes:

Selection
Liaison with publishers/creators
Quality control and problem solving problems
Cataloguing into the National Bibliographic Database

The selection process has wide ranging implications because it results in a commitment to preserve any given title for future use. To be selected, digital publications should be about Australia or Australians or written by an Australian on a subject of interest to Australia.
Cataloguing requires a persistent naming convention for digital resources in addition to policies and procedures related to metadata for access and future preservation management. Through this project, the Australians are cooperatively working with the CEDARS project in the United Kingdom and others metadata standards development. As of August 2000, there were 652 titles in the PANDORA Archive and approximately 35 titles are selected and archived monthly.[16]

The Royal Swedish Web Archiw3e

The Royal Swedish Web Archiw3e represents a comprehensive approach centered on archiving the entire Swedish web and, as of August 2000, contained sixty-five million items. The Royal Library chose this approach because of the difficulty of determining what information future generations would consider important. The project is economically feasible because computer storage is becoming cheaper, and it is now possible to identify and collect these pages through web snapshots and robot harvesting technologies. In this way, a complete copy of the Swedish web is stored after each snapshot, which takes a couple of months to complete. Access to the archive is through surfing and free-text searching, and low priority is given to more traditional library methods such as cataloging. As of August 2000, unfortunately there was no public access to the archive because of possible copyright infringement.[17]

[16] Cliff Law, "PANDORA – Towards a National Collection of Selected Australian Online Publications." Paper presented at the IFLA General Conference, August 2000. Available at http://www.ifla.org/IV/ifla66/papers/174-157e.htm

[17] Allan Arvidson, Krister Perssson, and Johan Mannerhiem, "The Kiulturarw3 Project – The Royal Swedish Web Archiw3e – An example of "complete" collection of web pages." Paper presented at the 66th IFLA General Conference, August 2000. Available at http://www.ifla.org/IV/ifla66/papers/154-157e.htm

Wayback Machine

The Internet Archive just announced the "opening" of its Wayback Machine, a free Internet tool that enables access to ten billion cataloged web pages (or over 100 terabytes of data) archived during its ongoing sweeps of the Internet dating back to 1996. The archive continues to grow at a rate of 10 terabytes per month. At the unveiling ceremony for the Wayback Machine on October 24, 2001, Brewster Kahle, the founder of the not-for-profit Internet Archive, said, "we created the Internet Archive (http://web.archive.org) because we felt it was critical to preserve a permanent record of this historically significant new medium for the public. To date, the Archive has catalogued over ten billion web pages that might otherwise have been lost, giving us both a record of the origins and evolution of the Internet, as well as snapshots of our society as a whole around the turn of the century."[18]

Since its founding as a permanent collection of digital materials for the public, the Internet Archive has been storing and recording digital material for the public. The Archive has collaborated with the Library of Congress and the Smithsonian Institution and the result is the largest known database in existence. Brewster Kahle founded the Archive to build a digital library and other cultural artifacts in digital form with the purpose of offering permanent and free access to researchers, librarians, and the general public. Recent special collections include a September 11 television and online catalog; archived movies from 1903-1973; and the US 2000 election.[19]

[18] Jack Lynch. E-mail dated 24 October 2001 (19:38:19 EST) that includes an announcement "Internet Archive Launches WayBack Machine."

[19] *Ibidem.*