Models for e-journal archives: future pathways into the past

MARILYN GELLER

Collection Management Librarian

Lesley University Library, Cambridge, Massachusetts, USA

The E-Resources Management Handbook – UKSG

The task of preserving our scholarly history has always been an important component of the mission of libraries. It is not just one particular library or type of library that carries this mission; many libraries serving a broad range of institutions and organizations strive to preserve our history, our heritage and our culture by preserving the intellectual artefacts that represent society. In the last century, concerted work was done to preserve our print history, and these experiences are instructive for how we approach the preservation of our digital history. This article reviews some of the significant lessons of print preservation and discusses how they are being used today in furthering our digital preservation initiatives.

Between the writing and the publishing of this article, an open letter to the scholarly research community, 'Urgent Action Needed to Preserve Scholarly Electronic Journals' (http://www.diglib.org/pubs/waters051015.htm) has been issued and is gathering support from such organizations as the Association of Research Libraries, the Association of College and Research Libraries, the Association for Collections and Technical Services, and the Medical Library Association, and from many prominent library directors.

"Whereof what's past is prologue, what to come,
In yours and my discharge."

(William Shakespeare: *The Tempest*, Act Two, Scene One)

A sampling of mission statements confirms what we intuitively understand about libraries: we take as a given that libraries are charged with collecting, organizing, providing access to and preserving material that supports the values and goals of their larger parent organizations and the social institutions they serve.

'The mission of the Library of Congress', according to its Strategic Plan, 'is to acquire, preserve, and make accessible the world's knowledge for the Congress and for America's use and to maintain a universal collection for future generations.'[1] The British Library states on its web site, 'It is not enough just to preserve knowledge; our task is to enable it to be used now and in the future. When knowledge is used, it adds to the sum of human understanding.'[2] Koninklijke Bibliotheek includes among its responsibilities that 'The KB, as the national library of the Netherlands, is responsible for the preservation, management, retrieval and accessibility of the national cultural heritage in so far as it has been laid down in written, printed or electronic documents.'[3] Of its collections, Harvard University says, 'Harvard's library holdings range from traditional print collections to rapidly expanding inventories of digital resources. It is the work of the Harvard libraries to provide the University's faculty, students, and researchers – now and in the future – with comprehensive access over time to all of these materials.'[4] Oxford University's Bodleian Library has the stated mission 'to maintain and develop collections and services in support of the present

and future teaching and research needs of the University of Oxford, and of the national and international scholarly community. In order to carry out this mission, the Library will always aim to: … (c) preserve the collections for future users …'[5] Boston Public Library proclaims that its mission is 'to preserve and provide access to the historical record of our society and to serve the cultural, educational and informational needs of the people of the City and the Commonwealth.'[6]

How libraries have operationalized this mission to consider, develop plans for and collaborate with each other on the preservation of print materials is instructive for how we should consider, develop plans for and collaborate with each other on the preservation of digital content. The successes of the print preservation effort in the latter half of the twentieth century are lessons for those planning the preservation of digital materials now and in the future. However, while these lessons in successful print preservation can provide some guidance for the digital preservation campaign, there are some things about digital materials that present new and very different preservation challenges.

For example, in print preservation work in the last twenty-five years, a collaborative effort was made to develop a standard that would guide producers in creating 'paper that will last several hundred years, without significant deterioration, under normal use and storage conditions in libraries and archives.'[7] The American National Standard for Permanence of Paper for Printed Library Materials, Z39.48, was originally published in 1984, revised in 1992 and reaffirmed in 2002. This new paper standard allowed publishers to issue their materials in a physical format that would prevent acidic deterioration and correct mistakes of the past going forward. It was prophylactic preservation in lieu of the former model of remedial preservation. The collaborative effort of developing a standard for paper that does not contain damaging acid has allowed publishers to produce reliably preservable materials. In 1995, Sagraves and Welsh wrote that 'staff in the Preservation Department of Northwestern University Library conducted a survey to determine what percentage of new acquisitions were printed on acid-free paper … Of all the new acquisitions tested, 89% were printed on acid-free paper.'[8] This was a dramatic improvement over pre-standard acid-free paper percentages. The lesson here is that collaborative development of technical standards enhances preservation progress.

It is in the aspects of the business model for digital archives that we might think differently from our print experience in order to answer such questions as:

Who pays for the creation and maintenance of the archive?

Who can use the archive and under what circumstances?

How will we know that the archive is doing what it has been built to do?

Different approaches are currently being tested for digital archiving, and this is generally a good thing. In a situation where failure could mean the permanent and unrecoverable loss of a wealth of information, having a variety of archiving options that are based on different technologies and different business models gives us some measure of assurance. Because an assortment of different kinds of digital materials exists today, preservation becomes even more complex. For the purpose of this chapter, we will limit ourselves to electronic journals. Given that the genre is evolving dramatically, it will not be long before we need to think more broadly about a range of file types and applications delivered by authors and their publishers in support of articles in these electronic journals. Today's archiving projects for e-journals focus predominantly on HTML, PDF and a variety of image formats; tomorrow's e-journals will include data-sets, different types of animated and interactive imaging formats, and things at which we can only guess.

One digital archiving model currently in use is the LOCKSS (Lots of Copies Keep Stuff Safe; http://lockss.stanford.edu/) Program which allows individual libraries to pull publishers' content down to local servers, match it with the same content on other library servers for assurance of correctness and completeness and deliver it to end-users with the same look and feel as the content was intended to have when delivered by publishers. Begun in 1999 and with heavy support from the Mellon Foundation and much testing, LOCKSS was released in production mode in the spring of 2004. A major philosophical tenet of the LOCKSS Program is that libraries have always been responsible for archiving our scholarly record, and it is the program's mission to make this easy and inexpensive for libraries in the digital environment. Paralleling the print library universe, and as the name suggests, LOCKSS depends on many libraries safeguarding the same content to assure that the content will always be available. In support of the translation of this concept to the digital world, a report was issued in August 2005 in which the authors

of *A Fresh Look at the Reliability of Long-term Digital Storage* who are associated with LOCKSS discuss a variety of issues related to long-term storage of data in the categories of hardware, software, and human administration of these systems. The authors suggest a 'reliability model of replicated storage systems designed to address long-term storage threats.'[9] Also paralleling the print universe, LOCKSS allows participating libraries to acquire publishers' digital content instead of accessing that content on publishers' web sites.

On the library side, LOCKSS requires the installation of software for gathering and storing data on a simple and commonplace personal computer. An administrative module with a web interface allows the library to configure a web crawler to access publishers' web sites for which they have permissions. The web crawler pulls down and caches the designated pages from the publishers' web sites and then compares these pages with the same pages in other LOCKSS caches at other libraries and at the publishers' sites. This comparison allows for a damaged cache to be repaired and for all LOCKSS caches to be synchronized. Delivery of content for end-users is from the publishers' web sites unless these sites are unavailable, in which case a cached copy is delivered. On the publishers' side, LOCKSS participation requires permissions to be granted both legally and technically to allow the LOCKSS crawler to access the appropriate material. A LOCKSS manifest must be made for each archival unit the publisher makes available, and this manifest lists the top-level URLs from which the crawler starts its collection activity. While the files from publisher sites are stored as the publisher has formatted them, there is a presumption that these file formats will likely become obsolete eventually and will be migrated 'on access'.[10] The LOCKSS environment as described is potentially thousands of separately controlled machines that are the technological equivalent of thousands of separately controlled print library collections. LOCKSS has recently announced that there are 50 library members of the LOCKSS Alliance although the software is installed and working at even more libraries.

Another model currently being tested is Portico (www.portico.org) and this assumes that the infrastructure required to support reliable long-term digital storage and future access as an insurance against loss is massive, labor-intensive and prohibitively expensive for any one library. The quantity of material to be stored is huge and grows rapidly as more publishers and authors produce work in electronic formats. Beyond these reasons is the simple fact that content in a digital environment is simultaneously shareable in ways that physical objects cannot be simultaneously shared. It is no longer necessary for every library to have quick and direct access either through ownership or local partnership to content that supports its patrons. In the print environment, libraries buy and take physical possession of objects. In the digital environment, libraries buy access to content. They may or may not have rights to that content in perpetuity, but they certainly do not have physical control of the content. Many licensing agreements include a clause that suggests there will be continuing access for authorized users to that part of the licensed material which was available during the period that the subscription was active, either from publishers' servers, or from a third party's server, or by supplying electronic files to the licensee. However, many libraries do not want the files or the inherent responsibility for storing and preserving files and migrating formats and would be reckless to confuse rights in perpetuity to actual archiving. Digital archiving is clearly something that libraries need and want, but it is also something that may not be within their individual power to create and maintain. Some publishers may be large enough and technologically advanced enough to take on the archiving responsibility, but for them the question is one of guaranteeing the archive's continuing mission in a changing business environment when, in fact, archiving is not the organizational mission. Other publishers may not be in a position to assure a digital archive of their own materials any more than individual libraries could. So, individual libraries find themselves unable or unwilling to archive large quantities of digital content, and publishers may not be willing or able to do it either, and that leaves the digital archiving community with fewer options. Unlike LOCKSS, this model suggests that we can create a separate third party whose sole responsibility is the creation and maintenance of a digital archive. The benefit is to libraries who fulfill their organizational mission and publishers who make their content more marketable to libraries (as purchaser) and to researchers (as users and contributors).

Portico began in 2002 as a Mellon Foundation-funded initiative of JSTOR (http://www.jstor.org/) and was launched in 2005 under its current name with a mission to build a sustainable electronic archive.

Long-term financial viability, according to Portico's model, is through support from those who benefit, and that includes both publishers and libraries. There is an assumption that governmental and charitable organizations would lend support also. Portico has some guiding principles for its archive that are aimed at assuring the integrity of the published scholarly record. First, like its older sibling JSTOR, Portico will only accept content as published with no post-publication correction and, second, Portico will work to preserve the intellectual content of that published material. To do this, Portico accepts what they call source files which are comprised of all the graphics, text, or other material that encompass an electronic journal. Unlike the LOCKSS model of migration on access, Portico assumes that an initial migration of all files to a normalized standard will take place and that migration will occur over time when file formats begin to approach obsolescence.

This strategy to create normalized files that can be stored and migrated for large-scale archiving and future accessibility has already been developed and tested. The National Library of Medicine's National Center for Biotechnology Information in collaboration with Mulberry Technologies and Inera Inc. has created the Archiving and Interchange DTD to provide a 'common format in which publishers, aggregators, and archives can exchange journal content.'[11] This open standard is prophylactic in that it allows a digital archive to ingest large quantities of content that are already formatted or can be formatted in the same way. If the content is formatted according to the standard, this assures us that we can seek an automated way forward, and automating the process allows us to manage such a massive project. In the same way that we have prepared print documents on paper that is designed to support archiving, that is, non-acidic paper, Portico can now also create digital files in a format that prepares them for their archival future.

Portico's approach relies heavily on new and developing standards. Much of the work that has been done thus far on standards, guidelines and best practices for digital archiving, including the Archiving and Interchange DTD, has its roots in the Reference Model for an Open Archival Information System (OAIS) originally proposed by NASA's Consultative Committee for Space Data Systems 'to provide a framework and common terminology that may be used by Government and Commercial sectors in the request and provision of archive services.'[12] The OAIS Reference Model is the starting point for many international efforts to design and implement digital repositories for an array of different types of organization including libraries. The Cedars (CURL Exemplars in Digital ARchives) Project based in the United Kingdom quickly adopted the OAIS Reference Model as its starting point.[13] NEDLIB, the Networked European Deposit Library, which ran from 1998 through 2000 and included partners from the major national libraries of Europe, used the OAIS Reference model to design the infrastructure for a European deposit library.[14] Another standard currently receiving attention in the archiving community at large and from Portico is the Metadata Encoding and Transmission Standard (METS) which 'provides a flexible mechanism for encoding descriptive, administrative, and structural metadata for a digital library object, and for expressing the complex links between these various forms of metadata.'[15] METS files are like packing slips in that they tell us what pieces are being transmitted and how these pieces fit together. Portico representatives are also active in a variety of other archiving standards initiatives.

The question for the third-party archive is who uses the archive and under what circumstances? Publishers spend a great deal of resources on designing platforms for delivery of their scholarly material that are functional and displayed to maximize usability. An archive that is based on the principle of migration of formats is less concerned about the finer details of display. It would seem that it is in the best interest of libraries and their users also to prefer the publisher's platform for accessing their scholarly digital materials. The archive, then, is our insurance that the content will always be available and is not really meant as the preferred platform for delivery. An archive that is designed for insurance and not for everyday use is generally referred to as a 'dark archive' in opposition to a 'light archive' that is allowed to be accessed and used by a variety of people. To keep an archive permanently dark defeats the purpose of the archive, but there can be a range of events that might trigger the lighting of a dark archive. The definitions of these triggering events are very likely to be a careful balancing act among the participating publishers, the library members supporting the archive and the archival organization itself.

Portico's archive is set up as a dark archive, a form of insurance that it is kept safe until it is needed. Library members will not generally have access to the archive unless some trigger event occurs. A trigger event might be defined as some occurrence of either a business or technological nature that permanently

inhibits a publisher from delivering content. One example might be bankruptcy of the publisher or some other mechanism that causes the publisher to go out of business. When the publisher can no longer deliver content as agreed in a license agreement with libraries, Portico's archive can be used instead, but only by those libraries that previously had rights to access the content. The danger of a dark archive is that vast amounts of data may sit unused for extended periods of time, and it may become corrupted over time without anyone knowing. To counteract this, strong auditing standards will be applied to all of Portico's content. In fact, auditing of an archive goes well beyond monitoring data for corruption. In 2003, the Research Libraries Group (RLG) and National Archives and Records Administration (NARA) created a joint task force to document the process by which any digital repository could be audited and certified. A draft of this Audit Checklist was recently released for comment from the community and touches on every aspect of digital repository work from high level policy decisions and staffing to procedures and technical infrastructure.[16] Portico will participate in the testing of the RLG-NARA Audit Checklist. Portico is an example of a third-party archive which assumes responsibility for both publishers and libraries and stores content in a standard format to be migrated en masse as necessary.

A third model for defining the environment in which to archive our digital assets is represented by the collaboration between Elsevier and Koninklijke Bibliotheek (KB), the National Library of the Netherlands. This project is based on the work done by KB and other national libraries during the NEDLIB project. One outcome of this project was the concept of a 'safe place' for deposit and storage of digital materials, 'an institution separate from the publishing environment, which is committed to digital preservation and possesses the right infrastructure, resources and skills for the task.'[17] Both of these organizations, Elsevier and KB, come to the task with a solid background in digital preservation. For its part, Elsevier has delivered to KB digital copies of all its 1,800 journals and, as its previously published materials are digitized, these too are added to KB's collection by Elsevier. Since beginning this project with Elsevier, KB has been able to negotiate similar arrangements with other publishers including Kluwer Academic, Blackwell, and BioMed Central. To create the appropriate technology environment, KB partnered with IBM.

KB's e-Depot (http://www.kb.nl/dnp/e-depot/dm/dm-en.html) receives files from publishers in an agreed format and with bibliographic metadata, and validates it. This receipt occurs in what e-Depot calls its 'electronic post office'. The content and metadata form the Publisher Submission Package. This is ingested into e-Depot's Batch Builder where the metadata is converted from the publishers' format to KB's format, which is an extended version of Dublin Core. The bibliographic metadata is then stored in a KB database while the content files remain in e-Depot. Content stored in e-Depot is not directly accessible, but it is accessible through KB's online database. Patron rights to use this material are currently restricted to on-site access. In this regard, the archive is not dark.

The Preservation Subsystem that works with e-Depot has three objectives: to identify file formats in danger of becoming inaccessible; to put into action a plan to preserve such files; and to supply the technical metadata that will allow the digital object in danger to be delivered. The system includes two types of technical metadata for each file format to be stored, a View Path and a Preservation Layer Model (PLM). 'A View Path consists of the software (including versionnumbers [sic] and required *patches*) that can be used to view a stored document. The PLM describes the different layers on which that software runs. A possible PLM could be: data format, viewer application, operating system, and hardware platform.'[18] More than one View Path should be available for each format type and either emulation or migration may be used depending on the View Path. In an article published in 2001, Holdsworth and Wheatley defined emulation as 'the re-creation on current hardware of the technical environment required to view and use digital objects from earlier times.'[19] Granger defines migration as 'the process of transferring data from a platform that is in danger of becoming obsolete to a current platform.'[20] Simply stated, do we change the machine that displays the object or do we change the file that packages the object's information? The question at the core of this debate is the nature of what we want to archive. Are we trying to preserve the object or the intellectual content within the object? If it is the object we want to preserve, then we must choose emulation, which requires the building of backward compatible software. If, instead, we want only to preserve the intellectual content, then the more important issue is to normalize files in some way that will allow us to store and manipulate them to assure readability in a future software

and hardware environment or to use a mechanism that will allow us to migrate files at the time of use. The rationale behind e-Depot's multiple View Paths is that each of these options has its advantages and that we should keep all of these options available.

LOCKSS, Portico and e-Depot represent three different models for digital archiving. What they have in common is a high degree of collaboration among the partners to identify needs and, based on those needs, to identify viable methods for fulfilling those needs. Clearly, these groups and others are working on a variety of technical and business aspects of the digital archiving problem in an effort to sort out what needs to be done and how best to do it. For standards, guidelines and best practices to succeed, they must be developed by the broadest possible constituency and with input from all materially affected parties. This is precisely what happened in the case of the permanence of paper standards. We are seeing the movement towards standards in the digital archiving community also.

These three projects approach archiving from very different perspectives and take very different stands on important questions such as who has control of and responsibility for the archive, who can use the archive and under what conditions, how content will be stored, and how file formats that are in danger of becoming inaccessible will be treated to assure viability in the long term. Their differences are both in technical and business areas, and they are significant. LOCKSS, Portico and e-Depot differ on a number of points. LOCKSS is under the complete control of individual libraries; Portico is a separate entity from both publisher and library, and e-Depot is under the jurisdiction of one national library. LOCKSS allows any licensed user to access content; Portico restricts access to all content and keeps it safe for insurance purposes, and e-Depot only allows on-site users to access the archived content. LOCKSS stores files in the format delivered by the publisher while both Portico and e-Depot ingest content in a specified format. Where LOCKSS uses migration on access to assure that obsolete files are still viable, Portico plans to migrate en masse files in any format that are in danger of becoming unreadable and e-Depot chooses to use either migration or emulation when and if necessary.

These three projects only highlight the range of possibilities, but there are many more ways to answer the important questions about digital archives and in fact there are other projects that are currently attempting to help us refine our thinking. Perhaps one of the best places to look on the web to keep abreast of issues and initiatives in digital preservation is the National Library of Australia's PADI (Preserving Access to Digital Information) web site (http://www.nla.gov.au/padi/). For now, there is progress and there is much to do for libraries, publishers and the scholarly world to accept as part of the routines of our jobs that digital preservation will take place and will be available reliably into the future.

References

1. Library of Congress, *Mission of the Library of Congress, Strategic Plan*: http://www.loc.gov/about/history/pdfs/04-08StrategicPlan8-14.pdf (18 April 2006).
2. The British Library, *Explaining our mission and vision:* http://www.bl.uk/about/strategic/explainmissvis.html (18 April 2006).
3. Koninklijke Bibliotheek, *Responsibilities of the Koninklijke Bibliotheek*: http://www.kb.nl/bst/taken-en.html (18 April 2006).
4. Harvard University Library, *About HUL*: http://hul.harvard.edu/about.html (18 April 2006).
5. Bodleian Library, *Mission and objectives*: http://www.bodley.ox.ac.uk/mission.html (18 April 2006).
6. Boston Public Library, *Mission statement*: http://www.bpl.org/general/trustees/mission.htm (18 April 2006).
7. *Permanence of Paper for Publications and Documents in Libraries and Archives*, 1993, Bethesda, Md.: NISO Press, p. 4.
8. Sagraves, B. and Welsh, J., The Acid-Free Paper Pledge Six Years Later. In: *Abbey Newsletter*, 19(4) p. 31, September 1995.

9. Baker, M. *et al., A Fresh Look at the Reliability of Long-term Digital Storage*: http://www.arxiv.org/PS_cache/cs/pdf/0508/0508130.pdf (18 April 2006).
10. Rosenthal, D.S.H. *et al*., *Transparent Format Migration of Preserved Web Content,* 22 Nov 2004: http://arxiv.org/PS_cache/cs/pdf/0411/0411077.pdf (18 April 2006).
11. National Library of Medicine, National Institutes of Health. *Public Domain XML DTD Describes Standard Content Model for Electronic Archiving and Publishing of Journal Articles:* http://www.nlm.nih.gov/news/electronic_archiving.html (18 April 2006).
12. NASA/Science Office of Standards and Technology, *ISO Archiving Standards – New Work Item*, June 10, 1995: http://ssdoo.gsfc.nasa.gov/nost/isoas/nwi.html (18 April 2006).
13. The Cedars Project, *Cedars Guide to The Distributed Digital Archiving Prototype*, March 2002: http://www.leeds.ac.uk/cedars/guideto/cdap/guidetocdap.pdf (18 April 2006).
14. Networked European Deposit Library, NEDLIB Homepage: http://www.kb.nl/coop/nedlib/index.html (18 April 2006).
15. Library of Congress, *METS: an overview and tutorial:* http://www.loc.gov/standards/mets/METSOverview.v2.html (18 April 2006).
16. RLG, *An Audit Checklist for the Certification of Trusted Digital Repositories: Draft For Public Comment*, August 2005: http://www.rlg.org/en/pdfs/rlgnara-repositorieschecklist.pdf (18 April 2006).
17. Adams, G., Partners go Dutch to preserve the minutes of science. In: *Research Information,* 13, September/ October 2004: http://www.reedelsevier.com/media/pdf/3/d/article.pdf (18 April 2006).
18. Koninklijke Bibliotheek, *The Preservation Manager for the e-Depot*: http://www.kb.nl/hrd/dd/dd_onderzoek/preservation_subsystem-en.html (18 April 2006).
19. Holdsworth, D. and Wheatley, P., 'Emulation, Preservation, and Abstraction'. In: *RLG DigiNews*, 5 (4), 5 December 2001: http://www.rlg.org/preserv/diginews/diginews5-4.html (18 April 2006).
20. Granger, S., Emulation as a Digital Preservation Strategy. In: *D-Lib Magazine* 6 (10), October 2000: http://www.dlib.org/dlib/october00/granger/10granger.html (18 April 2006).

■ **Marilyn Geller**
E-mail: marilyn.geller@configuredinfo.com

Correspondence address:
Collection Management Librarian
Lesley University Library
Cambridge
MA 02138, USA

Biographical note

Marilyn Geller began her professional library career as a cataloger at the Tozzer Library at Harvard University, and later was a serials cataloger at the Massachusetts Institute of Technology Libraries. After many years in academic libraries, she went on to work for Readmore, Inc., a wholly owned subsidiary of Blackwell's Information Services, where she was responsible for Internet product development. She also spent several years as an independent consultant and was involved in a variety of projects for subscription agencies, service providers, publishers, non-profit organizations, and libraries. Marilyn is a member of the American Library Association and several ALA divisions including ACRL, ALCTS and LITA, and is also a member of NASIG. She has been a member or chair of many committees within these organizations and presents and writes on a range of topics related to scholarly digital communications.

She is currently Collection Management Librarian for the Lesley University Library in Cambridge, Massachusetts, where her responsibilities include collection development and digital services support.