Archiving Electronic Journals

Research Funded by the Andrew W. Mellon Foundation

Edited, with an Introduction,

by Linda Cantara, Indiana University.

The Digital Library Federation
Council on Library and Information Resources
Washington, DC.
2003

Published by

The Digital Library Federation
Council on Library and Information Resources
1755 Massachusetts Avenue, NW, Suite 500
Washington, DC 20036
http://www.diglib.org/

Introduction

Scholarly research and communication depend upon perpetual access to the published scholarship of the past. Before the advent of electronic journals, research libraries subscribed to printed journals, provided access to, and preserved these bibliographic resources in continual support of the research, teaching, and learning needs of their constituent communities. The introduction of electronic journals has transformed scholarly communication in extraordinary ways — making it possible to disseminate research results more quickly, to provide hyperlinked access to cited publications, and to amplify text with images, audio and video files, datasets and software — but it has also created a dilemma for libraries which now license access to rather than own the journals to which they subscribe. Clearly, a model of collaboration involving scholars, publishers, and librarians is required to ensure that the e-scholarship of today will be accessible to researchers of the future.

The seminal report on digital preservation, *Preserving Digital Information: Report of the Task Force on Archiving of Digital Information*, commissioned by the Commission on Preservation and Access (now the Council on Library and Information Resources) and the Research Libraries Group (RLG) in 1994 and published in 1996, issued the following list of major findings that have served as the guidelines for more recent research:[1]

- The first line of defense against loss of valuable digital information rests with the creators, providers, and owners of digital information.
- Long-term preservation of digital information on a scale adequate for the demands of future research and scholarship will require a deep infrastructure capable of supporting a distributed system of digital archives.
- A critical component of the digital archiving infrastructure is the existence of a sufficient number of trusted organizations capable of storing, migrating, and providing access to digital collections.
- A process of certification for digital archives is needed to create an overall climate of trust about the prospects of preserving digital information.
- Certified digital archives must have the right and duty to exercise an aggressive rescue function as a fail-safe mechanism for preserving valuable digital information that is in jeopardy of destruction, neglect, or abandonment by its current custodian.[2]

Equally influential in the development of digital archiving strategies has been the *Reference Model for an Open Archival Information System (OAIS)*, an initiative of the Consultative Committee for Space Data Systems (CCSDS) which began in 1995.[3] The OAIS Reference Model is the conceptual framework for virtually all international digital

archiving efforts,[4] including the seven e-journal archiving planning projects funded by the Andrew W. Mellon Foundation and reported in this publication.

In October 1999, the Council on Library and Information Resources (CLIR), the Digital Library Federation (DLF), and the Coalition for Networked Information (CNI) convened a group of publishers and librarians to discuss responsibility for archiving the content of electronic journals.[5] A series of meetings led to the publication in May 2000 of the document, "Minimum Criteria for an Archival Repository of Digital Scholarly Journals" (1.2).[6] Soon after, the Andrew W. Mellon Foundation solicited proposals for one-year e-journal archiving planning projects which would incorporate the minimum criteria outlined in this document. Seven institutions were awarded grants for projects carried out from January 2001 through early 2002: the libraries of Cornell University, Harvard University, Massachusetts Institute of Technology (MIT), Stanford University, the University of Pennsylvania, and Yale University, and the New York Public Library (NYPL). Cornell and the NYPL took a subject-based approach, with Cornell addressing issues related to agricultural journals and the NYPL addressing those related to electronic resources in the performing arts. Harvard, Pennsylvania, and Yale took a publisher-based approach: Harvard worked with Blackwell Publishing, the University of Chicago Press, and John Wiley & Sons; Pennsylvania worked with Oxford and Cambridge; and Yale worked with Elsevier Science. MIT investigated the issues presented by "dynamic" e-journals, that is, those in which the content changes frequently,[7] while Stanford focused on the development of tools to facilitate local caching of e-journal content. While the approach of each library was unique, a number of key issues were addressed by all.

Development of sustainable economic and business models

As Brian Levoie recently noted, "preservation objectives must be aligned with the incentives for relevant decision-makers to carry them out."[8] In the case of e-journals, the "relevant decision-makers" include authors, publishers, and librarians. Although the grantees propose several economic models — from charging authors an archiving fee upon publication, to setting up endowments to ensure perpetual funding, to charging publishers for archiving services (charges which would undoubtedly be passed on to subscribers), to charging libraries for access to archived content — no one means of financing digital archiving of e-journals was identified, and in fact, a combination of funding models will most likely be required. Further, whereas smaller publishers have a strong incentive to have their electronic content archived, larger commercial publishers are reluctant to provide potential archives unrestricted access to their electronic content, fearing loss of control over presentation as well as loss of future revenues.[9] On the other hand, libraries are reluctant to delegate e-journal archiving to publishers alone for fear that bankruptcies or mergers or simply a publisher's decision that it is no longer economically beneficial to support a particular journal could result in loss of access to the scholarly record. In addition, as Donald Waters has noted, "the concern about the viability of publisher-based archives is whether the material is in a preservable format and can endure outside the cocoon of the publisher's proprietary system."[10] Nevertheless, although research libraries and their constituents would be the beneficiaries of e-journal archives (and thus, have a strong incentive to archive e-journals), the

grantees almost unanimously acknowledge that the costs of long-term archiving — which are still unknown, given rapid changes in technology — cannot be assumed by individual libraries on behalf of the wider library community.

Identification of what should be archived

The grantees had considerable differences of opinion concerning what should be archived, ranging from the "look and feel" of original e-journal issues to bit-stream-only preservation. Whereas Stanford's LOCKSS project focused on caching Web pages, other grantees outlined protocols for requesting that publishers deposit SGML/XML source files and the document type definitions (DTDs) required to validate them. Also addressed was specific content that should or could be archived as well as the range of file formats anticipated and supportable. In addition, nearly all the reports discuss the need for metadata, both publisher-provided and archive-created, for ingesting, documenting, maintaining, and accessing archived materials.

Guidelines for accessing e-journal archives

One of the most controversial issues addressed by the grantees concerned when and how archived journals might be accessed. Debate over what constitutes a "trigger event," that is, a predefined occurrence that would permit an archive to disseminate content, remained unresolved. Nearly all suggested a JSTOR-like "moving wall"[11] as a potential trigger event, but many publishers were reluctant to agree to permit access until after a resource had no more commercial viability. Equally uncertain was the question of whether an archive should be "dark," that is, one that allows no access for routine scholarly use, or "light," that is, fully accessible.

Recent Developments

When available, each report in this publication is followed by a brief postscript on related activities in-progress since the submission of the final report. Meanwhile, the Mellon Foundation has provided development funding for two projects which take two very different approaches to e-journal archiving, Stanford's LOCKSS project and JSTOR's Electronic-Archiving Initiative.

As outlined in Stanford's report, LOCKSS (Lots of Copies Keep Stuff Safe) uses low-cost tools to crawl the Web to cache "redundant, distributed, decentralized" e-journal presentation files for which a library has a subscription or license. LOCKSS supports the traditional model whereby individual libraries build and maintain local collections of journals, and work is underway to develop a user interface for local collection management of e-journals cached using the LOCKSS system. A LOCKSS Alliance of participating libraries has been formed and the system is currently in beta test mode.[12]

Taking a different approach, the JSTOR Electronic-Archiving Initiative is focusing, among other things, on preservation of publishers' source files. As Eileen Fenton, Executive Director of the Initiative reports:

> As the academic and publishing communities have moved into the twenty-first century with ever-increasing reliance on digital content, the infrastructure for preserving this content has not yet been created. Recognizing that establishing a production-level archiving system is a matter of increasing importance, JSTOR, with support from the Andrew W. Mellon Foundation, has launched the Electronic-Archiving Initiative. Known informally as "E-Archive," the mission of this Initiative is the long-term preservation of and access to electronic scholarly resources. The goal is to develop all of the technical and organizational infrastructure elements necessary to ensure the longevity of important scholarly e-resources. At a practical level this includes developing a business model that can support the ongoing work of the archive; establishing relations with producers of electronic content, with librarians, and with scholars; and developing the technical and content management infrastructure necessary to support a trusted archive of electronic materials.
>
> Currently E-Archive is engaged in collaborative discussions with publishers and libraries and is focused on developing a sustainable business model and a prototype archive. E-Archive has also launched a study of the economic impact increasing reliance on e-journals is having on library periodical operations. This study, which focuses on the non-subscription costs of print versus electronic periodicals, is nearing completion, and the findings are expected to be available for broad distribution in late 2003. [13]

The Mellon Foundation's support for two very different approaches to e-journal archiving is based on acknowledgment that "overlapping and redundant archiving solutions under the control of different organizations with different interests and motives in collecting offer the best hope for preserving digital materials...It would be unwise at the outset to expect that only one approach would be sufficient."[14] Noteworthy e-journal archiving approaches and developments initiated since the submission of the final reports in this publication include:

- In cooperation with IBM Global Services, the Koninklijke Bibliotheek (KB), the National Library of the Netherlands, has developed a large-scale Digital Information Archiving System (DIAS). In August 2002, the KB became the first official digital archive for Elsevier Science e-journals; in May 2003, the KB also signed a long-term digital archiving agreement with Kluwer Academic Publishers.[15]

- In June 2003, the National Library of Medicine (NLM) announced the public domain availability of a Journal Archiving and Interchange Document Type Definition (JAIDTD) for publishing online articles. If widely adopted, the JAIDTD would considerably streamline the process of archiving e-journals.[16]

In related digital preservation activities, work is underway to develop a global digital format registry to provide finer granularity of format typing than the current MIME Media Types registry provides, and to standardize representation information about document formats.[17] In addition, OCLC Research and RLG have formed a new

working group which will build on their previous research to develop recommendations and best practices for implementing preservation metadata. The projected time frame for the PREMIS (PREservation Metadata: Implementation Strategies) working group's activities is twelve months (June 2003-June 2004).[18] And, in December 2002, the United States Congress approved funding for the National Digital Information Infrastructure and Preservation Program (NDIIPP), a collaborative project under the leadership of the Library of Congress to develop an infrastructure for the collection and preservation of digital materials. The first of three calls for proposals was announced in August 2003, for projects to begin in early 2004.[19]

The seven Andrew W. Mellon Foundation e-journal archiving planning project reports in this publication represent a significant body of research upon which future endeavors to ensure long-term access to the electronic scholarly record will build. For their efforts to identify, develop, and test the archival practices and tools that will facilitate long-term preservation of and access to electronic journals, the scholarly community owes many thanks to the seven institutions that carried out the projects, to the Digital Library Federation (DLF) and the Coalition for Networked Information (CNI) for initiating discussion of the issues, and to the Andrew W. Mellon Foundation for providing the funds necessary to accomplish the required research.

Linda Cantara
Indiana University, Bloomington
October 2003

Endnotes

[1] For example, see RLG-OCLC Working Group on Digital Archive Attributes, *Trusted Digital Repositories: Attributes and Responsibilities*, An RLG-OCLC Report (Mountain View, CA: Research Libraries Group, May 2002), online at http://www.rlg.org/longterm/repositories.pdf; and OCLC-RLG Working Group on Preservation Metadata, *Preservation Metadata and the OAIS Information Model: A Metadata Framework to Support the Preservation of Digital Objects* (Dublin, OH: OCLC Online Computer Library, June 2002), online at http://www.oclc.org/research/projects/pmwg/pm_framework.pdf.

[2] John Garrett and Donald Waters, co-chairs, *Preserving Digital Information: Report of the Task Force on Archiving of Digital Information*, The Commission on Preservation and Access and The Research Libraries Group, 1 May 1996, 40. Online at ftp://ftp.rlg.org/pub/archtf/final-report.pdf.

[3] Consultative Committee for Space Data Systems, *Reference Model for an Open Archival Information System (OAIS)*, Blue Book, Issue 1, CCSDS 650.0-B-1/ISO 14721:2002 (January 2002). Online at http://wwwclassic.ccsds.org/documents/pdf/CCSDS-650.0-B-1.pdf. For an overview of the development of the OAIS Reference Model, see http://ssdoo.gsfc.nasa.gov/nost/isoas/overview.html.

[4] For example, see CEDARS (Curl Exemplars in Digital ARchives) at http://www.leeds.ac.uk/cedars/, NEDLIB (Networked European Deposit Library) at http://www.kb.nl/coop/nedlib/, and PADI (Preserving Access to Digital Information) at http://www.nla.gov.au/padi/.

[5] See http://www.diglib.org/preserve/presjour.htm.

[6] Dan Greenstein and Deanna Marcum, "Minimum Criteria for an Archival Repository of Digital Scholarly Journals," Version 1.2 (Washington, DC: Digital Library Federation, 15 May 2000). Online at http://www.diglib.org/preserve/criteria.htm, and included as Appendix I of this publication.

[7] For a discussion of e-journals as "dynamic collections of dynamic entities," see Patsy Baudoin, "Uppity Bits: Coming to Terms with Archiving Dynamic Electronic Journals," *The Serials Librarian* 43:4 (2003), 63-72.

[8] Brian Lavoie, *The Incentives to Preserve Digital Materials: Roles, Scenarios, and Economic Decision-Making*, white paper published electronically by OCLC Research (Dublin, OH: OCLC Online Computer Library, April 2003). Online at http://www.oclc.org/research/projects/digipres/incentives-dp.pdf.

[9] This is a significant issue since the majority of commercial scholarly publications are produced by a very small number of publishers. For example, Maggie Jones of the Joint Information Systems Committee (JISC) recently reported that in 2002, 80 percent of the 5,025 journal titles licensed by JISC/NESLI (National Electronic Site Licensing Initiative) were from six publishers: Elsevier, Blackwells, Springer, Kluwer, Taylor & Francis, and Wiley. See Maggie Jones, *Archiving E-Journals Consultancy: Final Report*, report commissioned by the Joint Information Systems Committee (JISC), October 2003, 12. Online at http://www.jisc.ac.uk/uploaded_documents/ejournalsfinal.pdf .

[10] Donald Waters, "Good Archives Make Good Scholars: Reflections on Recent Steps Toward the Archiving of Digital Information," *The State of Digital Preservation: An International Perspective*, Conference Proceedings, Documentation Abstracts, Institute for Information Science, Washington, DC, 24-25 April 2002, Publication 107 (Washington, DC: Council on Library and Information Resources, July 2002), 86. Online at http://www.clir.org/pubs/reports/pub107/pub107.pdf.

[11] The "moving wall" is "the time period between the last issue available in JSTOR and the most recently published issue of a journal." See "JSTOR: The Moving Wall" at http://www.jstor.org/about/movingwall.html; see also, Roger C. Schonfeld, *JSTOR: A History* (Princeton and Oxford: Princeton UP, 2003), 134-138.

[12] For a discussion of the philosophical underpinnings of the LOCKSS model, see Michael A. Keller, Victoria A. Reich, and Andrew C. Herkovic, "What is a Library Anymore, Anyway?," *First Monday* 8:5 (May 2003). Online at http://firstmonday.org/issues/issue8_5/keller/index.html.

[13] Email correspondence from Eileen Fenton to author, 20 October 2003. See also "JSTOR: The Challenge of Digital Preservation and JSTOR's Electronic-Archiving Initiative" at http://www.jstor.org/about/earchive.html.

[14] Waters, 89.

[15] For more information, see Anne Katrien Amse, "Safeguarding the Historic Resources of the Future: Digital Archiving at the Dutch National Library," Parallel Session 3: Historical Resources of the Future, Bibliopolis Conference: The Future History of the Book, 7-8 November 2002, The Hague (Netherlands), Koninklijke Bibliotheek, online at http://www.kb.nl/coop/bibliopoliscongres/amse.html; Johan F. Steenbakkers, "Permanent Archiving of Electronic Publications," *Serials* 16:1 (March 2003), 33-36; Koninklijke Bibliotheek, "National Library of the Netherlands and Kluwer Academic Publishers Agree on Long-Term Digital Archiving," (19 May 2003), online at http://www.kb.nl/kb/resources/frameset_kb.html?/kb/pr/pers/pers2003/kb-kap-en.html; and the IBM/KB Long-term Preservation Study Reports Series at http://www-5.ibm.com/nl/dias/preservation.html.

[16] For more information, see the Postscript to Harvard University's report in this publication.

[17] See Stephen L. Abrams and David Seaman, "Towards a Global Digital Format Registry," Meeting 165, Information Technology and Preservation and Conservation Workshop, World Library and Information Congress: 69th IFLA General Conference and Council, 1-9 August 2003, Berlin, Germany. Online at http://www.ifla.org/IV/ifla69/papers/128e-Abrams_Seaman.pdf.

[18] See http://www.oclc.org/research/projects/pmwg/.

[19] See http://www.digitalpreservation.gov/.