Chapter 7

Special Topics

Electronic Journals

The long-term preservation of electronic journal content is a topic of great concern to academic and research libraries, many of which have been reluctant to invest in all-electronic subscriptions without a believable guarantee that the content will be responsibly archived. This is a matter of no small consequence: ARL statistics for 2003–2004 show member libraries spent $301,699,645, or roughly 31 percent of their total materials expenditures, to license electronic materials, almost exclusively journals.

The preservation of electronic scholarly journals presents special organizational and technical challenges. Organizationally, the normal model for delivery of electronic journals, where a for-profit publisher hosts the content in a central system and sells subscription access to it, means that libraries, the parties with the most stake in preserving the content for future generations, neither own nor physically control it. Preservation requires the negotiation of agreements between the parties defining rights and responsibilities on both sides, including conditions for access and limitations on access.

Technically, the source for online journals can be anything from images of printed pages to highly marked-up text in XML or SGML. Different journals, even from the same publisher, can have different markup formats. Receiving preservation systems must control and relate content at article, issue, volume, and title levels. Quality control is a big issue, as issues often fail to meet the publisher's own documentation standards. Many items of content, from masthead information to advertisements, are now dynamically provided by the online delivery system and may not be part of the journal source files at all.

Early studies suggested that publishers would not be trusted by the library community to provide long-term preservation and access capability for their own content. At the same time, many publishers were finding that the ability to promise long-term preservation and access for their content would constitute a significant competitive advantage. In 2000 the Andrew W. Mellon Foundation funded seven major research libraries to plan and pilot the development of e-journal archives and address business models for sustainability. One major conclusion from these studies was that the costs of archiving could not be assumed by individual libraries on behalf of the wider library community.[1] Mellon subsequently focused its funding on two projects with quite different approaches, LOCKSS development at Stanford, and the JSTOR Electronic-Archiving Initiative (E-Archive). LOCKSS is discussed in Chapter 6. The JSTOR E-Archive project ultimately evolved into Portico.

The Portico archive is a centralized preservation repository of scholarly electronic journals. Unlike LOCKSS, which harvests Web content and so stores access versions by definition, Portico technology is based on acquiring the publishers' source files and archiving both original and normalized versions. Normalization occurs at the time of ingest, and involves converting from the publishers' proprietary SGML and/or XML formats to the National Library of Medicine's Journal Archiving and Interchange DTD, a de facto standard.[2] Portico will also perform forward migration as file formats threaten to become obsolete. Because its preservation strategy is based on normalization and migration, Portico promises to preserve intellectual content only, not necessarily the original look and feel of the journals. However, much attention is paid to the integrity of the scholarly record, and Portico works directly with individual publishers to ensure the appropriate content is archived.

Portico is supported by a combination of grant funding, publisher fees, and library subscription fees. In return for an annual support payment based on the total materi-

als expenditure of the library, the institution obtains campus-wide access to archived journals when certain trigger events occur, such as when the publications are no longer available from the publisher or other source.

Portico Web site
www.portico.org

Globally, national libraries are investing in means to preserve their national journal output. Electronic journals are included in Australia's PANDORA archive, Germany's kopal project, the Koninklijke Bibliotheek's e-Depot, and other initiatives worldwide.

Readings

- Evan Owens, "Automated Workflow for the Ingest and Preservation of Electronic Journals," www.portico.org/news/Archiving2006-Owens.pdf
- Amy Kirchhoff and Eileen Fenton, "Archiving Electronic Journals: An Overview of Portico's Approach," Portico Papers, 2006, www.portico.org/news/PapersFromPortico.1.Overview.pdf.
- Dale Flecker, "Preserving Digital Periodicals," in *Building a National Strategy for Preservation: Issues in Digital Media Archiving,* CLIR pub 106, 2002, http://www.clir.org/pubs/abstract/pub106abst.html. A clear exposition of the difficult issues related to archiving electronic journals.
- Anne R. Kenney, et al., *E-Journal Archiving Metes and Bounds: A Survey of the Landscape,* CLIR pub 138, 2006, www.clir.org/pubs/abstract/pub138abst.html. A review and comparison of journal archiving programs.

Records and Archives

Preservation of digital records has not been emphasized in this report so far, but internationally archivists have been at the forefront of preservation theory and practice. To archivists, a record is a bit of recorded information created or received by an organization in the course of doing business. The connection between the record and the organizational activity is key, and the implicit or explicit documentation of relationships demonstrating this connection is as important to preserve as the content of the record. The evidentiary value of records is that they document the activities which produced them. For archivists and records managers, maintaining the demonstrable authenticity, reliability, and accuracy of electronic records is paramount.

Archivists emphasize the difference between the digital bitstream and the digital record, which requires software to render for use. "Electronic records are stored in forms that differ substantially from those in which they can serve their intended purpose as records."[3] Records can be used only as their bitstreams are interpreted through some rendering process. Andrew Wilson, of the National Archives of Australia, calls this rendering a "performance," and notes that it is the performance, not the bitstream itself, which must be preserved. For this reason, archivists have been quick to abandon the stored form of the record and preservation strategies aimed at keeping original data intact. They have concentrated instead on defining significant properties of various forms of records, developing transformations which preserve these properties, and establishing metadata and procedures which together ensure the authenticity of the record through these transformations.

InterPARES (International Research on Permanent Authentic Records in Electronic Systems) is an international, multidisciplinary research project involving participants in twenty-one countries. It is arguably the longest and most influential digital preservation initiative ever undertaken. InterPARES 1 (1999–2001) addressed methods for maintaining the authenticity of digital records in administrative and legal databases and record-keeping systems. Deliverables included two sets of requirements for assessing and maintaining the authenticity of digital records, one intended for the creators of records creators, and one for preservers. InterPARES 2 (2002–2006) expanded the focus to include reliability and accuracy in addition to authenticity, and to records produced by governmental, scientific, and artistic activities. InterPARES 3 (2007–2012) is intended to translate earlier research into concrete action plans that archives with limited resources will be able to implement.

Perhaps the most mature digital record keeping initiative is the Victorian Electronic Records Strategy (VERS) begun in 1995 by the Public Record Office Victoria (PROV). Because the PROV receives records from state agencies throughout Victoria, VERS has developed specifications for the functions that local agency record-keeping systems must support, and mechanisms to be used to reliably export records from agencies to the PROV. The VERS preservation approach is based on normalizing records at the time of accession to one of several long-term preservation formats, depending on the format of the original record. To ensure that context, authenticity, and reliability are maintained, VERS specifies a standard for metadata which must be provided, and encapsulates the metadata with the record content in a single object, called the VERS Encapsulated Object (VEO).

The National Archives of Australia has also adopted a digital preservation archive based on normalization. The Archives developed an open-source tool called Xena that

encapsulates the original file in XML and creates a second, normalized version in an open format (see Chapter 4).

In the United States, the National Archives and Records Administration (NARA) is taking a slightly different approach in building the Electronic Records Archives (ERA), a system to capture, preserve, and provide access to digital federal records. In late 2003 NARA issued an RFP and requirements document, following which two finalists were selected to participate in a one-year design competition. In September 2006 Lockheed-Martin was awarded a six-year development contract worth $308 million. Progress can be followed on the ERA Web site.

ERA Web site
www.archives.gov/era

Readings

- "The Long-Term Preservation of Authentic Electronic Records: Findings of the InterPARES Project" [n.d.], www.interpares.org/book/index.cfm. Report of InterPARES 1.
- "A Framework of Principles for the Development of Policies, Strategies and Standards for the Long-Term Preservation of Digital Records," version 1, Aug. 1, 2007, www.interpares.org/ip2/display_file.cfm?doc=ip2(pub)policy_framework_document.pdf. One of several important deliverables from InterPARES 2.

Web Harvesting

The idea of preserving the Web has taken on almost mythical proportions because of the vast size of the Web and the notorious ephemerality of Web content. Even apart from spatial and temporal concerns, Web archiving presents a number of thorny technical challenges. With blogs, RSS feeds, and easy authoring tools provided as "Web 2.0" technologies, Web sites are updated with increasing frequency and share with databases the difficulty of selecting the appropriate frequency for capture. The interconnectedness of the Web means every harvest involves a decision about how broadly and how deeply to pursue links. Many Web sites or parts of sites are excluded from crawling by requiring password or IP authorization or by convention such as robots.txt files. Dynamically generated Web pages returned in response to queries are a technical problem for harvesters, as are pages returned through client-side JavaScript links.

Despite, or perhaps because of, these challenges, the community that specializes in Web archiving has pretty much settled on a common set of tools and approaches. Web pages are downloaded using a web crawler (spider) similar to those used by Internet search engines. By far the most popular tool for archiving is Heritrix, an open-source crawler developed by the Internet Archive. Capture results are written in records in the ARC or WARC formats. ARC (not an acronym) is a simple format developed for the Internet Archive, while WARC (Web ARC), is an updated, expanded version still under development. The archived content is indexed with a tool such as NutchWAX, an extension of the open-source Web search engine Nutch customized to work with ARC files. Some archives are cataloged and some are not. End-user access can be provided by a number of tools including the Internet Archive's Wayback Machine or WERA (Web ARchive Access), a newer tool that supports full-text searching.

Within this general technical framework there is much variation in the scope and focus of archiving and in how the addresses to archive are obtained. Most Web archiving is either site-centric, topic-centric, or domain-centric. Site-centric archiving is most commonly done by an organization to preserve its own Web site(s). Site-centric archives are narrowly focused but comprehensive, because the URLs to pursue are known in advance and access is granted to any restricted content. Internal links are followed vertically as many levels deep as necessary to capture the entire site, but external links may be not be pursued. Topic-centric archives focus on a particular subject (e.g., medical Web sites or Chinese studies) or event (e.g., an election or Hurricane Katrina), while domain-centric archives focus on a particular location and are commonly specific to a country, state, or territory. In general, topic- and domain-centric archives tend to be broader but less deep than site-centric archives. A new model for Web archiving is emerging for preserving virtual environments, such as Second Life or World Without Oil.

Archive crawlers start from "seed" URLs used to fetch an initial set of Web pages, which are then parsed to obtain additional links. Seeds can be supplied manually, generally by subject experts, or automatically. A number of projects center around the development of tools for managing the crawl, such as providing seeds and setting harvest parameters. Tools are also needed for analyzing results in order to determine appropriate future parameters such as crawl frequency.

The Internet Archive, a nonprofit corporation founded by Brewster Kahle, has been critical both in archiving Web content and in providing open-source tools for other archiving efforts. The Internet Archive also offers the hosted service Archive It, which allows subscribers to archive, catalog, and search collections of their own. The Internet Archive was a charter member of the International Internet Preservation Consortium (IIPC) along with the national libraries of Australia, Canada, Denmark, Finland, France, Iceland, Italy, Norway, and Sweden, the British

Library, and the Library of Congress. The IIPC encourages the use of common tools, techniques, and standards for Web archiving, and has developed or vetted a set of open-source applications for crawling, managing crawls, managing storage, and providing access.

Web archiving has been a focus of NDIIPP grant awards, including the Web-At-Risk project, led by the California Digital Library, and the Echo Depository, led by the University of Illinois at Urbana–Champaign. Both projects are developing tools for selecting and capturing materials published on the Web.

Up to now, far more attention has been paid to capturing Web content than to its active preservation, and there is an acknowledged need for research into preservation methods for Web archives. The potentially huge volume of material and the diversity of source file formats make Web archives particularly problematic from a preservation point of view.

One of the best ways to keep current in Web archiving developments is to follow the annual International Web Archiving Workshop, which makes all papers and presentations freely available on the Web.

International Web Archiving Workshop
www.iwaw.net

Readings

- International Internet Preservation Consortium Web site, http://netpreserve.org.
- Julien Masanés, ed., *Web Archiving: Issues and Methods* (Springer, 2006), www.springerlink.com/content/u723352353416271/fulltext.pdf (login required). The author is the guru of Web archiving and coordinates the International Internet Preservation Consortium.

Databases

Databases are critical to the conduct of science and social science, not to mention business and government. Particularly in academic environments, there is an understanding that the long-term preservation of databases begins with good data curation practices while the database is being created and actively used for research. Data creators must be responsible for creating or collecting data in accordance with appropriate technical and procedural standards, and for ensuring that data is authentic, reliable, and of high quality. The digital provenance of the data, which in this case includes documentation of the methods, techniques, and instruments used in data collection, is especially important.

Optimally, data no longer being actively collected and used for primary research should be passed over to a curation center for reuse and preservation. Curators are responsible for creating metadata for discovery and access, maintaining documentation, and enhancing annotation and linkages. The extent to which curators must be subject specialists, as opposed to generic information professionals, is still under debate.

As the curatorial framework is emerging, so are best practice techniques for database preservation. Nonetheless, at this time there are more questions than answers. A 2007 International Workshop on Database Preservation held in Edinburgh, Scotland listed the following issues:[4]

- What are the salient features of a database that should be preserved?
- What are the different stages in the database preservation's life cycle?
- How do we keep archived databases readable and usable in the long term (at acceptable cost)?
- How do we separate the data from a specific database management environment?
- How can we preserve the original data semantics and structure?
- How can we preserve data while it continues to evolve?
- How can we have efficient preservation frameworks, while retaining the ability to query different database versions?
- How can multi-user online access be provided to hundreds of archived databases containing terabytes of data?
- Can we move from a centralized model to a distributed, redundant model of database preservation?
- What documentation is preserved together with a database, and in what format?
- What are the legal encumbrances on database preservation?
- What can be learned from traditional archival appraisal for the selection of databases for preservation?
- To what extent can the preservation strategies, and procedural policies developed by archivists be adapted for databases?

In the sciences in particular, millions of dollars are spent on research projects that generate and rely on massive amounts of digital data, making the stewardship and preservation of these databases a matter of major impor-

tance to scientific programs and funding agencies. Both the U.K. e-Science Programme and the U.S. National Science Foundation (NSF) consider mechanisms for data curation to be a part of the necessary cyber-infrastructure for the 21st century. The U.K. Core e-Science Programme funds the Digital Curation Centre as one of its six key activities (see Chapter 5). The NSF issued in June 2006 a call for proposals for "Sustainable Digital Data Preservation and Access Network Partners (DataNet)" with initial funding of $100 million. The project aims to establish a few exemplars of a wholly new type of organization that integrates library and archival sciences, cyber-infrastructure, computer science, information science, and domain science expertise to "provide reliable digital preservation, access, integration, and analysis capabilities for science and/or engineering data over a decades-long timeline."[5]

Reading

- National Science Board, "Long-Lived Digital Data Collections: Enabling Research and Education in the 21st Century" (Sept. 2005), www.nsf.gov/pubs/2005/nsb0540/nsb0540.pdf. A strategy for the National Science Foundation.

New Media Art

New media art is defined in Wikipedia as "a genre that encompasses artworks created with new media technologies, including computer graphics, computer animation, the Internet, interactive technologies, robotics, and biotechnologies." Many of the objects created by these technologies are usually addressed with the preservation strategies of migration (as for videos) or emulation (as for executables). The success of these strategies in reproducing the performance of an artwork is largely untried, however. "Seeing Double: Emulation in Theory and Practice" is a noteworthy exception. That 2004 exhibition at the Guggenheim Museum showed original and emulated versions of a set of media art installations side by side. A symposium held in conjunction with the exhibition explored the process of emulation and the success of the attempted recreations.

Another preservation strategy applicable to new media art is reinterpretation, or total recreation in current media. Regardless of preservation strategy, establishing the significant properties of the artwork is key, preferably with the advice and consent of the artist. Jon Ippolito of the Guggenheim Museum and the Variable Media Network has developed the Variable Media Questionnaire, an interactive tool to help document an artist's conception of significant properties and identify appropriate strategies for preserving the work. Building on this, Richard Rinehart, from the Berkeley Art Museum/Pacific Film Archive, has proposed an XML-based language called Music Art Notation System (MANS) for recording the information obtained. This XML "score" can enable an artwork to be performed again using different technologies.

For artists and curators, the preservation of new media art is a matter of some urgency. As James Coddington, the chief conservator for the Museum of Modern Art, said, "If future generations are to understand the art of our time, they need to have real examples presented in authentic manner to understand what we and our artists were talking about. And that is very difficult."[6]

A major challenge in the preservation of new media art is simply establishing what preservation means in a context where the significant properties of an installation may differ for each beholder.

Readings

- "The Archiving and Preservation of Born-Digital Art Workshop" (2003), www.erpanet.org/events/2004/glasgowart/briefingpaper.pdf. A good overview prepared as a short briefing paper and bibliography.
- Alain Depocas, Jon Ippolito, and Caitlin Jones, "Permanence through Change: The Variable Media Approach," Guggenheim Museum, 2003, www.variablemedia.net/pdf/Permanence.pdf.
- Richard Reinhart, "A System of Formal Notation for Scoring Works of Digital and Variable Media Art" (2007), www.bampfa.berkeley.edu/about/formalnotation.pdf.

Personal Collections

Nearly everyone in the developed world has some interest in the longevity of their personal digital collections. According to statistics by Nationmaster.com, in 2004 the United States had 762 personal computers per 1000 people (compared with 740 television sets).[7] Individuals accumulate personal content using computers in their homes and at work, digital cameras and video recorders, and camera phones. They store downloaded music, photos of their children, tax returns, schoolwork, and lots and lots of e-mail.

Librarians and archivists have a professional as well as personal stake in the preservation of personal digital collections. Today's personal collections are the source material for tomorrow's historians and biographers, and many will be desired and/or acquired by cultural heritage institutions. Moreover, some preservationists believe that the general public's interest in the longevity of their own digital materials can be leveraged to achieve a greater awareness of and support for institutional digital preservation initiatives.

"The Long Term Fate of Our Digital Belongings: Toward a Service Model for Personal Archives" was one of the first in-depth field studies of how ordinary people provide for the longevity of their own digital files.[8] Among the many interesting findings, the authors identified environmental factors that complicate archiving for individuals, such as the pervasiveness of malware and reliance on ad hoc IT support from friends and relatives (and even the occasional ex-husband). In addition, they list four central challenges of personal archiving that make the digital arena more difficult than the artifactual and analog one:

- Digital materials are accumulated at a much faster rate than physical belongings, making it harder to judge their future worth.
- Personal assets tend to be scattered on multiple machines and many different types of online and offline media.
- Personal curation practices are "in many ways a direct consequence of benign neglect coupled with an incomplete understanding of heterogeneous file systems and digital formats."
- The desktop metaphor does not support facilities for long-term access.

The Universities of Oxford and Manchester investigated the preservation of digital personal collections in the Paradigm project (Personal Archives Accessible in Digital Media), which ran through 2007. Paradigm produced best-practice guidelines for curators in the form of an online workbook, which also includes a useful set of guidelines for individual creators of personal data. In 2007 the (U.K.) Arts and Humanities Research Council funded the Digital Lives research project to run from 2007 through 2009. Spearheaded by the British Library, Digital Lives is studying the intersection of personal digital collections and research repositories. The study will attempt to ascertain more about the behavior and attitudes of individuals, investigate legal and ethical issues that impact personal digital collections and their acquisition by repositories, look into promising technologies for preservation of personal objects, and finally, assess the curatorial workflows of the British Library in light of the above.

Digital Lives Web site
www.bl.uk/digital-lives/index.html

Readings

- Paradigm Project, *Workbook on Digital Private Papers,* 2007, www.paradigm.ac.uk/workbook.
- Neil Beagrie, "Plenty of Room at the Bottom? Personal Digital Libraries and Collections," *D-Lib Magazine* 11 no. 6 (June 2005), www.dlib.org/dlib/june05/beagrie/06beagrie.html.

Notes

1. Linda Contera, ed., *Archiving Electronic Journals: Research Funded by the Andrew W. Mellon Foundation,* 2003, www.diglib.org/preserve/ejp.htm (accessed Nov. 17, 2007).
2. NLM Archiving and Interchange Tag Suite Web site, http://dtd.nlm.nih.gov (accessed Nov. 17, 2007).
3. *The Long-Term Preservation of Authentic Electronic Records: Findings of the InterPARES Project,* Appendix 6, www.interpares.org/book/index.cfm (accessed Nov. 17, 2007).
4. International Workshop on Database Preservation 2007 Web site, http://homepages.inf.ed.ac.uk/hmueller/presdb07/index.html (accessed Nov. 17, 2007).
5. Sustainable Digital Data Preservation and Access Network Partners (DataNet), NSF Office of Cyberinfrastructure Web site, www.nsf.gov/funding/pgm_summ.jsp?pims_id=503141&org=OCI&from=home (accessed Nov. 17, 2007).
6. Terry Schwadron, "Preserving Work That Falls Outside the Norm," *New York Times,* March 29, 2006, www.nytimes.com/2006/03/29/arts/artsspecial/29digital.html (accessed Nov. 28, 2007).
7. NationMaster.com Web site, www.nationmaster.com/index.php (accessed Nov. 17, 2007).
8. Catherine Marshall et. al., "The Long Term Fate of Our Digital Belongings: Toward a Service Model for Personal Archives," 2006, www.csdl.tamu.edu/~marshall/archiving2006-marshall.pdf (accessed Nov. 17, 2007).