Library and Information Services in Astronomy V
ASP Conference Series, Vol. 377, 2007
S. Ricketts, C. Birdie and E. Isaksson (eds.)

Digital Preservation and Electronic Journals

Evan Owens

Portico, 100 Campus Drive, Suite 100, Princeton, NJ 08540, USA

Abstract. Since the last LISA conference, digital preservation has become a major area of interest world-wide. This talk will survey the current landscape in the long-term preservation of electronic information, looking at the organizational and technical components of preservation and then consider recent developments in the preservation of electronic journals.

1. Introduction

I spoke on the topic of archiving electronic journals at LISA III in 1998 and I touched on that topic again briefly at LISA IV in 2002. Today I'm going to return to the topic but try to put it in the larger context of the explosion of interest and activities in digital preservation. My original topic for this presentation was focused on e-journal preservation projects; however, the Council on Library and Information Resources (CLIR) is doing a major study on that topic (Kenney 2006), better and more impartial than anything that I can do, so I have adjusted my talk accordingly.

I had the pleasure of participating in the LISA II, III, and IV conferences representing the University of Chicago Press. Today I represent Portico, a new, not-for-profit electronic archiving service established in response to the library community's need for a robust, reliable means to preserve electronic scholarly journals. Portico was initiated by JSTOR and has been developed with the initial support of Ithaka, The Andrew W. Mellon Foundation, and the Library of Congress. Portico's mission is to preserve scholarly literature published in electronic form and to ensure that these materials remain accessible to future scholars, researchers, and students.[1]

To set the stage, I'd like to take us back to around 1995–1996, when web delivery of e-journal content was brand new. Peter Boyce, then Executive Director of the American Astronomical Society, in a paper in *Serials Review* asked the question: what are the consequences of electronic publishing, and suggested answers from five different points of view (Boyce 1997):

- Author – "No more proofreading. Immediate publication of my work!"
- Publisher – "Production tasks can be automated."
- Librarian – "Cheaper serials, more complex operations."

[1] More information about Portico can be found at our web site: `http://www.portico.org`.

- Reader – “Access to the full literature for free!”
- Archivist – “Everything will be lost!”

I wish that I could say to you that the archivist doesn’t need to worry, that all the problems have been resolved in the last 10 years, but I can’t. The reality is that, although there has been a tremendous amount of work done on digital preservation, especially since 2002 or so, there is still a lot more to do. This is not yet a solved problem.

Today I’m going to try to share with you some highlights of the digital preservation scene as I have seen it based on my explorations in the last three years of working full time in this area. I can’t possibly do a truly systematic survey so I’m going to sample instead: touch on the role and influence of the astronomical community on digital preservation and talk about some of the projects and activities in which I have been involved that represent the range of the technical problems. The point that I would like you to carry away is that digital preservation is both a social/political problem and a technological problem. Then I will focus in on the particular problem of e-journal preservation.

2. Introduction to Digital Preservation

Digital preservation, simply put, is ensuring the long-term viability of digital objects: twenty, fifty, one hundred or more years from now, will we be able to read the files, understand the structure of the files, and be sure that we have an authentic copy of the work? Preservation must address the physical layer (storage media), the logical layer (file formats and data structures), and the conceptual or intellectual layer (the “work”). The major approaches to preservation include emulation or maintaining the original technology, migration of objects to currently supported formats, or simple preservation of the bits for future digital archeologists.

I have heard digital preservation defined as “interoperability with the future.” I don’t know who first said that but it is an immensely helpful concept because it points to the need for standards, protocols, best practices, and clear roles and responsibilities. If we can’t interoperate with the present, we are unlikely to be able to interoperate with the future. Astronomy is strong on interoperability and standards and therefore the preservation risk may be lower than in some other areas. Journal publishing is not at that level yet as there are few standards in use, though that is beginning to change finally. In the discipline of digital archiving itself, standards are just beginning to be defined.

3. The Landscape of Digital Preservation

To state the obvious: digital preservation is everyone’s problem. There is electronic content all over the place, in enormous quantities, growing by leaps and bounds. Among the players in digital preservation are memory institutions such as libraries, archives, and museums; research organizations such as universities, laboratories, and data centers; government agencies; corporations because of regulatory compliance and business continuity needs; and even private individuals who increasingly have substantial personal collections of digital objects.

What is less obvious (or at least easy to forget) is that the issues (technical, organizational) are not exactly the same in all cases. Some want or need to preserve content for ever; for others 75 years is adequate (some times referred to as "medium-term preservation"). That it is or soon will be a general problem suggests that some solutions will eventually come from the commercial sector.

The variety of digital preservation projects is tremendous and each has different technical characteristics. Library and media digitization projects for books, journals, and manuscripts are controlled environments with at least the potential for good metadata. Website harvesting is a completely different preservation challenge: an uncontrolled environment with minimal metadata available. Electronic records retention as practiced in government and industry has the potential for lots of control with mandatory metadata and formats. Published electronic content such as e-journals is less controlled, with good descriptive metadata but variable or no technical metadata. Scientific data is yet another set of challenges: enormous quantities of information and high expectations for long-term usability. The bottom line is that one size does not fit all and that different approaches will be needed to ensure appropriate long-term preservation of all content.

There will be variations in the business or organizational motivations for preservation, in the stability of the content, in the value of the content, and in the choice of an appropriate approach, as is already the case today for conservation of physical objects. Some high value objects may warrant extraordinary measures, as in the case of an 18-foot long scroll from Thailand at the British Library that took six person-months to conserve. Other objects may not be worth saving: data generated by simulation software rather than by observation is likely to be cheaper to regenerate than to place under managed preservation. Truly, one size does not fit all!

The best web resource for exploring the digital preservation landscape is the PADI (Preserving Access to Digital Information) website of the National Library of Australia,[2] which has done pioneering work in this field and maintained this incredibly useful site since 1996. Web links for all the projects mentioned here can be found through PADI. Digital preservation includes traditional players such as the national archives and libraries and the digital library communities such as CNI (Coalition for Networked Information), DLF (Digital Library Federation), JCDL (Joint Conference on Digital Libraries), and ECDL (European Conference on Research and Advanced Technology for Digital Libraries); but also a wide range of new organizations and projects such as the Digital Preservation Coalition (DPC), Erpanet, the Digital Curation Center (DCC), and the National Digital Information Infrastructure and Preservation Program (NDIIPP), to name but a few.

Digital preservation is now beginning to take on the feel of a separate discipline, with its own courses and workshops. Specialized conferences have also emerged: the IS&T Archiving Conferences (from 2004), the International Conference on Preservation of Digital Objects (from 2004), and the International Digital Curation Conference (from 2005). There is also attention being given

[2] `http://www.nla.gov.au/padi/`

to defining the profession of digital preservation: what skills and expertise are required and what sort of education and certification might be appropriate.

4. Some Milestones in the History of Digital Preservation

Now, jumping back to 1996, I'd like to point out some landmarks or milestones in the history of digital preservation. The first historical thread is a set of documents that helped define what digital preservation is.

1. Preserving Digital Information (Garrett & Waters 1996). In 1994, the Commission on Preservation and Access and RLG created the Task Force on Archiving of Digital Information, charged with investigating and recommending means to ensure "continued access indefinitely into the future of records stored in digital electronic form." In May 1996 the 21-member task force, co-chaired by Donald Waters and John Garrett, completed their final report. This report set out the conceptual framework through which we now understand this problem.

2. The OAIS Reference Model (CCSDS 2002). Developed by the NASA Consultative Committee for Space Data Systems, this standard represents the culmination of a series of international workshops that began in 1995. The influence of the astronomy community on digital preservation through OAIS can not be overstated. The OAIS reference model has become an ISO standard and is now ubiquitous in the preservation community: it seems like every talk at every conference on archiving must have a copy of the OAIS Functional Model Diagram or an equivalent, or must apologize for its absence from the presentation! The importance of the work has been in providing a neutral vocabulary to describe roles and responsibilities, the flow of content, and how the archive meets the needs of its "designated community", the consumers of the content. The reference model makes it clear that there is more to preservation than just the storage or management of bits, including preservation planning, administration, and management. The ISO standard version of OAIS is now five years old and the first revision is coming up soon. It will be interesting to see what changes will be made. There has been some push back about OAIS on the grounds that it is biased toward data and data sets (large amounts of very controlled data) and that it doesn't apply as well in messier and less controlled environments such as web harvesting or e-journal publishing.

3. Trusted Digital Repositories (RLG-OCLC 2002). Building on the OAIS reference model, an international working group organized by RLG and OCLC began studying the characteristics and responsibilities of a trusted digital repository in 2000, leading to a report in 2002. "A trusted digital repository is one whose mission is to provide reliable long-term access to managed digital resources for its designated community, now and in the future." Some of the attributes of such a repository include compliance with the OAIS Reference Model, administrative responsibility, organizational viability, financial sustainability, technological and procedural suitability, system security, and procedural accountability. It is noteworthy that this list is as much or more about organizational characteristics as it is about technology issues.

4. Audit Checklist for the Certification of Trusted Digital Repositories (RLG-NARA 2005). Building on the previous document, a committee organized by RLG and NARA developed a 60 page check list to use as the basis for certification of repositories. The text of the checklist is available online. Typical items include "A1.1 Repository has a mission statement that reflects a commitment to the long-term retention of, management of, and access to digital information on behalf of depositors" or "B1.5 Repository obtains sufficient physical control over the digital objects to preserve them" or "B3.10 Repository has mechanisms to change its preservation plans as a result of its monitoring activities."

Auditing and certification of digital repositories is a hot topic right now. The Andrew W. Mellon Foundation has funded a project by the Center for Research Libraries to conduct a pilot study, not to certify particular repositories but to learn from experimental audits what such a certification process using the checklist might entail.[3] Participating archives in this pilot included the Koninklijke Bibliotheek, Portico, the Inter-university Consortium for Political and Social Research (ICPSR), and LOCKSS. The CRL report is expected to be released in late 2006. As one of the guinea pigs in the study, I can report that in practice the audit checklist is very broad and very subjective; more work needs to be done on defining specific evidence that should be presented to support the questions raised by the checklist. We are still a long way from having defined for archives the equivalent of "generally accepted accounting principles" (the basis for a financial audit in the USA).

5. The Technological Component of Preservation

The technological component of digital preservation has also been the subject of considerable study and research. The technological issues can be grouped according to the three layers mentioned earlier: physical (storage media, replication; preserving the bits), logical (file formats, structured data; preserving the organization of the bits), and intellectual (preserving usability/accessibility to the "work", primarily through metadata). Physical storage is reasonably well understood: hardware will fail so replace it early and often. For me, the most interesting questions are at the logical and intellectual level. To give you a sense of the kinds of issues that are currently under study, I will briefly describe three recent technology projects, primarily US-based. For a broader picture of the spectrum of current research world-wide, consult the NDIIPP project site, the DCC web site, and PADI.

1. JHOVE: JSTOR/Harvard Validation Environment.[4] JHOVE is a tool for format-specific identification, validation, and characterization of files. Developed by Harvard Library with funding assistance from the JSTOR Electronic-Archiving Initiative (now known as Portico), JHOVE addresses a key need in digital preservation. As noted on the JHOVE project page, "the concept of representation format, or type, permeates all technical areas of digital repositories. Policy and processing decisions regarding object ingest, storage, access, and preservation are frequently conditioned on a per-format basis. In order to

[3] The project website can be found at `http://www.crl.edu/content.asp?l1=13&l2=58&l3=142`

[4] `http://hul.harvard.edu/jhove`

achieve necessary operational efficiencies, repositories need to be able to automate these procedures to the fullest extent possible." JHOVE has been a tremendous success and has been widely adopted by libraries and archives. A proposal for second round development of JHOVE, to be undertaken jointly by Harvard University Library, Stanford University Library, and Portico, is expected to be funded shortly.

2. GDFR: Global Digital Format Registry. Another project from Harvard Library,[5] GDFR came out of a series of meetings in 2003 and has recently been funded by the Andrew W. Mellon Foundation. As noted on the website, the GDFR "will provide sustainable distributed services to store, discover, and deliver representation information about digital formats." One of the things that most surprised me as I began working in the archiving space was learning that there are no "official" names for file formats of sufficient granularity that would allow archives to talk to each other about the digital objects that they are preserving. GDFR would provide the framework for agreeing on names and definitions of formats as well as organizing the preservation of supporting documentation and specifications.

3. PREMIS: Data Dictionary for Preservation Metadata. The PREMIS project[6] was an attempt to develop a data dictionary that described a minimum set of required information for long-term preservation, what every repository must know about its objects, based on a survey of current preservation implementations world-wide. What became obvious to me during a year of weekly conference calls was how early we are in our collective understanding of digital preservation. The PREMIS data dictionary is now at the trial use stage and is expected to be revised based on experience from actual implementations.

6. The Archiving of Electronic Journals

Now I will follow another thread in the brief history of digital preservation: the preservation of published electronic content, specifically electronic journals. I have already mentioned the report by Garrett and Waters. In 1999, CLIR (Council on Library and Information Resources), DLF, and CNI convened a group of publishers and librarians to discuss responsibility for archiving the content of electronic journals. A series of meetings led to the publication in May 2000 of the document, "Minimum Criteria for an Archival Repository of Digital Scholarly Journals" (Version 1.2). Soon after, the Andrew W. Mellon Foundation solicited proposals for one-year e-journal archiving planning projects. Seven institutions were awarded grants for projects carried out from early 2001 through early 2002: Cornell, Harvard, MIT, Pennsylvania, Stanford, Yale, and the New York Public Library. The reports were published by DLF.[7] The partner publishers for the Harvard project were The University of Chicago Press (and therefore indirectly the AAS journals), John Wiley & Sons, and Blackwell Publishing. Of

[5] `http://hul.harvard.edu/gdfr`

[6] `http://www.loc.gov/standards/premis/`

[7] `http://www.diglib.org/preserve/ejp.htm`

the seven projects, one is still going (LOCKSS from Stanford); the work done by Harvard, Yale, and Cornell has strongly influenced the work of the JSTOR Electronic-Archiving Initiative, now called Portico.

The challenge of digital preservation of electronic journals has not changed much since my presentation on this topic at LISA III; the main difference is that what was true for *The Astrophysical Journal* in 1998 is now more widely applicable. Journal publishing models are still evolving: after ten years of delivery of e-journals on the web, there is still wide variation in practice and product. The possible relationships between print and online and between online PDF and online HTML are many; in effect, an e-journal article is a work with multiple "manifestations." This makes preservation an interesting challenge, particularly when the manifestation delivered via the web (the HTML) is a subset of richer content and information resources that exist behind the scenes, as it were.

At Portico we are working with a wide range of publishers, from small to very large. Based on our evaluation of publisher data and extensive interviews, it is evident to us that there are many outstanding problems in current e-journal publishing practice that need industry attention: in areas of content management and quality control; in documentation, naming, and packaging; in the handling of author-supplied supplemental content; in the use of persistent identifiers such as DOI; in control of versioning and policies regarding revisions or updates; and in control of issue-level content such as front matter and back matter. This is no surprise: publishing electronically is still very young. But as the industry moves rapidly away from print publication, it is important that the implications of that change are carefully considered. An example of a transition problem is the treatment of information about the editorial board of the journal. In a print journal this is commonly included inside the front cover of each issue and the history of changes to the board is thereby captured. In many online journals that information is on the journal home page rather than with the article content and is updated in place, making it impossible to determine who the previous editors were at any given point in the past. This is a small point, but illustrates the kind of problems that must be addressed if we are to make a successful transition from print to online-only publication.

Some standards for e-journal publishing are starting to become established. Publishers have begun to adopt the NCBI NLM Journal Article DTDs, moving away from proprietary DTDs; the NLM DTDs have in effect replaced the ISO 12083 Serials DTD, which was developed during the 1990s but never widely adopted. There is also wide-spread agreement on the utility of using standard identifiers by assigning DOIs and participating in CrossRef.[8] However, there are not yet standards for packaging and handling of e-journal content as it moves from system to system, either technically or from a business point of view. The UKSG (United Kingdom Serials Group) and ALPSP (Association of Learned and Professional Society Publishers) have organized a working group to study the problems that arise when electronic journals change publishers.[9] In another indication of how complex the landscape has become, NISO (National Informa-

[8] `http://www.crossref.org`

[9] `http://www.projecttransfer.org`

tion Standards Organization) and ALPSP have organized a working group to develop a vocabulary to describe all the different versions of journal articles that are now being distributed on the web.[10] It is clear that there is still a lot to be invented and agreed upon as best practices are developed for the electronic publishing of journals.

Given the importance of electronic journals as the primary means of publication for a significant portion of our intellectual heritage, it is no surprise that there are several initiatives underway around the world to attempt to preserve this body of content. The CLIR survey of e-journal archiving initiatives has now been released (Kenney 2006) and is available on the web. The projects evaluated included KOPAL, the Koninklijke Bibliotheek e-Depot, Los Alamos National Laboratory, LOCKSS Alliance/CLOCKSS, National Library of Australia, OCLC ECO, OhioLINK, Ontario Scholars Portal, Portico, and PubMed Central. The technical approaches taken by these projects vary and depend on what they understand to be the object of preservation: varying combinations of the PDF rendition, the HTML rendition, and the XML or SGML source files. The particulars are documented in the CLIR survey.

7. Visions for the Future and Visions from the Past

What can we expect for the future? Will everything be lost? Will what we wish to preserve be preserved? Will we be able to use it? Digital preservation is very new and there is a lot yet to be learned. We need standards and best practices; we also need diversity to hedge our bets. Dale Flecker of Harvard University Library put it neatly when he said that long-term preservation requires that we store content on at least three continents, using three different operating systems, and under three different political systems. It is dangerous to assume that there is a single perfect solution when the stakes are so very high. This does create a fundamental tension: to the extent that we try to establish best practices and uniform standards, are we in effect creating a monoculture and reducing diversity?

I often think of an earlier change in publishing practices, one that did not have happy consequences. In the 18th-century sheet music was printed on linen paper; that paper survives today as beautiful as it was 250 years ago. In the 19th-century sheet music was printed on acidic paper; today that paper is brittle and falls to pieces as the pages are turned. New technology doesn't always make things better. I hope that the electronic journals published today will not someday be seen as fragile, or that if so, we will have successfully de-acidified (as it were) our electronic content into stable formats and preserved them for future generations.

References

Boyce, P. B., Owens, E., & Biemesderfer, C. 1997, Serials Rev., 23, 1

CCSDS 2002, Reference Model for an Open Archival Information System – OAIS, by National Aeronautics and Space Administration Consultative Committee for Space

[10] `http://www.niso.org/committees/Journal_versioning/JournalVer_comm.html`

Data Systems (Washington, DC: NASA), 148, http://public.ccsds.org/publications/archive/650x0b1.pdf

Garrett, J., & Waters, D. 1996, Preserving Digital Information: Report of the Task Force on Archiving of Digital Information, 64, http://www.rlg.org/ArchTF/

Kenney, A. R, Entlich, R., Hirtle, P. B., McGovern, N. Y., & Buckley, E. L. 2006, E-Journal Archiving Metes and Bounds: A Survey of the Landscape (Washington, DC: Council on Library and Information Resources), http://www.clir.org/pubs/reports/pub138/pub138.pdf

RLG-NARA 2005, An Audit Checklist for the Certification of Trusted Digital Repositories: Draft for Public Comment (Mountain View, CA: RLG & NARA), http://www.rlg.org/en/page.php?Page_ID=20769

RLG-OCLC 2002, Trusted Digital Repositories: Attributes and Responsibilities (Mountain View, CA: RLG & OCLC), http://www.rlg.org/longterm/repositories.pdf