Stacy Kowalczyk **L702** **March 24, 2005**

Electronic Journal Archiving and Preservation Annotated Bibliography

Across the country and around the world, libraries are spending an increasing amount of their budgets to purchase electronic journals. An ever-increasing percentage of new knowledge is being documented and distributed in an exclusively electronic medium. What will happen to this information over the next 100 years? Without a concerted effort, these digital resources could become obsolete. This annotated bibliography, a total of 28 journal articles, white papers and grant reports, provides a glimpse into the complicated world of electronic journal preservation.

Since the seminal paper on digital preservation in 1996 (Waters & Garrett), a great deal of progress in digital preservation in general, and ejournal preservation specifically, has been recorded. Two very important developments have shaped this progress– the Open Archival Information Systems (OAIS) Reference Model (CCSDS, 2001) and the Andrew W. Mellon Foundation Electronic Journal Archiving Planning Grants (Cantara, 2003). The OAIS has provided a common language and framework for discussing the process of ejournal preservation, and the Mellon grants have provided an excellent base of experience that will enrich future planning.

Often, the discussion of digital preservation is overshadowed by technology – format obsolescence, media deterioration, standards development, and scale. What is lost in the discussion is the art and science of *preservation*. Libraries, museums, and archives have long understood that preservation begins with assessment – what should be preserved, under what conditions and to what result. Assessing - understanding that the levels of preservation – from conserving the total experience of the object to preserving the content only – needs to be applied to the digital world, just as it has been in the paper world (Arms, 1999; Pinfield and James, 2003). The preservation mainstays – content, fixity, reference, provenance and context – need to be part of the digital preservation discussion (Waters, 1996).

Technology is no longer the barrier that it seemed 10 years ago. Three working solutions are in operation today – eprint servers, digital library repositories and Lots of Copies Keeps Stuff Safe project, LOCKSS. Having multiple technology solutions is excellent. Should one prove fatally flawed, other options will continue to be viable. (Flecker, 2001) The issues of format obsolescence, media refreshing and multimedia objects are understood and solvable. Multiple institutions are working on technical solutions that will available in the near future. (Cantara, 2003; Falk, 2003; Harvard University Library, 2003).

Most of the issues that complicate ejournal preservation involve the societal and organizational environment of scholarly publishing (Kling, Spector, & Fortuna, 2003; Pinfield & James, 2003) – conflicts between the academic promotion process of publishing in peer reviewed journals and the great desire to self-publish and self archive

as well as the conflicts between the $9 billion scholarly publishing industry and the not-for-profit budgets of universities and libraries **(**Kaser, 2003) . These issues require a great deal more study. I think that changes to the scholarly publishing model will happen, but slowly.

Three significant issues of social informatics remain unresolved with no answers readily available:

- Who is responsible?
- Who will pay?
- Who can access the archived journals?

As the social and organizational conflicts resolve over time, it seems likely that solutions will arise.

Arms, W. (1999). Preservation of scientific serials: Three current examples. *Journal of Electronic Publishing, 5.* Retrieved March 4, 2005 from http://www.press.umich.edu/jep/05-02/arms.html

The *ACM Digital Library*, the Internet RFC series, and *D-Lib Magazine*: these three electronic publications were studied to determine the technology, the organization, and the preservation challenges of each. "This paper asks what can be done today that will help to preserve the information contained in these three example for scientists and historians a hundred years from now. The answers are partly technical and partly organizational."

The *ACM Digital Library* is an online publication of all of the articles published by the ACM. The articles are in a database and are marked up with SGML with a Web interface for access. The Internet RFC services are the "request for comment" that are the primary technical documentation of the Internet. Over the past 30 years, more than 2,700 RFCs were published online without any intellectual, structural or administrative metadata. There is no central organization that "owns" these documents. There is no central access mechanism for the RFCs. "Until recently the older RFCs were not collected systematically and some of the older RFCs have been lost." *D-Lib Magazine* is a monthly publication of CNRI supported by DARPA grants. Each article has a Digital Object Identifier (DOI) and a Dublin Core (DC) XML metadata record.

There are three levels of preservation – conservation (preserving look and feel as well as content), preservation of access (like the current systems of access – websites), and preservation of content ("a simple warehouse of the article with minimal metadata"). If a publisher is actively maintaining a serial there is no need for other originations to duplicate the technical work of preservation, but publishers cannot be relied on for long-term preservation – the 100-year issue. Publishers will not conserve the materials; when the access system changes, the former interface will be "lost".

With 80,000 members, the ACM will like survive as an organization. CNRI may not fare as well, but it has an extended community. Were CNRI to fold, the ARL community will

likely step in to maintain the *D-Lib Magazine* content. As each article is fully marked up, much of the look and feel could be preserved. The objects most at risk are the RFCs. Without real organizational support, preservation is perilous. While these documents are actively being used, they will persist. But when the Internet is no longer and these documents become the history of science, there is no provision for their preservation. "The solution lies in preparation during the period of active management, so that the technical and legal arrangements for subsequent preservation are already in place."

Bearman, D. (1999). Reality and chimeras in the preservation of electronic records. *D-Lib Magazine 5*(4). Retrieved March 4, 2005 from http://www.dlib.org/dlib/april99/bearman/04bearman.html.

This is article is an opinion piece in D-Lib Magazine. It is a reaction to the Jeff Rothenberg paper published by CLIR, the Council on Library and Information Resources. In that article, Rothenberg declares emulation as the only real solution. (Rothenberg is the champion of the "emulation" school of preservation.) Bearman refutes the Rothenberg paper. Bearman argues that emulation is not a sufficient strategy for preservation. Rothenberg dismisses four preservation solutions: printing to paper, relying on standards to maintain readability, preserving obsolete hardware and software, and migrating data to a new format. Bearman states that migration is the real solution for preservation of electronic records. The emulation supporters state that migration cannot be reliable. But Bearman argues the opposite – that migration is much more reliable for preserving the true meaning of the records and that emulation will introduce many more errors. While this article focuses mostly on electronic records, it is applicable for electronic journals.

Bergmark, D. (2000). Link accessibility in electronic journal articles. Retrieved February 17, 2005 from http://www.cs.cornell.edu/bergmark/LinkAccessibility/paper1.pdf

This article describes research resulting from a CNRI grant for the Open Citation project. The Open Citation project "aims to amplify [citation URLs] by looking up online locations for reference that are not accompanied by URLs." During the course of this project, the researchers determined that they needed to understand the longevity of reference URLs over time before they added new URLs to electronic journals. The 1996 Harter and Kim study showed that 50% of reference links were inaccessible within the same year. The author wanted to determine if the situation had improved in the intervening years. She examined the links from two electronic journals, D-Lib and the Journal of Electronic Publishing, for the entire life span of each publication. Using PERL scripts, she checked the links from all of the articles reference sections. The results of this study are quite different from the Harter and Kim study. For these two ejournals, over the 5 years, 86% of the references were viable and the rate of increase of broken links decreases over time.

Buckley, C., Burright, M., Prendergast, A., Sapon-White, R., & Taylor, A. (1999). Electronic Publishing of Scholarly Journals: A Bibliographic Essay of Current Issues. Retrieved February 17, 2005 from http://www.library.ucsb.edu/istl/99-spring/article4.html

This article is a collection of brief bibliographies on a variety of topics concerning ejournals: access, pricing, cataloging and indexing, licensing, and archiving. Chad Buckley wrote the section on archiving – the only section to be reviewed in this bibliography. Written in 1999, this bibliography looks at the issues of ejournal archiving from a librarian's viewpoint by focusing primarily on the fixity of paper and the mutability of the intellectual content of digital documents, the reasons that ejournals should be archived, and policy issues regarding organizational responsibility for archiving and long-term access rights. Like the Nisonger article below, this early foray provides more questions than answers.

Cantara, L. (2003). Introduction. In: L. Cantara, (Ed.) ***Archiving Electronic Journals: Research Funded by the Andrew W. Mellon Foundation.*** **Retrieved February 17, 2005 from http://www.diglib.org/preserve/ejp.htm**

In early 2001, the Andrew W. Mellon Foundation awarded one year ejournal archiving planning grants to 7 major research libraries: Cornell University Library, Harvard University Library, MIT University Library, New York Public Library, University of Pennsylvania Library, Stanford University Libraries, and Yale University Library. The seven libraries developed 4 different types of planning projects: Subject based – Cornell and NYPL; Publisher-based – Harvard, Penn, and Yale; Dynamic ejournals (frequently changing content) – MIT; Software tool development – Stanford. While all of the projects proposed an economic model, no two were even similar. All of the projects made significantly different decisions about what to archive – from web pages to PDFs to XML. Access to the archives presented another opportunity to diverge – from "dark archives" to fully accessible. Nearly all suggested a JSTOR-like "moving wall." Because of the enormous cost estimates from each of the 7 projects, the Mellon Foundation decided to fund only two projects – the Stanford LOCKSS project and the JSTOR Electronic-Archiving Initiative.

CCSDS. (2001). ***Reference model for an open archival information system (OAIS), draft recommendation for space data system standards, CCSDS 650.0-R-2.*** **Red Book. Issue 2. Washington, D.C.: Consultative Committee for Space Data Systems. Retrieved March 4, 2005 from http://www.ccsds.org/RP9905/RP9905.html.**

The Open Archival Information Systems (OAIS) Reference Model is the new buzzword in the digital preservation community. Like the OSI reference model, it is not a systems design but a set of high-level requirements built as a conceptual framework. It lays out what should be done in a digital preservation archive. It does not provide implementation instructions. The main contribution to the discussion is the concept that preservation is a planned process and needs to be designed in to a digital library from inception. OAIS uses the concept of an Information Package, a transaction and/or datastore that

accompanies a digital object. At ingest, the package is a SIP, a submission information package. It is modified as required to be come an AIP, an archival information package. When the object is distributed for use, the information package is transformed again into the DIP, the dissemination information package.

Crow, Raym. (2002). *SPARC institutional repository checklist & resource guide.* The Scholarly Publishing & Academic Resources Coalition, American Research Libraries. Retrieved January 20, 2005 from www.arl.org/sparc

This document is a blueprint for implementing a digital, institutional repository in an academic environment from institutional and faculty support to long term funding to technical and systems development. More than a digital library, an institutional repository is designed to take all intellectual output of the university. To fulfill the mission that the content of the system must outlive the system itself, digital repositories need to be content-centric. Therefore, managing the formats of the content is a primary concern to the repository. Not only does the repository need to know, understand and control formats, it needs to be able to understand the nature of the complete object – the conference proceedings in the document form, the presentation that accompanies it, and the poster that preceded it. Most of the extant repositories use METS (Metadata Encoding & Transmission Standard), an XML schema standard supported by the library community and managed by the Library of Congress. To insure long-term viability and the ability to preserve the content, preservation and technical metadata must be collected and attached to both the file(s) and the object. OCLC and RLG (traditional library services vendors) will be offering preservation services to organizations that do not want to build the infrastructure for digital preservation. A repository must also be scalable – to be able to continually grow both in capacity and functionality. Since ejournal are likely to be stored in either a digital library or an institutional repository over time, understanding the issues of repositories is important.

Falk, Howard (2003). Digital archive developments. *The Electronic Library*, *21*, 375–379. Retrieved January 20, 2005 from http://caliban.emeraldinsight.com

Falk attempts to survey the state of digital archives in universities and colleges in the US and abroad. In this very shallow article, he names very few of the innovators, MIT's DSpace, Cornell's FEDORA, and a number of national archive initiatives. He gives no definitions of archives and mixes institutional repositories – large shared file systems with public access to renderable objects – and preservation archives which are often invisible to the end user. No architectural or technical information is provided, leaving the reader to do the real work that this article promised.

Flecker, D. (2001). Preserving scholarly e-journals. *D-Lib Magazine, 7*(9). Retrieved January 20, 2005 from http://www.dlib.org/dlib/september01/flecker/09flecker.html

The traditional model of journal preservation is to have multiple institutions each save their paper copy. In the electronic world, each ejournal has a single instantiation at the publisher's site with one access system. In the early days of electronic journal, libraries

relied on the paper version as the archival copy. But "it in increasingly the electronic versions of titles that are the version of record, containing content not available in the print version." This article discusses the Mellon ejournal archiving projects. The key assumptions of all of these projects are:

- Archives should be independent of publishers
- Archiving should be abed on an active partnership with publishers
- Archives should address preservation over a long timeframe – more than 100 years
- Archives need to conform to standards
- Archives should be based on the OAIS reference model.

Questions that Harvard University Library wants to answer in their study:

- What is the publisher/archive/subscriber relationship?
- Is the archive content usually "dark"?
- When can archived content be accessed?
- Who can access archived content?
- What content is archived? (Advertisements, mastheads, etc.
- Should content be normalized at ingest?
- Should the archive preserve usable object, or just bits?
- Who pays for what?

Digital archiving will need redundancy – not at the same scale that the paper era had, but more than one institution needs to save each journal.

Gadd, E., Oppenheim, C., & Probets, S. (2003). The Intellectual property rights issues facing self-archiving – key findings of the RoMEO project. *D-Lib Magazine*, *9*(9). Retrieved January 20, 2005 from http://www.dlib.org/dlib/september03/gadd/09gadd.html

This article describes the RoMEO project. Part of the UK's Joint Information Systems Committee (JIST) Focus on Access to Institutional Repositories program, the Rights Metadata for Open archiving (RoMEO) project is a one-year initiative to look at Intellectual Property Rights issues of institutional repositories. The project archived academic research papers and provided access via the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) [not to be confused with the OAIS. The OAI-PMH is a server-to-server lightweight transaction set for requesting and sending metadata.] The project's goal was to develop simple rights metadata so that scholars and institutions could protect their research papers in an open-access environment and the OAI data and service providers could protect their metadata. The project developed a survey to determine what academic authors thought that their copyright obligations were and what they were willing to allow in the open-access system. 61% thought that they had copyright to their own work, but 32% did not know. The project team had three options for developing the rights metadata: develop a totally new "rights expression language", use and existing Digital Rights Expression Language (DREL), or to use the Creative Commons Initiative (CC). The Creative Commons initiative has developed at least 11 licenses that content creators may use to make their work available. The team decided to use the CC license metadata but to develop an ORDL (the open source rights language with a data dictionary) XML schema rather than use the CC's RDF/XML

implementation. This approach was used to describe the rights to the research papers as well as the metadata.

Gladney, H. (2004). Trustworthy 100-year digital objects: Evidence after every witness is dead. *ACM Transactions on Information Systems, 22*(3), 406-436. Retrieved January 17, 2005 from http://www.acm.org/dl/

Gladney proposes to build a system to encrypt and archive digital objects with their contextual metadata. Using existing infrastructure and theory, the author designed a workflow for data producers that would create a Trusted Digital Object (TDO). This object would be encoded to prove the authenticity of the underlying bit-stream. The validity of the data would need to be verified by humans.

The article mentions a number of very important concepts in digital libraries and digital preservation:

- the need for trust in the institutions that will archive digital objects
- the need to trust that the bit-stream has integrity
- end-user's basic requirements for digital objects
- the need for unique and persistent identifiers like URNs and URIs
- the Reference Model for an Open Archival Information System (OAIS) conceptual framework
- private/public key encryption
- Open Archives Initiative – Protocol for Metadata Harvesting (OAI/PMH)
- digital audit trail

But the article is provocative. It suggests a number of controversial practices and glosses over a great number of very difficult issues:

- storing the entire TDO vs. building it for delivery
- embedding metadata vs. binding metadata
- reusing identifiers for new editions
- centralized services for resolution and other important functions
- storing and migrating files
- hardware and software obsolescence

Granger, S. (2002). Digital preservation and deep infrastructure. *D-Lib Magazine 8*(2). Retrieved March 4, 2005 from http://www.dlib.org/dlib/february02/granger/02granger.html.

This article tries to develop the concept of "deep infrastructure" that was introduced in the digital preservation seminal paper, Task Force Report on Digital Preservation (see Waters and Garrett below). He proposes a collaborative model for different communities to preserve data. Data producers could work with data users to better define the needs of both. Scholarly communities could work together to save vital data for future research – medical researchers and botanists saving the records of a specific researcher important to both fields. He creates a view of the digital domain with three intersecting circles – libraries and archives, research communities and commercial entities. But within a few

sentences, he begins a diatribe against the commercial entities that make up a third of the domain. He accuses software vendors, hardware vendors, and content providers of undermining digital preservation by building new products. More than just bemoaning the rapid rate of technology obsolescence, he contends that it is by design rather than by innovation. This article does not further the discussion of "deep infrastructure" or preservation.

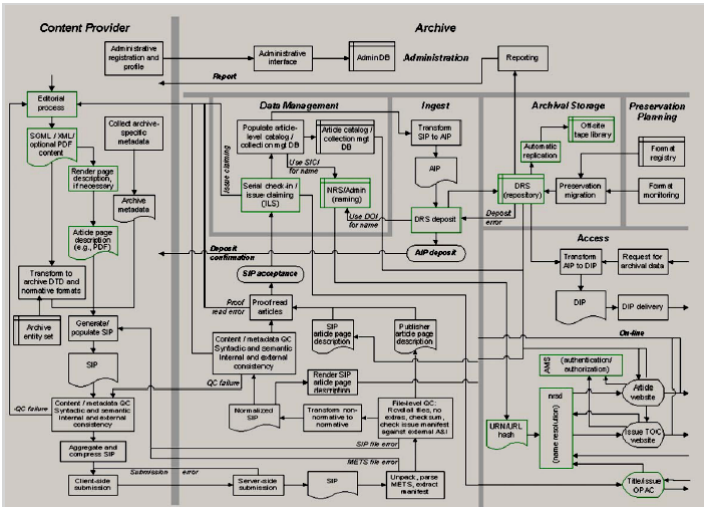
Harnad, S. (2001). The self-archiving initiative: Freeing the refereed research literature online. *Nature 410* (April 26), 1024-1025. Retrieved February 17, 2005 from http://www.ecs.soton.ac.uk/~harnad/Tp/nature4.htm

This article proposes to radically change the process by which scholarly work is disseminated. Rather than using the traditional method of turning over copyright to an organization that will charge others to read their works, authors would band together to review each others work and publish the papers electronically in what is now known as an eprint archive providing free access to all who are thirsty for knowledge. The author asserts that this will be cost effective for the entire scholarly community. He cites some startling figures – 20,000 refereed journals publish 2,000,000 refereed articles every year, and the world's academic institutions pay an average of $2,000 per paper. Only the institutions who can afford this "toll" get access to the knowledge. With the minimum cost of refereeing an article generally accepted to be $500, the author asserts that this cost is inflated and estimates the "true" cost at a maximum of $200 per article. He does not explain how he developed this estimate. With the example of the physics eprint archive, he urges academic institutions, libraries and the researchers to work together to effect the changes necessary to create this type of service for all refereed journals. The physics eprint self-archiving model benefits institutions in three ways – by maximizing visibility and impact of their own research output, by maximizing other researchers access to that information, and by reducing 90% of the costs of traditionally published journals. While the author had a number of innovative ideas, he only considers the refereeing costs and does not address any of the other, very real costs, of self-archiving – technology to support the service, the people to support the service, the overhead of the technology and people (office space, electricity, management, etc.), ongoing upgrades and migration as servers and storage age, back ups, disaster planning and recovery and all of the other costs associated with a production system. When these types of issues are ignored, the innovations are often dismissed as pipe dreams.

Harvard University Library. (2003) Report on the planning year grant for the design of an e-journal archive. In: L. Cantara, (Ed.) *Archiving Electronic Journals: Research Funded by the Andrew W. Mellon Foundation.* Retrieved February 17, 2005 from http://www.diglib.org/preserve/ejp.htm

This is the Harvard University Library's report to the Andrew W. Mellon Foundation for its one year planning grant. This planning grant yielded a report of 33 pages. This report documented an entire process to take raw journal files from publishers, transform them into archival information packets (per the OAIS reference model), provide unique and persistent identifiers to each component of the journal, store the files in the digital

repository, create issue level metadata and access control, and control for the quality of each issue submission. This process is captured in the image below:



Harvard estimated 3 – 10 FTE to run a production operation. While no further funding for this project was obtained, the process itself was useful. This study prompted a number of other valuable studies – a SIP schema, the PDFA initiative, the DLF Format Registry project, and JHOVE, a joint Harvard JSTOR project to develop an object format validation infrastructure.

Inera, Inc. (2001). *E-journal archive DTD feasibility study.* Prepared for the Harvard University Library, Office of Information Systems, E-Journal Archiving Project. Retrieved March 4, 2005 from http://www.diglib.org/preserve/hadtdfs.pdf.

As part of its Andrew W. Mellon Foundation one year ejournal archiving project, Harvard University Library commissions Inera, a consulting company, to determine the feasibility of designing an industry wide DTD for ejournal archiving. Working collaboratively, the two institutions determined the methodology. They selected 10 publisher DTDs for review: American Institute of Physics, BioOne, Blackwell Science, Elsevier Science, Highwire Press, Institute of Electrical and Electronics Engineers (IEEE), Nature, Pubmed Central, University of Chicago Press, John Wiley & Sons. Each of these publishers provided their DTD, documentation and a sample of document instances from multiple journals and issues. Inera determine that a common DTD *could* be developed. They recommended that rather than a DTD, an XML schema be developed, both for longevity of the technology and the additional functionality that schema provides. The Archive Schema (AS) should be less restrictive in structure and more streamlined in element selection than the publishers DTDs. The AS need not provide for all of the publishers' specific markup that does not add to the understanding of the intellectual content. The AS should use public standards wherever possible. The archival XML files that are generated using the AS should include generated text and markup to provide a more effective method to render the content. The report provides recommendations for transformation of the publishers' data at deposit (SIP in OAIS speak) and retrieval (DIP).

Kaser, D. (2003). The future of journals. *Information Today, 20*(3), 1–5. Retrieved January 20, 2005 from http://proquest.umi.com/

This article is an interview between the Dick Kaser, the *Information Today* vice president of content and Pieter Bolman, an executive with Elsevier. Bolman was formerly the CEO of Pergamon and Academic Press. Bolman was interviewed in his role as chair of the PSP Executive Council. The PSP is the Professional Scholarly Publishing Division of the Association of American Publishers (AAP). The main topic of conversation was the future of journals – more specifically, the future business models of scholarly publishing. The questions were clearly pro-business and somewhat anti-academic and anti-library. Bolman credits Elsevier with ending the spiraling serials crisis. Rather than continuing double and triple digit percentage increases in the prices of journals as the subscriptions decrease, Elsevier declared that they will not increase any journal price more than 7.5% per year. Other publishers followed suit. They discuss open access – publishers need to educate their customers (universities and libraries) on the benefits that publishers bring to scholarly communication. "Everyone thinks that once you have a PC, you can be a publisher." (Bolman) The article ends in an editorial (not marked as such) from the author laying out the big picture outline of the scholarly communications debate. He ends with this: "Wasn't it Locke who said that the noblest of causes are best achieved by each one of us pursuing our own selfish interest?...Despite its flaws, the current system of scientific communication somehow manages every year to get the research results of all the scholars in the world officially reported and available for access. It may not be a perfect system. And it may not be cheap. But in the support of intellectual advancement, society could certainly do worse." So much money is involved with scholarly publishing that any fundamental changes will be contentious, painful and slow.

Keller, A. (2001). Future development of electronic journals: a Delphi survey. *The Electronic Library, 19*, 383–396. Retrieved January 20, 2005 from http://hermia.emeraldinsight.com/

This article reports the results of an "international and interdisciplinary" Delphi survey on the future of electronic journals. The expert panel was made up of 45 publishers, scientists, librarians, journal agents and consultants. Scholarly promotion was a main factor in scholarly publishing – "as long as journals remain the main indicator for quality control, scholars will be forced to publish in high-quality journals in order to enhance their career." Even so, the panel thought that preprint archives will be the main competitors to traditional journals over the next 10 years. The ejournal of the future has a number of scenarios:

1. Ejournals will incorporate multimedia and interactive features. They will no longer be "digital *doppelgangers"*. (I love that phrase!)
2. Ejournals will represent customized collection of articles built on personal profiles
3. Ejournals will no longer be "envelopes for articles" but articles will be tagged with quality labels and stored in "large knowledge environments."
4. Articles will be replaced by a stream of dynamic information objects that represent different versions of a paper over its lifetime.

The survey showed no consensus on the possibilities of do-it-yourself publishing. While it is technically feasible today, the social environments do not support DIY publishing. Archiving was an important part of the survey. The consensus of the panel was that archiving will be much more expensive than archiving paper journals and that implementing standards will be the only way that the contents of the data will survive. Who should be responsible for archiving? The survey suggested 5 candidates (only two of which were pronounces likely; data not available on the remaining three):

1. National libraries or depositories (81% likely or very likely)
2. International discipline-specific archives (62% likely or very likely)
3. Publishers
4. Authors or authors' institutions
5. Special commercial providers

The study concludes that "it is not clear who will take the responsibility for archiving ejournals.

Kling, R., Spector, L. B., & Fortuna, J. (2003). The real stakes of virtual publishing: The transformation of E-Biomed into PubMed central. *Journal of the American Society for Information Science and Technology, 55*(2), 127–148. Retrieved January 20, 2005 from http://www3.interscience.wiley.com/

This article discusses the forces that changed the NIH's plan for E-Biomed, a biomedical preprint server and a published article archive, to PubMed Central, a delayed post published article database. The questions that this paper proposes to answer are:

- Why was E-Biomed reworked into a "format preventing self-publishing, preserving peer review, and allowing arbitrary posting delays?"
- Why didn't the final proposal include that features that the scientist who sent comments to NIH supported?
- Who influenced this transformation?

Even though Stevan Harnard and other influential proponent of preprint and self-publishing were proponents of the new format, the publishing industry, including the biomedical scientific associations opposed removing peer review as a requirement for publication. They felt that bad research published under a governmental masthead would provide a certification that would legitimize use of bogus information. Unlike physics, bad biomedical research could result in pain, suffering and death. A very influential player in this issue was the New England Journal of Medicine who introduced the "Ingelfinger Rule" that "defined 'prior publication' so broadly that article posted on Websites of any kind were prohibited from being peer-reviewed and published by the journal." So no pre-print at all for any biomedical journals. The major shift was from readers and authors interested to publishers' interests. The scientific societies wanted to retain membership and revenue by controlling the publication of their journals. Publishers had very similar concerns. This was a very thorough study, including discourse analysis of email postings to the NIH comments list, research in the literature, and content analysis of the numerous proposals.

Liu, Z. (2003). Trends in transforming scholarly communication and their implications. *Information Processing & Management*, *39*, 889-898. Retrieved January 20, 2005 from http://www.elsevier.com/

This article reports the results of a study to determine how scholarly communication has changed over the past century. Because these journals had been published for more than 100 years, American Journal of Mathematics, American Journal of Sociology and Journal of the American Chemical Society were studied. The study analyzed articles only from 11 years of the past century – each of the '00 years from 1900 – 2000. Additionally the study included an analysis of the age of the citations in the 2000 articles. The study focused on collaboration and volume of production. While the number of collaborators is interesting, the growth in volume has a major impact on preservation. If the trend continues, the cost will continually increase as well.

	Collaboration **Authors / Title**	**Volume** **Articles / pages**
American Journal of Mathematics	1900 – 1.04 1950 – 1.24 2000 – 1.45	1900 – 26 articles / 388 p 1950 – 66 articles / 867 p 2000 – 49 articles / 1308 p
American Journal of Sociology	1900 – 1 1950 – 1.13 2000 – 1.58	1900 – 42 articles / 864 p 1950 – 48 articles / 622 p 2000 – 40 articles / 1840 p
Journal of the American Chemical Society	1900 – 1.36 1950 – 2.35 2000 – 4.30	1900 – 107 articles / 414 p 1950 – 1415 articles / 5891 p 2000 – 1298 articles /13,040 p

Maniatis, P., Rosenthal, D., Roussopoulos, M., Baker, M., Giuli, T. J., & Muliadi, Y. (2003). Preserving peer replicas by rate-limited sampled voting. *ACM Symposium on Operating Systems Principles archive – Proceedings of the nineteenth ACM symposium on Operating systems principles table of contents*. Retrieved February 17, 2005 from http://portal.acm.org/

This is a very technical article describing in great detail the peer-to-peer opinion poll protocol recently implemented by the LOCKSS system (Lots of Copies Keeps Stuff Safe). LOCKSS uses a cooperative independent network of low cost persistent web caches to create a distributed ejournal archive. Based on a Byzantine-fault-tolerance model, the system polls a set of peers to determine who has the best copy of the file. In early version of LOCKSS, this process was very slow and did not scale either for additional peers or ejournals. LOCKSS has a set of design principles:

1. Cheap storage is unreliable (so have lots of copies at lots of peer servers)
2. No long-term secrets (long term secrets like private keys break easily)
3. Use inertia (prevent change to the data and be prepared to repair slowly)
4. Avoid third party reputation (don't rely on the last server that fixed your problem)
5. Reduce predictability (make it harder for malicious hackers to hurt you)
6. Intrusion detection is intrinsic

7. Assume a strong adversary.

The new polling protocol maintains a list of previously polled peers. This list is "churned" regularly so that the peers are evenly polled enforcing design principles 4 and 5. The protocol requires a proof of effort, a numeric computation, to prove that the peers have actually validated their version of the data and to discourage malicious peers enforcing design principles 2, 3 and 7.

The protocol itself is relatively simple. There is a poll initiator and a poll responder or voter. The poll initiator has an "inner circle" of peers that it sends the request to. These peers can then pass the request to other peers creating an "outer circle" of voters. After the poll has been completed, the initiator can request a repair for its journal using a bulk transfer protocol. The communication between peers, either for voting or repairing, is encrypted via symmetric session keys, derived using Diffie-Hellman key exchanges. After the transaction, the key is discarded. LOCKSS has great potential for being a major piece of the ejournal archiving for many years to come.

Nisonger, T. E. (1997). Electronic journal collection management issues. *Collection Building, 16*(2), 58–65. Retrieved January 20, 2005 from http://taddeo.emeraldinsight.com

This article is one of the first published on the topic of ejournals and preservation. While most of the article is concerned with the selection, acquisition, processing, evaluation, and management of ejournals, the author lays out the essential questions that still do not have clear answers:

- Which ejournals should be archived?
- What format should be used for archiving?
- Who will archive ejournals?

The author asserts that policy needs to be developed and formalized in a written document. He concludes by stating that libraries have dealt with new formats throughout history – from books to serials to microfilm and now to networked resources. "…[You] should have fun being a librarian because you are probably not going to get rich being a librarian."

Pearson, D. (2001). Medical history for tomorrow – preserving the record of today. *Health Information and Libraries Journal*, 18. Retrieved January 17, 2005 from www.blackwell-synergy.com/www.blackwell-synergy.com/

The future of the history of science depends on how well librarians and archivists preserve today's electronic records, ejournals and other digital scientific information sources. Using two case studies, the author effectively shows the issues. In 1994 Harold Cook published a book about a Dutch doctor in 1694 who went to prison because he prescribed a certain medicine to cure a patient. Becoming a medical "cause célèbre," the doctor was eventually exonerated. Cook was able to research this case using the medical literature of the time and was able to place it in the "context of the standard surgical and therapeutic practices of the time." Cook had the writings of the doctor himself, newspapers, medical text books, the medical law texts resulting in a 9 page bibliography.

The second case occurred almost 300 years later. Another "cause célèbre," the case of Jaymee Bowen, more commonly known as Child B, was a media circus. The child had leukemia. After years of treatment, including a bone marrow transplant, the child had only several months to live. The UK National Health Service denied the father's demand for another bone marrow transplant. After losing in the UK court system, an anonymous donor funded more chemotherapy and a new experimental treatment that prolonged the child's life for another year. For a future researcher to complete a similar history of this case, access to the Nation Health Services patient records will be required – the GP, the specialists, the private physician, the hospitals will all be part of the records. Will they be available? To find the treatment protocols, a Medline search produced over 30.000 citations. With an Internet search engine a search for leukemia produced a result set of 150,000 pages. (Over the course of writing the article, Pearson states that the number of WebPages increased 35%) The researcher will be overwhelmed. One has to ask, however, if there are not textbooks that would synthesize and summarize all of these articles and websites for the future history of science researcher.

Regardless of the availability of a summary text, a number of questions are still important. What of this current data should be retained? Who will be making the curatorial decisions and preservation investments? Pearson suggests that libraries and archives work together to preserve "the contemporary record for tomorrow."

Pinfield, S., & James, H. (2003). The digital preservation of e-prints. *D-Lib Magazine, 9*(9*).* Retrieved January 17, 2005 from http://www.dlib.org/dlib/september03/pinfield/09pinfield.html

Should e-prints be preserved? This is the question. The authors define "e-prints" as "electronic versions of research papers or similar research output," either pre-prints or post-prints. A number of open-access online repositories are often referred to as "archives" (sometimes because of the misnamed Open Archives Initiative Protocol for Metadata Harvesting), but this term does not "necessarily imply a curation or long-term preservation function."

The case against worrying about long-term preservation of e-prints (a summarization of the opinions of Stevan Harnad):

- e-Prints are duplicates of published literature.
- e-Prints repositories were created to provide immediate, free access.
- The focus should be on filling repositories.
- With the focus on volume, preservation will be a distraction
- Preservation will be a barrier because of all of the submission requirements that authors will have to deal with.
- Preservation efforts should be focused on the conventionally published materials that are truly endangered.

The case for preserving e-prints (a summarization of the opinions of Peter Hirtle):

- Preservation is necessary to continue to provide open access.

- Preservation is necessary to ensure citations will continue to be valid.
- e-Prints have additional information that will be lost if e-prints are not preserved.
- Preservation is important when a repository has a coherent collection of e-prints.
- Guarantees of preservation may attract authors.

The authors propose that both are possible. Filling repositories is important, and efforts to preserve the data can be accomplished as well. Describing their project, UK SHERPA, the authors assert that libraries are the "natural home for [preservation] since they have a tradition of managing access to information resources and preserving them into the future." But libraries cannot do this important work without the support of the entire academic community.

Preserving e-prints is not as technically difficult as other digital objects. Because they are "paper documents made electronic," they normally contain only text and still images. It will be important for repository managers to know versions of HTML and types of images. Over time, e-prints will evolve to include moving images, sound and more external links. These issues will need to be resolved. But rather than focusing on the technical issues, the authors addressed many of the social and organizational issues. Institutional commitment is vital as the scholarly communications process shifts. Funding needs to become a stable stream, and costs need to be better understood. Preservation appraisal is important because digital preservation, just like physical preservation, is not an "all or nothing" proposition.

This article presents a refreshing pragmatic approach to digital preservation. Understanding the tradeoffs between access and preservation, costs and benefits, "good enough" and perfection is the only way to make appropriate decisions

Reich, V., & Rosenthal, D. (2004) Preserving today's scientific record for tomorrow. *BMJ: British Medical Journal*, *328*(7431) Retrieved January 20, 2005 from http://www.bmjjournals.com

This article opens with a quote from Thomas Jefferson: "Let us save what remains, not with vaults and locks which fence them from the public eye and use in consigning them to the waste of time, but by such a multiplication of copies, as shall place them beyond the reach of accident." This is the theoretical basis of the Stanford University/Hotwire Press ejournal archiving project called LOCKSS – Lots of Copies Keeps Stuff Safe. With the premise that distributed copies reduce the likelihood of permanent loss, the LOCKSS system has a mechanism to distribute multiple copies of electronic journals on very low cost hardware linked via the Internet. Ejournal copies are routinely checked for damage. If a file is damaged, the server queries other servers to find an undamaged copy. The damaged copy is then repaired. This system is in production with 80 libraries and 50 publishers of academic journals participating.

Rosenthal D., Lipkis T., Robertson T., & Morabito S. (2005). Transparent format migration of preserved web content. *D-Lib Magazine, 11*(1). Retrieved January 20, 2005 from http://www.dlib.org/dlib/january05/rosenthal/01rosenthal.html

Yet another article on LOCKSS. This article describes the format migration capabilities of the LOCKSS system and provides an excellent summary of the possible strategies for keeping data viable over time. For years, the big issue has been emulation vs. migration. As more applications have become web enabled, emulation is no longer an option – with the majority of the functionality of a system on the client platform, the environment if out of the control of the preservation system. Within the migration path, a number of options are discussed:

1. Migration on ingest – done when the data is newest and is well known. Does not preclude the need for future format migration and might still need either batch or migration on access.
2. Batch migration – the most common working model (let's move everything from gif to jpg) Computationally expensive. Just in case rather than just in time. But it is done and the knowledge of the specific process does not need to be kept current.
3. Migration on access – seems to be less computationally intensive, Just in time model rather than the just in case of the batch model. But migrations can compound. Operational knowledge of the migration process must be maintained over a longer period. The migration process itself might need to be migrated to a new OS or programming language if the process extends over years.

LOCKSS will use a migration on access model. Format conversion processes will register the input and output MIME types and the LOCKSS Web proxy code will invoke them at access. It will also use a feature of http headers to compare Accept: and determine if the format is acceptable. If not, LOCKSS will search for a conversion program as above and perform the conversion. The registry of converters will be distributed via the normal LOCKSS polling techniques. The migration process was tested by successfully converting a number of GIFs to PNG.

Waters, D. (2002). Good archives make good scholars: Reflections on recent steps toward the archiving of digital information. Retrieved February 17, 2005 from http://www.clir.org/pubs/reports/pub107/waters.html

Don Waters is undoubtedly the best writer working in Library and Information Science today. He brings a lyricism to ejournal archiving that one could never anticipate. Using Robert Frost's poem "Mending Walls" as a paradigm, he bridges from "good fences make good neighbors" to good archives make good scholars. Rather than being a clumsy metaphor, mending fences frames the issues – the stone fence falls apart over the winter; neighbors find the flaws; they work together to fix the problem; they go back to their normal activities. Like the neighbors in the poem, libraries, publishers and scholarly communities are a community with common boarders with shared responsibilities for maintaining their common resource. He summarizes the findings of the seven Mellon foundation ejournal archiving grants (see above)

1. Archiving seems technically feasible in multiple modes (the LOCKSS model of harvesting web pages and the publisher source file capture models)
2. Publishers are coming to view archiving journals as a competitive advantage
3. Ejournal archives will make it possible to view ejournals as the publication of record and to allow libraries to consider abandoning print.

He spends a fair amount of time discussing the political economy of public goods, economic models, and organizational options. He concludes that a mixed model will emerge – a combination of government, publisher, and university funding. "…what makes good neighbors is the very act of keeping good the common resource between them – the act of making and taking the time together to preserve and mend the resource. So too it is with digital archiving."

Waters, D., &Garrett, J. Eds. (1996). *Preserving digital information: Report of the task force on archiving of digital information.* Washington, D.C. and Mountain View, CA: The Commission on Preservation and Access and the Research Libraries Group. Retrieved March 4, 2005 from http://www.rlg.org/ArchTF/

This is the seminal paper on digital preservation. Commission by the Commission of Preservation and Access and the Research Libraries Group, this paper has framed the discussion of digital preservation for the past decade and will continue to do so for the next decade. While many of the issues surrounding digital preservation had been addressed by the business community, none of the issues were framed with a library time line. Beyond its own operational needs, a business maintains electronic data for regulatory purposes. Often a business is required to destroy the file after the mandated time has passed. Libraries and other archival agencies have no such time limits. Describing the challenges of archiving digital information, the article discusses, technological obsolescence, migration issues, legal and institutional issues, infrastructure needs, and develops a conceptual framework.

The section of "Information Objects in the Digital Landscape" is, I think, the core of the paper, laying out the functional requirements for digital repository and archives that has yet to be bettered or implemented. "For digital object, no less than for object of other kinds, knowing *how* operationally to preserve them depends, at least in part, on being able to discriminate the essential features of *what* needs to be preserved…Whatever preservation method is applied, however, the central goal must be to preserve information integrity: that is, to define and preserve those features of an information object that distinguish it as a whole and singular work… including: content, fixity, reference, provenance and context."

A primary concept in this paper is that of the "certified digital archive", one that has been accredited by an outside agency. This idea is alluded to in other writings but has not been championed by the Library of Congress, OCLC, RLG, ARL or any of the other major library organizations. Certification of digital archives is not likely to be a reality for many years to come.

Wusteman, J. (2003). XML and e-journals: the state of play. *Library Hi Tech, 21*(1), 21–33. Retrieved January 20, 2005 from http://hermia.emeraldinsight.com/

This article is a survey of the state of XML technology of the STM publishing community in 2002. The library and publishing communities were beginning to think about using XML as both a metadata and a journal archiving format. At the time of the survey, most of publishers used SGML as an internal processing format. While several "standard" DTDs had been developed by the library and publishing communities, none of the publishers used these "standards". Most had used a variation of one of the standard DTDs. Very few of the publishers had ventured into XML. Those who had merely converted their SGML DTD into an XML DTD. The author asserts that XML Schema, a proposal of the WC3, was too expensive to implement and would have a slow adoption rate. While the movement is slow, the author concludes that XML will likely be the future of publishing and ejournal archiving. One could imagine that a similar survey taken in mid-2005 would find a different state of the market with many publishers using XML and schema for their internal processing format and with the ability to export in other schemas or even DTDs with little effort.