Archiving scholarly material: one publisher's perspective

Based on a paper presented at the 29th UKSG Conference, Warwick, April 2006

We need large digital archives to preserve scholarly content for generations to come, and we need them now. Deciding who should do the archiving, what to archive, where in the scholarly community archiving should be undertaken, and how to do it successfully, is starting to come into focus. A brief review of the scholarly digital archiving efforts underway reveals some commonality of intent. This can be contrasted with a clear need for decisions regarding the scope of what should be digitally archived and, ultimately, how archives in their many forms might be best constructed. While exploring how we might archive digital material, we can turn our attention to which critical success factors to consider in adopting any archiving solution. These factors include their governance (community), economic stability, technical soundness, their community acceptability, and ease of use. Several current solutions are examined within this framework.

GORDON TIBBITTS
President
Blackwell Publishing

Introduction

Every day, more forms of digital scholarly content are being created and disseminated[1]. In the case of the scholarly journal and/or book (Type 1 content), having a digital copy with links is more the norm than ever. Short-term access to these materials is made easy through research libraries and institutional networks. Yet, the research material (Type 2 content) that supports these works, and the many related activities of scholars and researchers tying together pre- and postprints, reviews, lecture material, blogs and data, are not generally available online. Institutional repositories have honorably stepped into this void. They have helped to offer some medium-term solutions for storage of the supplemental material through inter-institution access, though links out to the books and journals are lacking. Open source discovery tools and indices are supplementing this void as well, though these too are in the early stage of being globally co-ordinated. Discovery engines (Google, MSN, Yahoo, A&I services and others) are providing more semi-permanent links between these items – in the case of Google equivalents, very temporary links. A basic challenge facing the scholarly community in this digital age is what we are going to do about digital archiving. This paper will discuss types of scholarly content we might archive, the emerging archiving solutions, and the critical success factors affecting their long-term viability.

What should we archive?

A philosophical debate on this subject might be enjoyable and perhaps would result in the call to store all information, everywhere, 'in every Planck instant' (at every miniscule moment), so as not to lose any relevant information. Regardless of the fact that this would violate Heisenberg's Principle (since it is impossible to know the positions and directions of the particles in the data store at the same time), the value of contemplating such massive storage of everything is out of our grasp. More practical decisions will need to be made. Commencing with the digital copies of all scholarly journals and books (Type 1 content above) is a reasonable start, and extending this to the material used to create these works is also a reasonable aspiration (Type 2 content above). Finally, some new work should be undertaken to capture the intellectual discourse

(Type 3 content) surrounding scholarly works, such as blogs, LMS (learning management systems), lecture notes, social networks[2,3], conferences, podcasts, message boards and web-interactive seminars. Defining what Type 3 content might entail is still a work-in-progress and standardization of the many initiatives underway needs to be improved. One important area of focus for Type 3 content will be the advertisements, classifieds, job postings and other supplementary works not currently being considered for archiving.

Methods to enable database interoperability and learning are spawning standards and tools as well[4]. The digital scholarly discourse will need to be parceled somehow; defining the beginning and end of a unit of exchange between scholars, clinicians and scientists is without definition today. Finally, there is a desperate need for permanence of links. URLs (even those in this paper) are fleeting; DOIs are a potential solution for permanent objects. However, how will Type 3 content be referenced? Once there are more standards for discourse (Type 3 content), it is likely that they, too, will be archived.

A note on the need for digital archives to ingest the scholarly material

In this age of interoperability and linking, it seems impractical that all information in the scholarly digital universe worthy of long-term storage should be part of a 'central-like' archiving solution. Most archiving activities and many long-term preservation activities, including conversion, testing and format migration through the ages[5], make it almost totally necessary to seek a solution that brings information in, separating it from the original data store. The types of scholarly content described in this paper are considered ingested into an archive separate from their original data store.

It is critical to note that not all scholarly content can be archived in this way. Initiatives like the Sloan Digital Sky Survey http://www.sdss.org/, the Whitehead Institute-MIT BioImaging Center, and the Computational and Systems Biology Initiative (CSBi) http://csbi.mit.edu/, require such a vast amount of storage that data is virtually non-transferable! Solutions for non-transferable content archives will have to be solved outside and in addition to the approach described in this paper.

Where are the current archiving solutions and how do they work?

National

One class of solution is the national archive. A good example would be the approach used by the National Library of The Netherlands *[see Appendix]*, which ingests material using best-practice guidelines and follows strict copyright owner requirements for access. They allow scholars on-site access and the costs of the site are somewhat absorbed by the government, though they have not yet described a method to allow access for libraries and their patrons should a catastrophic event occur. The British Library Digital Library System *[see Appendix]* also provides a similar national archive approach. They recover operating costs through a business-like approach which requires some consideration.

Product

Private not-for-profit and corporate for-profit organizations are offering archiving product solutions. The customers in these cases are separate from the organizations running the product solutions. A review of current publishers offering archiving product solutions (based on a limited amount of publicly available information) reveals that at present they fall somewhat short of providing comprehensive archiving solutions[1]. They seem to be focused on more commercial content-provision products rather than archives. Archiving product solutions provided by the private not-for-profit sector propose a more comprehensive archiving solution than publisher-based solutions. Portico *[see Appendix]* is one solution of this type. By accepting content using the open standards journal archiving and interchange DTD[6], Portico migrates content to a standard format more suitable for long-term archiving. As steward, Portico maintains the content, not the business systems and related platforms, and makes all decisions regarding production, compression, storage, file format, data migration, and the distribution and provision of archival versions. Upon defined trigger events for *authorized users* (only institutions – libraries and publishers – participating in Portico), the content can be made available for non-commercial use. Publishers retain the right to terminate the agreement but must leave all deposited material irrevocably in the archive.

Run your own

Institutions (such as university libraries, corporate and pharmaceutical libraries) are also building or buying and installing archives and archive software to operate 'run-your-own' archives. Current solutions include DSpace, EPrints and Fedora *[see Appendix]* as well as self-built varieties. At this time, the content stored is mostly of the Type 2 variety (Type 3 content is still not widely uniform or managed). Another type of solution available and built for Type 1 content (with the capability for more than Type 1 content) is the LOCKSS (Lots Of Copies Keeps Stuff Safe) solution *[see Appendix],* initiated by Stanford University Libraries. More than 150 institutions have started using LOCKSS so far. LOCKSS collects licensed content by connecting itself to library-run e-content delivery platforms. It continually compares its digital copy with other LOCKSS installations worldwide that have identical content, utilizing a shared global manifest system to repair any damage in its distributed network. It then acts as a web proxy or cache providing browsers in the library community with access as appropriate.

Community

Then there are community (-built and -managed) initiatives. Though global indexes and initiatives that link the silos of run-your-own solutions might evolve to be of this variety, these efforts are still nascent. One initiative that has just entered community solutions' space is called CLOCKSS (controlled LOCKSS) *[see above and Appendix].* The technology is based on the already-established and successful LOCKSS archiving technology, and the initiative is managed by a joint board of both librarians and publishers consistent with the JISC recommendations emphasizing community collaboration as key to the success of an archive[7]. The promise of success with this approach is that there are no single organizational ties (of either a not-for-profit or for-profit nature) that could compromise the long-term financial and operational viability of this initiative. At the outset there is a broad base of publishers committed, as well as foundational libraries supporting its development. Upon defined trigger events for *all users worldwide, not restricted to funding members of the CLOCKSS initiative*, the content can be made available for non-commercial use. Publishers retain the right to terminate the agreement but must leave all deposited material irrevocably in the archive.

How do archiving solutions work?

Keeping archiving requirements restricted to the content types described, a comparison of these archiving approaches (national, product, run-your-own, and community-based) can be analysed. The aspects that seem most salient to compare are *governance* (the level of community control and access), *economic stability*, *technical soundness* and, ultimately, the *community acceptability*.

National archives may ultimately serve the entire need. They may also end up being the most comprehensive archiving solution for scholarly content when there are no other avenues available (due to technical know-how, awareness of alternatives, and/or general ability of content providers). National archives may also introduce levels of censorship and access control that do not align with the level of freedoms to which scholars are accustomed. For instance, what might a US national archives access policy be for subjects such as stem cell research, women's rights, or foreign policy? Furthermore, it is unclear if content owners will cede rights to national archives allowing for access post-subscription, or due to orphaned or intentionally 'open' content. Over time, national archives, enabled with market-borne solutions, have the ability to make technical solutions work. (The Apollo missions and the Great Wall of China come to mind.) Communities are likely to accept national archives (cautiously) and will facilitate their continuance.

The governance of product solutions initiatives is entirely in the control of the selling organization. They remain in complete control of access, funding models and accountability. If priorities change, there are no barriers that prevent rate hikes and access restrictions from being implemented. In general, content providers participating in these solutions are not giving up rights in any way, which leaves questions of how they will serve the community with regards to access. Organizations tend to change continually. In the case of publishers, they are bought and sold. In the case of not-for-profits, they can change their charter at will. Providers of these solutions can charge 'over-public-good' rates for their services and eventually become monopolistic in nature, or they could be benign purveyors of the community. The important fact is that they, not the community, are in charge of this destiny. As with governance and business models, the technology can be shown or hidden from the community at will. In the case of publishers, the library community should consider what rights they have in using 'product' archiving solutions. Some publishers offer 'big deals' (providing access to large collections of content), which do not include archival rights. Though this is hopefully a dying practice, it is one more reason to listen cautiously when a publisher uses the word 'archive' while meaning the sale and use of an online content provision product[8]. Other product solutions have not yet proven themselves either technically or financially viable, and therefore the jury remains out as to community acceptance.

Run-your-own archives are governed by a variety of institutions and organizations. Though open source discovery tools, global catalogues and arranged sharing methodologies may eventually form into an archive serving all of the needs[7] of the scholarly community, it will be challenging to break the silos of this approach. A more organized approach, such as the LOCKSS initiative, does provide local governance and compliance with content provider access terms. Generally, this includes a restriction on content access through IP restrictions and a manifest of authorized content. This solution provides most of the archiving requirements[8]. However, it does not provide the scholarly community who are not using LOCKSS with access to content. Many publishers are not participating because of concerns about the immediacy of the access trigger (which can be the loss of a subscription), and believe it still needs more robust access security.

Run-your-own archives have proven their technologic worthiness (leaving some remaining concern over security). The costs of these solutions are also fairly low, though many institutions do not take into consideration the FTE charges and opportunity costs when considering run-your-own solutions – they are not no-cost solutions. The broad use of these solutions reinforces that they are widely accepted by the library and institutional market (and some governmental organizations are experimenting with them).

Finally, some emerging solutions are addressing the challenges of governance by launching with community governance in the first place. For example, the CLOCKSS initiative includes a board from both the content provider and library community. Most archiving solutions have trouble in being both comprehensive and robust enough. This initiative is starting out with the majority of large publishers (and their content) and a foundation of well-established and renowned libraries to form the core governance team and to establish public-good rates, oversee technical robustness, and hopefully provide a level of community acceptance. This solution remains unproven though the strong start, and testing over the next two years, will help demonstrate its viability.

Initiative type	**Governance and content censorship**	**Cost level for publishers/ institutions and risk**	**Technical soundness**	**Access control**	**Community acceptance**
national/ government	possible	minimal/ minimal	reasonable/ high	content rights holders can deny access	acceptable
private/public not-for-profit & corporate (archiving products and run-your-own)	moderate (single administrative control)	mid to high/ unbounded	unknown	content rights holders can deny access	mixed
community collectively run	no (multiple administrative control)	low/ minimal	CLOCKSS is LOCKSS-based and somewhat proven	Community-managed, broad access rights ceded in the case of orphaned and non-accessible content via the publisher's site	likely to be acceptable

A note on access and ease of use

All archives need to provide access to content at salient points in time. The long-term requirement of an archive vs. a content provision service is to ensure that the content is preserved and available. So, access for an archive is more about getting the content to a place where it can be provided rather than what one might call being greatly functional. In these early days of developing a long-term digital archive for scholarly content, there is much confusion about these concepts. Frequently, archiving providers talk about how well their archive services content. Practically, a level of ease of use (the ability to search, index, reference and display) needs to be built into an archive. For instance, LOCKSS agrees to display the content in a form similar to that in which it was rendered through the content provider's website or library interface.

A healthy development regarding access is the recent consideration of separating the functions of access and delivery. An archive is not a profit-making tool. It is not a content provider's website. It is not a place to do data-mining (perhaps?). It can be a location for all that has been preserved. It can be a place to get content which has been lost, purposely or voluntarily abandoned, or orphaned, or open. And it can be a point where services can access those scholarly items that they have rights to, so they in turn can ingest the content and provide easy-to-use functionality.

So what is next?

Which is more likely?

- Many approaches are believed to be the best, so that no solution (or solutions) will emerge in the long run.
- Or, an assortment of solutions will emerge, serving the scholarly community in an economically feasible and comprehensive way.

Of course we all believe the latter; however, we have made it hard to move forward on several fronts and we will need first to answer a few questions. First, if we are to build a comprehensive archiving landscape, perhaps built on several layers of services, is it not best to have a solution that allows for community-run governance and access? If we are to tie solutions together, linking national, run-your-own, private and community solutions, should we not evaluate the cost components and business models? Should we not agree that they should not be governed (or dominated) by any one self-interested party?

We already agree that technological solutions should be open to scrutiny. If the open source model is not feasible for sustaining a technological solution, then there still should be some oversight to ensure long-term viability by a community-led and supported technology team. Finally, content providers are protective of the content and intellectual product they manage for authors, societies and other owners, so regardless of the outcome of the open access debate, it is likely that some scholarly content will need access control to allow the scholarly content economy to continue to flow. In this light, should we not seek solutions that aim to meet both content provider access concerns and librarian and archivist needs for a long-term archive in a unified and comprehensive way? There are a lot of decisions yet to be made and I believe we are nearing a good time for archiving; we might actually start to build a comprehensive scholarly archive that meets all the needs of the community.

Annotated Appendix related to scholarly archiving

Access
AARL position on open access:
http://www.arl.org/scomm/open_access/framin.html

Why is access to information important?

- Society benefits from the open exchange of ideas. Access to information is essential in a democratic society. Public health, the economy, public policy, all depend on access to and use of information, including copyrighted works.
- Access to copyrighted materials inspires creativity and facilitates the development of new knowledge.
- Intellectual property is the lifeblood of progress in the sciences and arts.
- New knowledge is developed from existing information. Authors build on the intellectual products of others to create new works.

- Copyright exists for the public good. Copyright was intended to serve the public interest by encouraging the advancement of knowledge while protecting the rights of authors and copyright owners. It is meant to balance the competing interests of creators, publishers, and users, not stifle the free flow of information.
- Federal investment in R&D is leveraged by access to research results. The US federal government spent close to $50 billion on non-defense-related R&D in 2002. The government depends on the dissemination of the results of that research as a stimulus to further economic, scientific, medical, and environmental development.

DC Principals for Free Access to Science:
http://www.dcprinciples.org/

March 16, 2004 – Washington, DC – As scholarly, not-for-profit publishers, we reaffirm our commitment to innovative and independent publishing practices and to promoting the wide dissemination of information in our journals. Not-for-profit scientific, technical, and medical publishers are an integral part of the broader scholarly communities supporting scientists, researchers, and clinicians. We work in partnership with scholarly communities to ensure that these communities are sustained and extended, science is advanced, research meets the highest standards, and patient care is enhanced with accurate and timely information.

We continue to support broad access to the scientific and medical literature through the following publishing principles and practices.

1. As not-for-profit publishers, we see it as our mission to maintain and enhance the independence, rigor, trust, and visibility that have established scholarly journals as reliable filters of information emanating from clinical and laboratory research.
2. As not-for-profit publishers, we reinvest the revenue from our journals in the support of science worldwide, including scholarships, scientific meetings, grants, educational outreach, advocacy for research funding, the free dissemination of information for the public, and improvements in scientific publishing.
3. As not-for-profit publishers, we have introduced and will continue to support the following forms of free access:
    - Selected important articles of interest are free online from the time of publication;
    - The full text of our journals is freely available to everyone worldwide either immediately or within months of publication, depending on each publisher's business and publishing requirements;
    - The content of our journals is available free to scientists working in many low-income nations;
    - Articles are made available free online through reference linking between these journals;
    - Our content is available for indexing by major search engines so that readers worldwide can easily locate information.
4. We will continue to work to develop long-term preservation solutions for online journals to ensure the ongoing availability of the scientific literature.
5. We will continue to work with authors, peer reviewers, and editors for the development of robust online and electronic tools to improve efficiency of their important intellectual endeavors.
6. We strongly support the principle that publication fees should not be borne solely by researchers and their funding institutions, because the ability to publish in scientific journals should be available equally to all scientists worldwide, no matter what their economic circumstances.
7. As not-for-profit publishers, we believe that a free society allows for the co-existence of many publishing models, and we will continue to work closely with our publishing colleagues to set high standards for the scholarly publishing enterprise.

FAIR

http://www.jisc.ac.uk/index.cfm?name=fair_synthesisintro Focus on Access to Institutional Resources specifically focusing on access issues. (This is a JISC initiative.)

Google Scholar

http://scholar.google.com/ GS provides a simple way to broadly search scholarly material.

PALS

http://www.palsgroup.org.uk/ Collaboration (UK Publishers and JISC) around issues with electronic publishing.

Author archiving

Stevan Harnad, Department of Electronics and Computer Science, University of Southampton, Highfield, Southampton SO17 1BJ, United Kingdom, 'For Whom the Gate Tolls? How and Why to Free the Refereed Research Literature Online Through Author/Institution Self-Archiving, Now.' http://cogprints.org/1639/01/resolution.htm

Archives and Institutional Repositories et al

British Library Legal deposit: the United Kingdom requires deposit of each publication. The British Library Legal Deposit system http://www.bl.uk/about/policies/legaldeposit.html facilitates this.

CLOCKSS

http://www.lockss.org/clockss/ is a collaborative initiative by a group of organizations drawn from publishers, libraries and learned societies. Uses LOCKSS technology and a social model to support a 'large dark archive' that is both fail-safe and has an acceptable process for providing continuing access for orphaned materials.

DSpace

http://www.dspace.org/ The DSpace digital repository system captures, stores indexes, preserves, and distributes digital research material. See also D-Lib paper, 'Best Practices for Digital Archiving' *D-Lib Magazine* Jan 2000 for introduction http://www.dlib.org/dlib/january00/01hodge.html (MIT Library initiative.)

EPrints

http://www.eprints.org/ Supporting open access and archiving and offering Eprints Free Software for archiving.

Fedora

http://www.fedora.info/ An open source software for managing and delivering digital content. (A Cornell and University of Virginia initiative with Andrew W Mellon Foundation support.)

FIGARO

http://www.figaro-europe.org/history.html stands for the Federated Initiative of GAP and Roquade, representing the Dutch and German universities collaborating to establish an infrastructure for academic electronic publishing in Europe.

LOCKSS

http://lockss.stanford.edu/ Lots of Copies Keeps Stuff Safe archiving initiative established with over 150 institutions. LOCKSS is open source software designed to provide librarians with an easy and inexpensive way to collect, store, preserve, and provide access to their own, local copy of authorized content they purchase.

MUSE

http://muse.jhu.edu/about/muse/overview.html Affordable collection of humanities, arts and social sciences journals with online access. (Johns Hopkins University Press and the Milton S Eisenhower library with grants from the Andrew W Mellon Foundation and the National Endowment for the Humanities.)

Portico

http://www.portico.org/ The Portico electronic archiving service is an initiative to provide a permanent archive. (An electronic archiving initiative launched by JSTOR in 2002.)

National Library of The Netherlands (KB, Koninklijke Bibliotheek)

http://www.kb.nl/menu/bibliotheek-en.html

SPARC IR

http://www.arl.org/sparc/IR/ir.html#exec Proposed potential alternatives to traditional journals.

Copyright and facilitating access

Creative Commons

http://creativecommons.org/ Assistance and administration of self-archiving and automatic sharing of rights.

EDItEUR

http://www.editeur.org/ Co-ordinating the development, promotion, implementation of electronic commerce in the book and serials sectors.

The Zwolle Group, copyright management for scholarship:

http://www.surf.nl/copyright/

Metadata

JISC on Metadata uses:

http://www.jisc.ac.uk/index.cfm?name=fairsynthesis_metadata

OAI

http://www.openarchives.org/ Open Archives Initiative – the promotion of interoperable standards.

OMI-PMH

http://www.openarchives.org/OAI/openarchivesprotocol.html Open Archives Initiative Protocol for Metadata Harvesting.

Research

LIFE

http://www.ucl.ac.uk/ls/lifeproject/ Life-Cycle Information For E-literature – a project looking at the life-cycle of the collection and preservation of digital material.

SHERPA

http://www.sherpa.ac.uk/ Securing a Hybrid Environment for Research Preservation and Access is a three-year project funded by the JISC and CURL and hosted by Nottingham University. It aims to address issues surrounding the future of scholarly communication and publishing by creating a network of open access repositories to release institutionally-produced research findings onto the web.

References and notes

1. Best Practices for Digital Archiving, An Information Life Cycle Approach, *D-Lib Magazine,* January 2000, 6 (1). http://www.dlib.org/dlib/january00/01hodge.html
2. Social Bookmaking Tools (II), A Case Study – Connetea, Ben Lund, Tony Hammond, Martin Flack, Timo Hannay, *D-Lib Magazine,* April 2005, 11 (4).
3. Koku, E.F. and Wellman, B., *Scholarly Networks as Learning Communities: The case of TechNet*, University of University of Toronto, January 2002; Also *Designing Virtual Communities in the Service of Learning*, Barab, S.A., Kling, R. and Gray, J.H., Cambridge University Press, June 2004, DOI: 10.2277/0521520819
4. *Digital Library Management and Course Management Systems: Issues of Interoperability*. Report of a Study Group, Co-Chairs Dale Flecker Assoc. Dir. For Planning & Systems Harvard, McLean, N., Dir. IMS, Australia, July 2004. http://www.diglib.org/pubs/cmsdl0407/
5. The LOCKSS initiative does a good job describing the generally accepted features and functions required by digital archives: Collecting, Preserving and Auditing, Providing Access, Administering. http://lockss.stanford.edu/works/how_it_works.htm
6. http://dtd.nlm.nih.gov/
7. A Continuing Access and Digital Preservation Strategy for the Joint Information Systems Committee (JISC) 2002-2005 http://www.jisc.ac.uk/index.cfm?name=pres_continuing
8. Also see ALPSP guidelines for when a journal changes publishers http://www.alpsp.org/socjourn1.pdf
   Also, the JISC/NESLI model license clearly spells out the need to keep 'perpetual access'.

■ Gordon Tibbitts
President
Blackwell Publishing, Inc
350 Main Street
Malden, MA 02148, USA
E-mail: gtibbitts@bos.blackwellpublishing.com