

Developing an E-Repository for Trusted Backup and Orphaned Content

A Public Knowledge Project
Proposal to
The University of British Columbia Library

Submitted by

Chia-ning Chiang
Student, SLAIS
Public Knowledge Project

and

John Willinsky
Professor, Language and Literacy Education
Public Knowledge Project



June 2002

Developing an E-Repository for Trusted Backup and Orphaned Content

Summary of the proposal objectives and methods

Born digital journals that only exist in electronic format are vulnerable to rapid organizational, technical, and economical changes that characterize today's information environment. For research libraries, the long-term preservation of digital collections may well be the most important issue in digital libraries.

The Public Knowledge Project (PKP) at UBC is developing and testing the Open Conference Systems, Electronic Theses and Dissertations, and Online Journal Systems, etc. In doing so, PKP seeks to improve the scholarly and public quality of academic research. These systems are designed not only to assist in the management and publishing of scholarly work, but also to improve the indexing of research in online environments and create a richer context of connections for any given study, connections both within the scholarly literature and to the larger world of related online information. In providing publishing systems for conferences and journals, one responsibility for PKP that arises is the provision of archiving resources that will ensure that what is published is preserved no matter what happens to the journal or conference websites. PKP is thus approaching UBC Library with the idea of developing a long-term, inactive repository or archive for digital scholarship in need of reliable preservation.

The goal of this Developing E-journal Repository proposal is to develop a model for permanent e-journal archives to ensure the availability of these important scholarly publications for the future generations. There are two functions of such a system. It will need to provide what is known as Trusted Backup for scholarly publications, as well as access to "orphaned content," resulting from publishing sites that can not be sustained. These two functions will need to serve with the repository/archives of open access materials, beginning but not restricted to set up a pilot archives composed of scholarly papers from the PKP systems.

Introduction

Electronic journals, particularly the born-digital ones, raise numerous new issues about preservation that place new demands on digital libraries.¹ In the world of physical research materials, a great number of valuable research resources have been saved passively; acquired by individuals or organizations and stored in little-visited recesses. For digital research materials, changes in computing technology will insure that, over relatively short timeframes, both the media and technical format of old digital materials will become unusable. Keeping digital resources accessible for use by future generations will require conscious effort and continual investment.²

More spending on electronic resources, according to the investigation of Canadian Association of Research Libraries³ and Library Journal's 2001 academic library book buying survey,⁴ indicated academic libraries have dramatically increased their offerings of online resources. A survey of the 21 members of the Digital Library Federation revealed that 40% of their costs for digital libraries in 2000 went for commercial content.⁵ The big-ticket items were electronic scholarly journals that libraries license rather than own. Yet little direct evidence shows that publishers have developed full-scale digital preservation capabilities to protect this material, and research libraries continue to purchase the print versions for preservation purposes. However, none appears ready to forgo access to the licensed content just because its long-term accessibility might be in question.⁶

¹ Cornell University Library Project Harvest: developing a repository for e-journals, http://cidc.library.cornell.edu/about/project_harvest.htm ; The Commission on Preservation and Access and the Research Libraries group, Inc. Preserving digital information: Report of the Task Force on Archiving of Digital Information. May 1, 1996. <http://www.rlg.org/archtf/tfadi.index.htm>

² Dale Flecker "Preserving scholarly e-journals," D-Lib Magazine, 7:9, September 2001, <http://www.dlib.org/dlib/september01/flecker/09flecker.html>

³ Schofield, John "The fight for knowledge: Canada's university libraries are battling for access to cutting-edge research," http://www.macleans.ca/xta-asp/storyview.asp?viewtype=browse&tpl=browse_frame&vpath=/2000/12/04/Education/44105.shtml December 4, 2000

⁴ Hoffert, Barbara "Book Report 2001: The budget shifts," Library Journal, 126:3, Feb. 15, 2001, p.130-132.

⁵ D Greenstein, S and Thorin, D Mckinney "Draft report of a meeting held on 10 April in Washington DC to discuss preliminary results of a survey issued by the DLF to its members, 23 April 2001," <http://www.diglib.org/roles/prelim.htm>.

⁶ Anne R. Kenney et al "Preservation risk management for web resources: virtual remote control in Cornell's Project Prism," D-Lib Magazine, 8(1), January 2002, <http://www.dlib.org/dlib/january02/kenney/01kenney.html>

More open-access web resources that are not covered by licenses or other formal arrangements are included in the catalogs and gateways of research libraries. A spring 2001 survey of Cornell's and Michigan's Making of America (MOA) collections revealed that nearly 250 academic institutions link directly to the MOA collections, although neither university has committed to provide other entities with long-term access. Similarly, a review of the holdings of several research library gateways over the past few years indicates growth in the number of links to open-access Web resources that are managed with varying degrees of control. Approximately 65% of the electronic resources on Cornell's Gateway are unrestricted, and additional open resources are included in aggregated sets that are available only to the campus community.⁷ One of the links is to the University of California, Berkeley's CPU Info Center. This resource is notable because Tom Burd, the site manager, has done several things to advance its preservation, including establishing a mirror site, documenting changes, and providing a checksum in the source page. A recent note posted on this site, however, demonstrates how fragile such resources can be:

"I am no longer affiliated with U.C. Berkeley, and it has become very difficult to maintain this site. With the state of the web now, as compared to when I started this site in 1994, I'm not sure if it even warrants continuing on in light of many other online resources. As such, I will probably bring this site to a close in the coming weeks. If someone wanted to take over maintaining this site, I would be happy to tar up all the files and hand them over. Please drop me a line if you are interested..."⁸

The web citations analyses conducted by Philip M. Davis (Davis and Cohen 2001 and 2002) indicate:

Web citations checked in 2000 revealed that only 18% of URLs cited in 1996 led to the correct Internet document. For 1999 bibliographies, only 55% of URLs led to the correct document (Davis 2001)⁹. A follow-up in year 2000 discovered that within six months 13% of the citations were found at a different URL and 16% of the citations could not be found at all. He also noted a growing tendency to rely on Web resources: from 1996 to 2000, Web citations increased from 9% to

⁷ *ibid*, Notes & References #3 <http://www.dlib.org/dlib/january02/kenney/kenney-notes.html#3>

⁸ <http://bwrc.eecs.berkeley.edu/CIC/> This link is no longer valid as of June 18, 2002; however, this message can be viewed under Google's cached page information.

⁹ Davis, Philip M. and Cohen, Suzanne "The effect of the web on undergraduate citation behavior 1996-1999," *Journal of the American Society for Information Science and Technology (JASIST)*, 52:4, Feb 15, 2001, p.309-314.

22%. (Davis 2002)¹⁰. "creating and maintaining scholarly portals for authoritative Web sites with a commitment to long term access." Davis recommended (Davis 2001).

Preservation of electronic scholarly journals

In October 1999, the Council on Library and Information Resources (CLIR), the Digital Library Federation (DLF), and Coalition for Networked Information (CNI) convened a group of publishers and librarians to discuss responsibility for archiving the content of electronic journals. The group was asked to consider what would be required to ensure access to electronic journals for 100 years. A practical initiative to identify and build consensus around appropriate archival practices and to facilitate the development of lasting digital archival repositories for electronic scholarly journals, funded by the Andrew W. Mellon Foundation to plan long-term archival solutions for electronic scholarly journals. Seven major libraries have now received grants from the Andrew W. Mellon Foundation including the New York Public Library and the university libraries of Cornell, Harvard, Massachusetts Institute of Technology, University of Pennsylvania, Stanford, and Yale.¹¹

Preservation of orphaned content

Orphaned content is the digital collection that the original open-access publishing site is no longer functioning. This proposal intends to create an Archive Service for the preservation of these orphaned digital collections. The Archive Service will not operate the publishing site and keep as a "dark area", but will provide full open access to the orphaned content, based on metadata and indexing. Archive Service will also utilize the Trusted Backup to protect its archived materials.

"Dark" content is that which is not accessible for normal daily use. An archive that keeps its content dark poses less of a threat of competition to the publishers with whom it is working. A dark archive will also be relieved from having to maintain a current user interface, with all of the bells and whistles that users have come to expect, and from the complex task of maintaining information on who has access to what

¹⁰ Davis, Philip M. and Cohen, Suzanne "The effect of the web on undergraduate citation behavior – A 2000 update," *College & Research Libraries*, 63:1, Jan 2002, p.53-60.

¹¹ Preservation of electronic scholarly journals <http://www.diglib.org/preserve/presjour.htm>

content. And, insuring that content that is never used remains sound and free from degradation will be challenging.¹²

UBC Library's vision to digital repository initiatives

The University library is committed to supporting the learning and research needs through the acquisition, provision and preservation of information resources locally, in print, electronic and other formats, and through access to information resources beyond the campus. In meeting the objectives of 2003 of participating in initiatives to preserve electronic information,¹³ UBC Library envisions that the repository for e-journal is part of a suite of digital library infrastructure that supports research and teaching and has persistent value. Management of these digital contents will require effective use of library methodology, assimilating the way libraries are dealing with the printed materials. Library users will expect multi-form, professional and quality services. It is foreseen that deployment of network learning webs will prompt new demands for new forms of services, emerging as time evolves. Libraries will not only be a part of public services in the future, but also will link schools, communities, research institutions, enterprises, industries, businesses, and even striding across national boundaries, to become an important portion of international knowledge informatics.¹⁴

Objectives of this proposal

Researchers are utilizing online-publishing technologies to broaden global access to research through open-access (meaning free) e-print services and online journals.¹⁵ In fact, anyone with access to a library can and should have access to this breadth of information. The digital divide between the information "haves and "have-nots" can be overcome by making the majority of our culture's information available to everyone.¹⁶ The Public Knowledge Project (PKP) is developing and

¹² Project Harvest: developing a repository for e-journals, Cornell University Library
http://cidc.library.cornell.edu/about/project_harvest.htm

¹³ Furthering Learning and Research: Implementing the UBC Library's Strategic Plan 2000-2003, p.3, 26

¹⁴ Ho, Ted and Chiang, Chia-Ning "The spring of 2005: Four visions of Internet services," In the *IT and Global Digital Library Development*. West Newton: MicroUse Information, 1999, p.191-200.

¹⁵ Willinsky, John and Wolfson, Larry "A tipping point for publishing reform?" JEP: The Indexing of Scholarly Journals, July 2002. <http://www.press.umich.edu/jep/07-02/willinsky.html>

¹⁶ Kahl, Brewster Prelinger, Rick and Jackson, Mary E. "Public access to digital material," D-Lib Magazine, 7:10, Oct. 2001. <http://www.dlib.org/dlib/october01/kahle/10kahle.html>

testing a number of online research management systems, including Open Conference Systems, Electronic Theses and Dissertations, and Online Journal Systems, to improve the scholarly and public quality of academic research. These systems are designed to not only assist in the management and publishing of scholarly work, but to improve the indexing of research in online environments and create a richer context of connections for any given study, connections both within the scholarly literature and to the larger world of related online information.

This proposal asks the library to consider operating an electronic repository for digitally published materials that have some relationship with UBC. That relationship, in the first instance would be published materials that have utilized UBC's PKP publishing software, but it could apply to other UBC projects in this area as well. PKP would participate in supporting and developing this new library service.

The repository is envisioned to serve two functions. It is needed to offer access to orphaned content, in which the original site of publication utilizing PKP software is no longer able to operate, and the scholarly content of that site would be transferred to the UBC repository where it would be accessible through a basic access interface and database that meets, say, Open Archives Initiative standards. A second function would be for existing sites of publication utilizing PKP software which would on a regular basis place a copy of their scholarly content in a Trusted Backup conservation site, which would not be otherwise accessible. Should one of these existing sites fail for whatever reason, it could restore its content using the materials in the Trusted Backup. PKP would work with library officials and editors to build a framework for collaboration between a library-based archive service and e-journal content providers who wish to deposit material in the archive.

What are the benefits for the UBC Library?

1. Library will take leadership role in developing digital preservation policies

Deposit agreements may identify the detailed characteristics of the data and accompanying metadata that are deposited, the procedures

for the deposit, the respective roles, responsibilities, and rights of the repository and the data procedure with regard to those data, references to the procedures and protocols by which a repository will verify the arrival and completeness of the data, etc.

The repository proposal will define its mission with regard to the needs of scholarly publishers and research libraries. It will also be explicit about which scholarly publications it is willing to archive and for whom they are being archived. Criteria may include subject matter, information source, degree of uniqueness or originality, and the techniques used to represent the information.

2. Library will establish new boundary and new structures for collection management

Change in the scholarly communications system requires librarians to create new and expanded roles for themselves in the scholarly communications system. Librarians will have to play a much more active role in the creation of scholarly publications. They will have to assert aggressively their professional principles for free and unbiased access to the world of knowledge in the face of trends to commercialize and restrict access to information. The UBC Library's SWOT Analysis had found the threats:

*The Library does not have an exclusive campus mandate to provide information resources, and other campus departments, such as IT Services, could become competitors in the electronic environment. Vendors of electronic resources and databases can bypass libraries and market their products directly to users, and they may do so to a great extent. Many users already find information elsewhere, and others may be willing to go to alternate suppliers who are cheaper or who can provide faster service. The failure or inability to keep up with new developments in information technology could also lead to the loss of Library users.*¹⁷

Collection management practices and perspectives must change in the face of environmental shifts in information services and higher education. Librarians may achieve an ultimate goal: a freely accessible, integrated, and comprehensive record of serious scholarship and knowledge.¹⁸

¹⁷ **UBC Library SWOT Analysis** January 25, 2000 15 p. Accessed by Oct. 8, 2001

¹⁸ Branin, Joseph Groen, Frances and Thorin, Suzanne "The changing nature of collection management in research libraries," *Library Resources and Technical Services*, 44:1, January 2000, p.23-32.

3. Library will extend its contributions to the digital age

In the digital age, the "library model" for funding and sharing information will be scrutinized for its applicability in a world of access. Collection management librarians must take the lead in wedding print collection management to new storage and electronic access and delivery options to maintain and preserve the record of knowledge.¹⁹

In recent centuries, research libraries have played significant roles in research and education: selecting and organizing materials for collections; developing systems of intellectual access; organizing items for physical access and retrieval; and preserving items for long-term use. These attributes signified a durability that is now challenged in today's fast-paced digital environment of networks, web interfaces, and proliferating search engines. We cannot ignore the rapid acceleration of digital dependence in all aspects of education and research, nor can we overlook the researcher's need for permanence, reliability, and continuity in this digital environment. Thus as we look to the new century, we must shape an information environment that has sustainable systems of access to enduring information resources so that users, now and in the future, can rely on them with confidence. Defining this future calls for new combinations of talent and expertise, for short- and long-term collaborations, and for experimentation and risk-taking in order to develop the best strategies for managing the rapidly expanding amounts of digital information. Our challenge is to ensure the viability, the continuity, of information for the scholars of 2020, 2050, and beyond.²⁰

A number of key issues can be explored for the betterment of digital collection management, including technical issues (such as file formats, archive organization and maintenance, and access control), management issues (such as submission procedures and administrative staff support), economic issues (such as installation and support costs), quality issues (such as quality control criteria), policy issues (such as digital preservation and collection

¹⁹ Ibid p.32

²⁰ Cline, Nancy M. "Virtual continuity: The challenge for research libraries today." *EDUCAUSE Review* (May/June 2000). <http://www.educause.edu/pub/er/erm00/erm003/cline.pdf>

management standards), academic issues (such as scholarly communication cultures and publishing trends), and legal issues (such as copyright and intellectual property rights).

4. Library will develop a trusted digital repository program

The UBC Library can gain experiences from this model for handling the production system, preparing itself a trusted digital repository to fulfill the digital preservation mission that very soon to come. This trusted digital repository meets the needs of research resources. Long-term preservation means two distinct but equally important functions: long-term maintenance of a byte stream and continuing access to its contents through time and changing technology. This proposal will build a foundation to open a new feature in the public knowledge and open archive side.

Detail work plan

1. Archive Service: Trusted Backup

The first phase scholarly content from PKP publications will arise from its free distribution of the Open Conference Systems (OCS); Open Journal Systems (OJS) to conference directors and journal editors on a global basis. The OJS and OCS manage access and provide an open and working archive for the content of those who install these systems. Users of these systems may or may not regularly back up their archives, but those who contribute and use this content, also want reassurances that these materials are protected in multiple ways. So the various users of the OJS may send to our repository an update every 3-6 months any additions to their content, knowing that we will protect the integrity of this material, should anything happen to their site and backup systems. Users of the OCS may just send us one complete version of the conference papers, since their site will not change once the conference is over.

The individual component files of the Archive Service fall into two categories: content and metadata.

(1) Content

Content files are the primary carriers of the intellectual meaning of the issue, it may occur at two separate levels: issue- and item-level.

- **Issue-level content** includes bibliographic information and other editorial material such as the masthead, site logo, cover image(if any), table of contents, board of editors, and editorial policy statement.
- **Item-level content** includes citable scholarly material such as peer-reviewed articles, editorials, reviews, correspondence, and errata, as well as supplementary, though pertinent materials such as data files associated with individual items.

(2) **Metadata:**

Metadata files provide pertinent information about the content and the content files themselves. In compliance to Harvard's standard metadata may include:

- **Descriptive metadata** provides information useful for resource discovery at the issue and item level;
- **Administrative metadata** encompasses rights metadata, which provides information concerning the intellectual property rights associated with the journal issue and its individual item-level components;
- **Provenance metadata**, which provides information concerning the creation and fixity of those components; and
- **Technical metadata**, which provides information useful for the archival preservation and delivery of issue and item-level components.
- **Structural metadata** provides the necessary information to successfully re-aggregate the individual file components of the SIP into unified issue items and, ultimately, an issue.²¹

We may need to develop our own set of metadata requirements for this proposal.

2. File formats: Text File, PDF, and HTML.

In order to prevent suspicion that the record might have been tampered with and maintain the most authoritative, up-to-date, reliable or useful forms of access to Internet materials, the Archive Service file server contains read-only "originals" of each issue. In

²¹ Harvard E-Journal Archive Submission Information Package (SIP) Specification Version 1.0 DRAFT, December 19, 2001.

case there are questions about later "copies," there will be a place to find what every issue looked like on mailing day.²² There are procedures for backup and provisions need to be set up for the continuity that should make E-journal as "permanent" as anything on paper. However, the UBC Library will maintain the ASCII file for the ease of conversion purpose to maintain the continuity of digital information.²³

3. Mirroring Management

This repository holds papers and metadata about papers, as well as software that is useful to maintain archives. Everything contained in an archive may be mirrored by request. For example, if the full text of a paper is in the archive, it may be mirrored by request (mirroring policy about time and frequency should be further defined). If the archive does not wish the full text to be mirrored, it can store the papers outside the archive. The advantage of this "remote storage" is that the Archive Service maintainer will get a complete set of access logs to the file.

4. Access Control

Policy of different level of access ID and password should be set up for the security of this server as a trusted, reliable, and certified repository. We would seek "long-term maintenance," with very restricted access, limited to cases in which the main sites and their immediate backups have been lost.

Possible Implementation Process

Phase I: Form a Joint Task Force

UBC Library and the PKP form a Joint Task Force and define individual responsibilities.

Phase II: Define the repository policy

1. Define the policy of acquiring, preserving, and

²² Cameron, Robert To link or to copy?--Four principles for materials acquisition in Internet electronic libraries. Technical Report TR 94-08, School of Computing Science, Simon Fraser University, December 1994. <http://elib.cs.sfu.ca/project/papers/e-lib-links.html>.

²³ "Significant Properties-a simple example: A repository decides that the only significant property of an electronic journal published on the Web is the text within the journal, not its layout and formatting. There is no need to store information about the HTML environment, but only to include information about retrieving or rendering an ASCII text file."

accessing of what (content) and how (procedures).

2. Define the repository metadata set

The Joint Task Force defines the metadata requirements in regard to item-level and issue-level of the e-journal repository.

3. Define guidelines of file maintenance

The Joint Task force defines guidelines including file conversion, periodic data auditing, format migration, and backup procedures.

Phase III: Create a "dark area" to collect, store and maintain quality control of e-journal repository

The Joint Task Force will decide the storage space needs, set up a file server with access control to enable the content providers to make e-journal deposit, and perform the backup and maintenance of the e-journal repository.

Phase IV: Create citation linking

The citations encoded within journal articles can be parsed and harvested, by using the PKP indexing tool, that will allow researchers to search and locate papers to a given article by using common citation information and word similarity, to graph citation links, and compute "hubs" (articles that cite many highly cited articles) and "authorities" (highly cited articles).

Configuration

Hardware:

To develop a simple prototype, started from a simple PC.

Software:

To combine with apache, mySQL, and PHP can meet most of the requirements (OS can be windows or Linux, since mySQL and PHP support both). As for software, PHP can easily help the maintainer to support PDF, HTML and Text File format (PHP has API for PDF).²⁴

Backup

Since the Archive Service serves as a backup storage in "dark area" for

²⁴ The Word format is a problem. If one really has to create or edit Word file in the server side script, then one might have to consider Windows platform since there are more software module can help one with this matter.

e-journals published by systems on the Internet, the repository itself needs backup for the safety and permanence of the digital collection. The key reliability factors of the repository identified are:

- Avoidance of erasure (including deletion and overwriting by users) through write-once policies.
- Backup agreements that incorporate replication policies by the system provider.

A "reliability layer" within this distributed archival repository architecture encompasses a series of functions and mechanisms that results in a reliable environment for preserved objects. Some of the functions identified include:

- Detection and restoration of missing/corrupted information.
- Communications among trusted components include exchanges between parts of a system and also between federated member systems.
- User security, intellectual property management, query processing.
- Import/export facility to move objects into and out of the storage.

We need to develop the *backup policy* and *avoiding, detecting, and restoring lost/corrupted information*.

References:

1. Beagrie, Neil and Greenstein, Daniel *A Strategic Policy Framework for Creating and Preserving Digital Collection, Version 4.0, Final Draft*, (1998) ahds.ac.uk/strategic.pdf.
a more complete discussion of the roles and responsibilities of different stakeholders in the lifecycle of digital materials.
2. Atkinson, Ross "Managing Traditional Materials in an Online Environment Some Definitions and Distinctions for a Future Collection Management", *Library Resources and Technical Services*, 42 1, 1997, p.7-20.
3. Procedures and guidelines developed such as:
 - Cornell University Library *Report of the Digital Preservation Policy Working Group on Establishing a Central Depository for Preserving Digital Image Collections, Version 1.0* (March 2001)
www.library.cornell.edu/imls/image%20deposit%20guidelines.pdf;

- Arts and Humanities Data Services (AHDS) provides guidelines for each of its service providers in Visual Arts, Performing Arts, Electronic Texts, History, and Archaeology, www.ahds.ac.uk/dephow.htm;
- National Library of Australia, *Safeguarding Australia's Web Resources: Guidelines for Creators and Publishers* (2000) www.nla.gov.au/guidelines/2000/webresources.html.
- Steenbakkers, Johan *The NEDLIB Guidelines: Setting up a Deposit System for Electronic Publications*, NEDLIB Report Series, report 5 (NEDLIB Consortium, 2000), Networked European Deposit Library (NEDLIB) <http://www.kb.nl/coop/nedlib/>
- Public Knowledge Project (PKP) <http://pkp.ubc.ca/>