Strategies for Expanding e-Journal Preservation

Shannon Regan, Joyce McDonough,
Bob Wolven (co-PI), Columbia University Libraries
Oya Y. Rieger (co-PI), Cornell University Library
May 2016

Introduction

The evidence indicates that the extent of e-journal preservation has not kept pace with the growth of electronic publication. In 2012, the Keepers Registry compared the e-journal holdings from Columbia, Cornell, and Duke to seven preservation agencies and discovered that only 22-27% of titles were preserved.[1] Influenced by the findings of the Keepers Registry study, in 2013, Columbia and Cornell Universities (2CUL) secured funding from the Mellon Foundation for a 2-year project to specifically evaluate strategies for expanding e-journal preservation. This was a follow-up study to expand on the results of a Phase 1 2CUL project that looked into a number of pragmatic issues involved into deploying LOCKSS and Portico at Cornell and Columbia. The goal of this initial project was to explore the implications of local institutional practices for the respective preservation frameworks and to understand how the third party preservation services were being leveraged.[2]

During 2013-2015, Phase 2 of the 2CUL study intended to expand the preservation coverage of e-journals and to implement strategies that will sustain the initiative beyond the Project. The key research questions includes: What is not being preserved?; Why are they not being preserved?; and How do we get them preserved? While quantitative goals are difficult to set until methods have been tried and proven, our objective was to increase the number of preserved e-journals by 1,500-2,000 through direct action and initiate methods and procedures that would preserve a total of 5,000-7,000 titles within two years. We applied qualitative and quantitative analyses to characterize the range of non-preserved titles. In consultation with Ivy Plus consortium[3] and preservation agencies, we identified a group of high-priority content within this group for investigation and action. These titles were drawn from the holds of Columbia and Cornell and identified as not

[1] This 2012 case study by Adam Rusbridge explores the benefits of the Keepers Registry, comparing an institution's e-journal holdings catalogue against the archiving agency metadata held by the Keepers Registry: http://thekeepers.blogs.edina.ac.uk/2012/11/05/a-trial-holdings-comparison. There is also a 2013 update by Peter Burnhill: http://thekeepers.blogs.edina.ac.uk/2013/10/28/generating-actionable-evidence-on-e-journal-archiving/

[2] 2CUL eJournal Preservation Study, Phase 1: https://www.2cul.org/files/2CULLOCKSSFinalReport.pdf (accessed May 1, 2016)

[3] The Ivy Plus consortium includes a group of twelve US libraries including the eight Ivy League institutions plus Duke, Johns Hopkins, MIT, and the University of Chicago.

preserved through The Keepers Registry. We explored the feasibility of developing specific, concrete procedures that would serve to sustain and expand the project's accomplishments following its completion.

Throughout the project we also met and consulted with the major parties currently engaged in e-journal preservation, including CLOCKSS, LOCKSS, Portico, the Keepers Registry, the Directory of Open Access Journals (DOAJ), the Public Knowledge Project (PKP), and the Center for Research Libraries. In addition, careful attention was given to outreach efforts to expand the understanding of issues around preservation for e-journals, as a first step towards expanding the number and range of active participants in this work.

Project Strategy

Recognizing that preservation status is not a binary value but covers a range of conditions, one of the first things we did was defining a set of preservation statuses: preserved, protected, in process, and not preserved. For an e-Journal to be considered truly preserved, it must meet the following criteria:[4]

1. Reliability of the Repository: The title has been deposited with a trusted digital repository that manages long-term archiving based on the OAIS model and/or has passed TRAC or equivalent evaluation.

2. Access to the Preserved Content: Terms have been established and agreed upon as to how, when, and under what circumstances libraries can discover and access the content and make it available to their patrons.

3. Nature of the Preserved Content: The repository has made available a written definition of the information it preserves (e.g. source files from a publisher, presentation files from a website, ancillary files, etc.).

4. Extent of the Preserved Content: There is confirmation that all retrospective content has actually been deposited, and that deposit is continuing on an on-going basis if the title is still being published. It is also highly desirable for there to be public notification through one or more sources that the journal has been preserved so that libraries can find it (e.g. a note in the OCLC record, listing by Keepers Registry, etc.).

The project focused on five main categories in an effort to develop methods for expanding the coverage of e-journal preservation: content produced or made available by major publishers, small publishers, aggregators, universities acting as publishers, and open access

[4] This set of preservation status criteria was developed in consultation with the 2CUL Mellon eJournal Advisory Group. In addition to the project team, the group included Janet Gertz and Breck Witte from Columbia and Jesse Koennecke and Kizer Walker from Cornell.

journals.[5] Taken together, these categories gave us a target set of roughly 30,000 active e-journal titles in need of preservation. We also worked with these groups to validate priorities, including the decision to exclude several categories of e-journals that were either considered low priority for preservation, or as offering little chance of success within the project's duration. [6] Realizing that the corpus of un-preserved e-journals is large, we wanted to characterize what is considered most valuable from a scholarly perspective and thus poses a greater risk of loss.

Major publishers

Title lists from nine major publishers were evaluated for the preservation coverage of their holdings. The Keepers Registry and data provided by Portico were integral in determining the preservation status of these titles. Shannon Regan, eJournal Preservation Librarian, contacted seven publishers with at least one non-preserved title to request further information and to encourage steps to complete preservation. Five of the seven provided detailed responses. In general, these publishers deliberately excluded certain categories from third-party preservation initiatives, mainly: indexing and abstracting titles, newsletters, and continuing education e-journals. Two publishers did not respond to repeated inquiries; however, subsequent analysis revealed that virtually all of their titles in question were later committed to preservation through Portico, though we are unable to say whether our inquiries prompted that action.

Through this analysis two additional factors relevant to the preservation of e-journals from major publishers became evident. First, the extent of preservation coverage for major vendor backfiles varies considerably. Second, the transfer of titles from publisher to publisher can have an adverse effect on the preservation of a title, even if both publishers involved participate in a third party preservation initiative.

[5] In earlier discussion with Portico, we came up with some possible categories for analysis: Large publishers (defined by number of titles published), Small publishers, 3rd party publications (e-journal provided by a publisher other than the original publisher), Aggregators (agencies that provide e-journals from many publishers), Open access, freely accessible journals, Databases (collections of e-journal content not organized by journal title and issue), US Government publications, East Asian journals (mostly in large aggregations), Historical collections (non-current journals, usually digitized from libraries), Newsletters, trade publications, Book series (monographs published under a series title) Broadly speaking, we intend to focus efforts on the first five groups and exclude the others. Some of these categories overlap, and large aggregators in particular will include examples of multiple types.

[6] The following categories of e-journals were excluded from detailed analysis (and from action during the project): Historical collections of older journals digitized from print, such as HathiTrust; Gallica; 19th Century UK Periodicals Series; Large collections of East Asian collections journals (either commercial or open-access) published in China, Japan, and Korea such as DBPIA and Open J-Gate; Aggregations of e-journal content designed for searching, without provision for browsing and reading individual titles and issues such as Factiva; Lexis-Nexis; Serial titles published by US government agencies such as MARCIVE record set; Newsletters, trade publications intended for a narrow audience such as Business & Company Resource Center, Collections of monographs published under a series title such as lecture notes in computer science.

Small and Mid-Size Publishers

In consultation with the Ivy Plus advisory group and with Portico staff, we identified 50 small- to mid-size publishers for detailed analysis and action: 30 already participating in Portico, CLOCKSS, or LOCKSS to some extent, and 20 not participating. The eJournal Preservation Librarian analyzed the preservation status of their e-journals and contacted publishers with non-preserved titles. The obstacles to preservation for small and society presses are in stark contrast to those of major vendors. Small and society publishers face significant cost barriers to participation in third party preservation initiatives. Similarly, most do not have the technological expertise or funds to implement new technologies supporting preservation strategies. Most evident in conversation with small and society presses is a clear lack of understanding as to what digital preservation is and how it may be accomplished. Overwhelmingly, the majority of small and society publishers contacted in conjunction with this project did not participate in any preservation initiative simply because they were not aware of the initiative or did not understand how it worked.

Of the 20 non-participating publishers in this group, two have now joined Portico (with 51 titles preserved or queued for preservation) and two others have been purchased by participating major publishers.

Aggregators

During the second quarter of the project we initiated discussions with two major aggregators of e-journals, EBSCO Information Services and ProQuest, regarding their potential role in preservation and the barriers to preserving the e-journal content they provide. These discussions continued throughout the course of the project. With EBSCO we entered into a deeper collaboration to explore potential business models for supporting a preservation service and to develop workflows for obtaining the rights to preserve the content distributed by the aggregator.

Through this collaboration the project manager contacted over 350 publishers. Most of these publishers publish one or two titles. Overall, there was a 30% response rate in which over half of the respondents agreed to participate in the pilot preservation service. *One of the key takeaways from this project is that these small publishers are not resistant to preservation but simply do not know about digital preservation initiatives or the expectations of the library community in regard to the archival status of journals subscribed.* Outreach and education have the potential to go a long way in securing preservation, but this is very labor-intensive work, requiring repeated conversation and explanation to preserve a relatively small number of titles.

In order to assess interest in an aggregator-provided preservation service, EBSCO and staff from our project held discussions with EBSCO's Academic Library Advisory Group and with the Ivy Plus Collection Development Group. As of this writing EBSCO has chosen not to pursue offering preservation as a service, but has begun to include the right to preserve content in future agreements with publishers, and to pass that content onto a third-party

preservation agency under defined conditions. They are continuing to explore ways in which to act on these agreements so that the content which they distribute is preserved.

Our discussion with Proquest resulted with a different outcome. The Proquest staff involved in the conversations with the 2CUL team were intrigued with the concept of "preservation." They decided to develop and administer a survey to assess the opinions and expectations of member libraries and publishers (current status unknown). Another related discussion was the pilot work CRL and Proquest have been involved in (within the context of the CRL e-newspaper preservation project) to experiment with a light-TRAC process. The purpose of this concept is to assess publishers/aggregators' technical infrastructures, policies, work flows, and business models in order to assess their preparedness and reliability for performing preservation responsibilities.

Open Access

Freely accessible e-journals comprise the largest, most diverse, and in all likelihood most problematic category for preservation in our target set. They vary widely in importance, in content, and in publication methods and source. There is considerable overlap with other categories addressed in the project, and we excluded from this group open access titles provided through major publishers and aggregators. There are also other active preservation efforts focused on specific groups of open access e-journals. LOCKSS has developed a plug-in that works with the latest version of the Open Journal System (OJS) platform and the Public Knowledge Project is in the process of formalizing their private LOCKSS network. In addition, Portico has a helpful export plugin for OJS users to ensure Portico receives the most accurate files. After consultation with the directors of LOCKSS, the DOAJ, and the PKP program, we decided to avoid overlap with their work and focus on small, independently published titles.

After further conversation with the directors of LOCKSS and CLOCKSS, we decided that the most effective method for immediate action would be to capture the e-journal content through web archiving, with the aim of using either a Private LOCKSS Network or a Fedora repository for long-term preservation. Through its pre-existing web archiving program, Columbia has already preserved 77 e-serials in the Human Rights and Avery Library collections. To test this approach further we selected 77 titles that had been identified as important by selectors within 2CUL. eJournal Preservation Librarian contacted the publishers with details regarding the project and proposed the use of web archiving tools for e-journal preservation. The response was swift and overwhelmingly positive, only one publisher asked to be excluded. Columbia's web archiving team then harvested the content and analyzed the results to identify and resolve issues of scope and quality control.[7] We believe this method can easily be extended to a large number of freely accessible e-journals.

[7] The collection is available at https://www.archive-it.org/collections/5921

University Publications

A relatively small but growing number of e-journals are published by universities, often through units associated with their libraries. Many are published through the OJS platform noted above, though not all are open access. We decided to devote special attention to this category, recognizing that publishers in this group often have active digital preservation programs whose content is not tracked by Keepers Registry. We examined the websites of a number of other university publishers for statements about archiving practice and contacted a few of the major publishers directly, to inquire about their preservation practices and policies. We found that most of these publishers have not made formal, open preservation commitments, but do plan to preserve their e-journal content.

Impact of the Project

One major objective of the project was to secure within two years the preservation of 5,000-7,000 e-journal titles that had not been preserved at the project's outset: 1,500-2,000 through direct action, and the remainder by applying methods devised through the course of the project. It is impossible to say with any precision how many e-journals have been preserved as a direct result of this project. As an example of the complexity involved, one major publisher did not respond to our initiative identifying titles among their online publications that were not preserved; shortly after, however, this publisher deposited over 200 of these titles with Portico. Similarly, Portico engaged directly with some smaller and mid-sized publishers based on our identification of high-priority titles. At least one university library decided to assume archiving responsibility for the journals it publishes after discussion with our project staff. Other titles have made progress towards preservation as a result of our work, but are not yet fully preserved. These include titles from publishers who have now granted archiving rights to an aggregator, and titles in Columbia's web archive; we would consider these to be in "protected" status. Taking all these factors into consideration, we can say with confidence that at least 700 titles we explicitly addressed through this project are now either preserved or "protected," and an additional 5,000-7,000 titles can secure protected status by applying the methods devised for aggregator-based and freely accessible journals.

It is even more difficult to measure progress toward a goal of "full" preservation. As we have repeatedly noted, the definition of e-journals (not to mention preservation) is imprecise. We selected e-journals in Columbia's and Cornell's catalogs as a corpus to analyze and focus on. In the course of our work, however, we encountered and preserved titles that had not been cataloged, while the total number of e-journals added to our catalogs over the past two years appears to exceed the number preserved during the course of the project.

While we can't directly link cause and effect, we were able to analyze the net change in preservation status since the project's beginning for 58,000+ e-journals held by Columbia as of July 2013. This data set was selected for analysis because it had been analyzed by Keepers Registry just prior to the start of the project. In August 2013, 44,889 titles out of 58,556 had not been preserved by any agency tracked by Keepers Registry. At the end of

June 2015, 4,043 titles in this set had been preserved by at least one agency. Thus, the percentage of preserved titles increased from 23.3% to 30.2%. The situation improves further when certain exclusion categories are removed from the list. For example, if Chinese, Japanese, and Korean titles are excluded, 34% of the remaining titles are preserved. Three similar analyses were conducted on data from Cornell: a set of 54,700 titles that had been analyzed in 2012, using two different sets of ISSNs as match points, and a new set of 63,422 titles from Cornell's current holdings. In the merged data sets from 2012, an additional 2,986 titles have now been preserved, increasing the rate of preservation from 28.1% to 33.5%. Of the full set of current holdings, 31.8% of the titles are now preserved.

Beyond the information tracked by Keepers Registry, the impact of the project has been smaller in numbers, but perhaps greater in effect. One major aggregator is now acquiring preservation rights in all of its licenses with publishers. Web archiving has been tested and proven feasible as a means of preserving journals that are unlikely candidates for CLOCKSS and Portico. Recognition of the complexity and diversity of the preservation landscape will increase opportunities for action by more parties, resulting in protection of more of the total e-journal corpus.

Challenges

Several challenges were encountered during the course of the project. Perhaps the biggest impediment was simply the time required to explain the purpose of the project, including libraries' expectations and needs regarding preservation of e-journals, to many parties with diverse backgrounds and perspectives. Publishers, editors, and aggregators each had different degrees of awareness of issues, but also different understanding of the meaning of terms such as "preservation" and "archiving." All of those working on the project spent considerable time answering questions. This became somewhat easier as we refined scripts for communicating, developed FAQs, and clarified the language in proposed licenses, but the process of negotiating preservation agreements remains labor intensive.

Adding to this challenge was the fact that preservation is not the highest priority for most of the parties we worked with. Staff at Portico and EBSCO were extremely helpful in providing contact information that helped us reach the right people at the publishers involved; even so, capturing and sustaining the attention of those individuals was difficult. The same proved true in attempting to work with staff at the two aggregators; while our interactions were cordial and informative, it was not always easy to get the attention of those needed to resolve issues.

Changes in the e-journal publishing landscape also required adjustments in project plans. Publishers we planned to work with merged with, or were purchased by others. E-journal titles under investigation transferred to new publishers. New preservation initiatives such as the PKP work on OJS titles were begun. DOAJ expanded the journal description data gathered from the OA publishers, now requesting information about the preservation status of journals listed in the directory. In general, these were positive developments for e-

journal preservation, but they did cause a few false starts and course corrections in our work.

Perhaps the most surprising challenge was the degree of questioning we encountered within the library community itself regarding the importance of taking action to preserve e-journals. This was expressed as a combination of (in our view, misplaced) confidence that publishers and aggregators can be relied on to archive their own content, plus doubts about the technical and economic reliability of existing third-party preservation agencies. While these views were by no means universal, we did find a number of librarians questioning the costs of preservation, when no major losses have occurred to arouse immediate concern. The reluctance from librarians to aggressively pursue e-journal preservation may be influenced by confusion as to where the responsibility for preservation lies: with publishers, third party agencies, or libraries. Actions to further expand e-journal preservation will bring additional costs. Assessment of the economic impact of this work lies outside the scope of the project, but will be an important area for further attention.

Individual libraries, despite their concern for preservation, often lack effective means for taking action. Selection and acquisition processes may not involve any direct interaction with the publisher; many titles are acquired as parts of large packages, with no comprehensive provision for preservation. Preservation, formerly a distributed activity for printed material controlled at the local level, has come to rely on centralized infrastructures and action in the case of digital material, without clearly defined roles for those staff charged with responsibility for preserving library collections. Some libraries have sought to include provisions for archiving in their e-journal licenses, either through direct deposit of content with the library or, more often, through third-party agencies. However, many existing licenses pre-date the development of current preservation options, and some licenses do not specify which titles are excluded from preservation agreements. In addition, many e-journals are available without the need for a license agreement.

At the outset of the project, one of our goals was working with the Ivy Plus Collection Development group to expand the current model license terms to incorporate preservation-related expectations and requirements. It was during this project that, with funding from the Andrew W. Mellon Foundation, the Center for Research Libraries announced the newly revised model license, incorporating the recent best practices. Although preliminary in nature, the new model license incorporates some licensee rights related to third party archiving of licensed journals.

Recommendations for Further Action

As noted earlier, e-journals are of many types, and emanate from diverse sources. Some are made available through the publisher; others only through an intermediary or aggregator.

Many are not commercial publications, but are issued by governmental and non-governmental agencies, professional societies and academic departments, with some of the latter issued as Open Access. The state of preservation varies widely among these categories, as does the importance of individual titles to scholars and researchers. Different groups of journals are also likely to require different strategies for preservation. With this caveat, we would like to offer the following recommendations:

Outreach and Advocacy: E-journal archiving responsibility is distributed and elusive. It has been more than seven years since the call for urgent action to preserve e-journals.[8] Although there have been some promising developments, the core challenge remains to be addressed in a comprehensive and persevering manner. Therefore, an integral part of the project is work with libraries, archiving organizations, publishers, and societies in order to develop and promote best practices such as model license agreements and practical steps leading to the deposit of e-journal content with recognized preservation agencies.

Major Publishers: E-journals provided by major publishers are in general well covered by existing preservation initiatives. However, most publishers have some titles that are deliberately excluded from preservation, because they are not considered important or suitable for current preservation models. As libraries and licensing agencies negotiate new licenses or renew existing licenses, publishers should be asked to specify any licensed content excluded from the license's provisions for archiving. We need to engage CRL to explore how the recently revised model license can be further enhanced by broadening the archival information section.

Ensuring Continuity: The preservation status of e-journal titles may change as titles move from one publisher to another, either because the new publisher does not have provisions for preservation, or simply through failure to initiate procedures for the newly-acquired content (as was demonstrated in this project.) The Enhanced Transfer Alerting Service maintained by the UKSG (http://etas.jusp.mimas.ac.uk) provides information that could be effectively used to monitor such changes. Libraries should work with preservation agencies, the UKSG, and the Keepers Registry to explore means of automatic data transfer, confirmation of preservation renewal, and notification of "dropped" titles.

University/Library Publishers: Several university and library publishers of e-journals already have provisions for archiving their content through a local preservation repository. University libraries engaged in publishing should develop a consistent approach to preservation, including open declaration of their archiving policies and practice. This work should help to inform, and be informed by, CRL's exploration of a "TRAC light" certification.

Freely Accessible E-Journals: Columbia and Cornell will work with members of the Ivy Plus group of libraries to assess the feasibility and cost of implementing a Private LOCKSS Network to preserve the pilot collection developed in Archive-It, and will pursue this model

[8] Donald J. Waters, ed., "Urgent Action Needed to Preserve Scholarly Electronic Journals," 2005, http://old.diglib.org/pubs/waters051015.htm

further in other initiatives currently under discussion around collaboration in web archiving.

Technical Development: As digital formats become more complex and new research methods emerge (e.g., text mining), just-in-case dark archiving solutions will be harder to justify from cost-effectiveness and return-on-investment perspectives. It will be beneficial for the stakeholders to reconsider the current assumptions that underlie significant initiatives such as CLOCKSS, LOCKSS and Portico.

Information Exchange: At present, up-to-date information about preservation status is not included in the systems and knowledge-bases libraries use to manage e-journal content. This inhibits libraries' ability to consider preservation as a factor in collection development and collection management. Keepers Registry has significantly enhanced the ability to query the preservation status of individual titles, but libraries should encourage more systematic exchange of preservation information among preservation agencies, subscription agents, and e-resource management systems.

Setting Priorities: One barrier to effective action has been the sheer number of e-journal titles that are not preserved. The analysis performed during the project helps to identify broad categories that can be used to set priorities, but some of these categories are themselves quite large. More discussion among libraries is needed to build consensus around priorities for action on titles provided through aggregators and on freely-accessible e-journals. Although it is important to note that scholarly communication takes increasingly diverse forms on the web, this project is centered upon what is formally issued as e-journals, also including resources for scholarship issued in e-serial form such as government documents.