Columbia University Libraries/Information Services

Strategies for Expanding e-Journal Preservation

Final Report

to the Andrew W. Mellon Foundation

APPENDICES

Contents:

Appendix A: Preservation Definitions & Categories

Appendix B: E-Journal Preservation: Quantitative Analysis

Appendix C: Recommendations for Further Action

Appendix D: Final Financial Report

Appendix A: Preservation Definitions & Categories

Criteria for a Fully-Preserved e-Journal Title

For an e-Journal to be considered truly preserved, it must meet the following criteria.

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.
☐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.).

Levels of Meeting the Criteria

This project will use the following terminology in evaluating the preservation status of titles. In the categories below, where there is a letter, (a) is the best case and quality declines as list goes on. Where there are bullets, no quality judgment is implied.

1. Reliability
a. Secured = meets both criteria.
b. Protected = is in a repository that has a sound track record but that has not met the full requirements (yet). More work is needed, but at the very least there are multiple backups of different kinds, reasonable authenticity and integrity measures, and the repository is sufficiently mature that there is not fear that it may disappear for lack of support any time soon.
c. In process = Efforts are being made, but so far do not come close to meeting the criteria.
d. Not preserved = No one is attempting to preserve this title.

Type of agency managing the digital repository
☐Preservation service such as Portico, LOCKSS, Hathi, etc.
☐National library
☐Academic institution
☐Publisher
☐Aggregator
☐Other

2. Access
a. Available to all based on trigger events or moving walls.
b. Available to libraries that had/have subscriptions based on trigger events or moving walls.
c. Available on-site or within a specific country only (e.g., some national libraries).

3. Nature of holdings
☐Presentation files from website (retains look and feel).
☐Source files from publisher.
☐Includes ancillary files.

4. Extent
a. Complete retrospective holdings and, for live titles, ongoing of new content.
b. Substantially complete but with gaps.
c. Only a small portion deposited.
d. Nothing deposited.

Appendix B: E-Journal Preservation: Quantitative Analysis

The 2CUL e-Journal Preservation project is based on the recognition that only a minority of the e-journal titles in our collections were preserved as of the project's inception. We have cited figures ranging from 12% to 27%, depending on which data set is used, which preservation agencies are used for comparison, and the time of the analysis. These variables show that no single figure can do justice to the complexity of e-journal preservation.

That complexity is compounded by the wide variety of publications that may be characterized as "e-journals." The term is often, if loosely, used to refer to current (or at least recent) scholarly, peer-reviewed journals published online. The 2CUL analysis, however, included all serials in digital form. It thus embraced many types of publications, with different issues, risks, and priorities involved in their preservation.

The current 2CUL project does not aim to address preservation of all e-journals. To have maximum impact, we are focusing attention and action on only a few categories, and on selected publications within those groups. Beyond that, we plan only to characterize the full data set of digital e-serials in more detail, to propose priorities for action, and to suggest preservation strategies that may be appropriate.

We seek input from the library community, and particularly from our partners in BorrowDirect, on our initial categorization of types of e-journal publication and on our identification of priorities for action.

Analyzing the corpus

Our initial 2CUL analysis was based on e-journal (more precisely, e-serial) titles in our catalogs. The data sets from Columbia and Cornell were not identical, but had high overlap and produced similar results. A separate analysis of data from Duke produced results in the same range. None of these data sets includes all existing e-journals, so we should acknowledge that our project is limited to e-journals "collected" (in some sense) by large research libraries.

On the other hand, our catalogs contain many records for e-journal titles that have not been individually selected by our libraries. They also contain records for historical (non-current) journals whose content is preserved in a variety of ways – print, microfilm, and sometimes digital. The importance of digital preservation is not uniform.

Parsing categories

Ideally, we might use the 2013 Keepers Registry analysis of our e-journal holdings as a starting point, since it provides the most complete analysis of preservation status across multiple agencies. This is probably not feasible because the data provides few clues to the nature of each title (only journal title and ISSNs). We would need to match ISSNs against more complete bibliographic and knowledge base information.

The only practical way to sort ca. 200,000 e-journal titles into categories may be using the source information in the SerialsSolutions knowledge base. Those 200,000 titles are grouped in perhaps 1,000 databases. Information on the provider, together with our own knowledge, makes it fairly easy to categorize each database broadly.

In earlier discussion with Portico, we came up with some possible categories for analysis:

a) Large publishers (defined by number of titles published)
b) Small publishers
c) 3rd party publications (e-journal provided by a publisher other than the original publisher)
d) Aggregators (agencies that provide e-journals from many publishers)
e) Open access, freely accessible journals

f) Databases (collections of e-journal content not organized by journal title and issue)
g) US Government publications
h) East Asian journals (mostly in large aggregations)
i) Historical collections (non-current journals, usually digitized from libraries)
j) Newsletters, trade publications
k) Book series (monographs published under a series title)

Broadly speaking, we intend to focus efforts on the first five groups (a-e) and exclude the others (f-k). Some of these categories overlap, and large aggregators in particular will include examples of multiple types. We can begin by agreeing on a hierarchy of exclusion: i.e., first exclude from detailed analysis collections that consist largely of a single category deemed of lower priority (or feasibility) for preservation.

Exclusion Categories

Proposal: exclude from detailed analysis (and from action during the project) collections of e-journals of the following types:

Historical collections: (ca. 50,000 titles)

Description: Collections from either commercial publishers or libraries and non-profit organizations of older journals digitized from print.

Examples: HathiTrust; Gallica; 19th Century UK Periodicals Series

Rationale: Many are already preserved (HathiTrust, Gale collections in Portico); the content is often still preserved in print and/or microform.

East Asian collections: (ca. 11,000 titles)

Description: Large collections of journals (either commercial or open-access) published in China, Japan, and Korea.

Examples: DBPIA; Open J-Gate

Rationale: Securing preservation agreements is likely to be difficult due to language barriers, different legal and publishing environments; individual titles in these collections have not been selected or evaluated.

Databases: (ca. 20,000 titles)

Description: Aggregations of e-journal content designed for searching, without provision for browsing and reading individual titles and issues.

Examples: Factiva; Lexis-Nexis

Rationale: The value of these collections lies in the database as a whole; securing preservation agreements is not likely to be a priority for the database provider; individual titles in these collections have not been selected or evaluated.

US Government Publications: **(ca. 21,000 titles)**

Description: Serial titles published by US government agencies.

Example: MARCIVE record set.

Rationale: Several other efforts are under way to develop programs for preserving digital government publications.

Newsletters, trade publications: (ca. 13,000 titles)

Description: Collections of minor publications intended for a narrow audience, often providing only a summary of current news within an association or industry.

Examples: Business & Company Resource Center

Rationale: As with the "Databases" category, the value lies in the collection as a whole; most libraries would not choose to collect or preserve the individual titles included.

Book series: **(ca. 2,000 titles)**

Description: Collections of monographs published under a series title.

Examples: Lecture notes in computer science;

Rationale: The current preservation strategy for titles in these series is to treat them as e-books. Many are deposited in Portico in that form.

Appendix C: Recommendations for Further Action

One of the main objectives of the project was to develop methods and procedures that could be effectively employed to further extend the range and number of preserved e-journals after the project's completion. Given the diversity of content and the range of agents involved in publishing, disseminating, collecting, and preserving e-journal content, future actions will need to match this complexity and involve many different parties.

Outreach and Advocacy: In addition to publishing information about the project and its outcomes and recommendations on the 2CUL website, Columbia and Cornell will seek opportunities to present this work at professional meetings, and to promote continued action through library organizations, archival agencies, publishers/societies, and aggregators, with actions tailored to specific groups for maximum effect.

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. 2CUL will 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.

Aggregators: The project has developed scripts and model license language that aggregators may use in working with publishers to acquire rights to preserve content and pass that content to a recognized preservation agency in case of a trigger event. As yet, however, no mechanisms or procedures exist to effect such a transfer. Columbia and Cornell will continue to work with EBSCO and ProQuest to explore options, but will also seek to engage partners such as Ivy Plus, NERL, and CRL to carry on this work.

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 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.