

February 2009

Tracking E-Journal Preservation: Archiving Registry Service Anyone?

Peter Burnhill

University of Edinburgh, p.burnhill@ed.ac.uk

Follow this and additional works at: <https://docs.lib.purdue.edu/atg>



Part of the [Library and Information Science Commons](#)

Recommended Citation

Burnhill, Peter (2009) "Tracking E-Journal Preservation: Archiving Registry Service Anyone?," *Against the Grain*: Vol. 21: Iss. 1, Article 13.

DOI: <https://doi.org/10.7771/2380-176X.2496>

This document has been made available through Purdue e-Pubs, a service of the Purdue University Libraries. Please contact epubs@purdue.edu for additional information.

is available only to **Portico** subscribers. The **CLOCKSS** model, on the other hand, makes triggered content publicly available to all under a Creative Commons license and will not recreate subscription access. While one could argue that a trigger event would be extremely unlikely, as most larger publisher's catalogs would be hosted by a successor, such an argument risks the creation of a scenario where eBook content is preserved but cannot legally be triggered, a situation which hardly fulfills the ultimate goal of digital preservation. Preservation of content within a library setting, such as that offered by the **KB** or the **DNB**, avoids the copyright issue but requires a scholar to travel to the Netherlands or to Germany to view needed titles.

A Commitment to Preservation

"There is always a question on archiving at each presentation I give, and rightly so" **Cynthia Cleto** notes. "Since the eBooks have a unique ownership model — customers that purchased a copyright year have perpetual access to that content — preservation becomes a concern. Customers want assurance that they will be able to access what is probably their largest library of eBooks in calm as well as turbulent times." **Craig Van Dyck**, Vice President of Global Content Management for **Wiley-Blackwell**, emphasizes that **Wiley's** strategy to preserve eBook content mirrors their approach to preserving journal content. "We are working with other parts of the industry (libraries, publishers, preservation archives, industry associations, technical experts) to come to terms with the issues, and to determine the best approaches."

The case for preserving eBooks in their digital form is a good one. In time, print archives physically deteriorate. Books can become damaged or be lost. A digital copy is more durable and takes up less space. Despite the challenges, **Springer** remains committed to finding a satisfactory solution in the near future. Knowing that **Springer** has robust preservation measures in place, our customers can rest assured when adopting our eBook content. As part of the **CLOCKSS** outreach committee, I speak regularly with publishers about their preservation strategies — or lack thereof. Recently, I have been describing a new **CLOCKSS** pilot project to ingest eBook content during 2009. As **Springer's** contact for **Portico**, I have proposed conference panels on digital preservation that feature the perspectives of the publisher, the library, and the preservation initiative. (Please look for us at **NASIG** in June). These opportunities naturally raise even more questions: can we preserve databases, electronic supplementary materials, whole Websites, files formatted for mobile devices?

We are still in the early stages of defining a comprehensive digital preservation strategy, one that requires the efforts of different entities with varying models. It is an exciting time to be working in publishing, alongside dedicated librarians and forward-thinking preservation initiatives. Future generations of researchers depend on the success of our collective efforts. 

Tracking E-journal Preservation: Archiving Registry Service Anyone?

by **Peter Burnhill** (Director, EDINA National Academic Data Centre, Causewayside House, 160 Causewayside, University of Edinburgh, Edinburgh, EH 9 1PR, Scotland, UK; Phone: +44 131 650 3301/2; Fax: +44 131 650 3308) <p.burnhill@ed.ac.uk> edina.ac.uk

The idea for a registry of archived scholarly publications has featured in various digital preservation and archiving discussions. In 2003 **Maggie Jones** highlighted the need for clarity on what the various digital preservation agencies were doing.¹ In 2006, **Kenney** et al went further and recommended a registry that would indicate which agencies were preserving which journal content, one that could be used to identify gaps in publisher or content preservation coverage.²

JISC, the agency for the UK higher education that funds initiatives such as these, acted on this and commissioned a "scoping study for a registry of electronic journals that indicates where they are archived."³ Having interviewed a range of stakeholders in the UK, including representatives from national and university libraries, publishers, and archiving organizations, **Sparks** et al (2007) concluded, "Almost everyone agreed that there was ... an overall lack of information about where e-journals were archived, but more particularly, the difficulty of finding the information across a range of sources."⁴

There was, however, a lack of consensus on the scope of the registry. There were differences of view relating to timing, implementation and sustainability, and at least one archiving organization wished funding to go more directly to sustaining archiving per se.

On the matter of organization, the scoping study suggested the registry should be attached to something else that already existed in order to leverage existing organizations and infrastructure. A pilot project was recommended, followed by phased development of the registry using **SUNCAT**, the UK serials union catalogue,⁵ as the possible master list against which to compare the current and planned "holdings" of archiving services.

Piloting an E-journals Preservation Registry Service (PEPRS)

The next step for **JISC** was to commission a pilot registry service from **EDINA**,⁶ the UK national academic data centre based at the **University of Edinburgh** responsible for **SUNCAT**. **EDINA** opted to partner from the outset with the **ISSN International Centre**, with which **EDINA** has had a long-standing association.⁷



Funding for a two-year e-journals preservation registry service pilot (nicknamed **PEPRS**) began

in August 2008. The aim of **PEPRS** is to investigate, build and pilot an online facility that would enable librarians and policymakers, principally in the UK but also worldwide, to ascertain the archival provision for e-journals, especially of scholarly work published in e-journals, and to identify the gaps in such provision.

Preliminary Thoughts on Design and Re-considerations of Scope

A registry for e-journal preservation would correlate what is being done by each preservation agency for each known e-journal. This puts the focus on (1) metadata for e-journals and (2) metadata for each agency and archiving action, both of which are addressed briefly below.

The intention is not just to build and keep a register but also to deliver a set of registry services, so the **PEPRS** project must establish the functionality for a registry service, including the review and testing of user requirements, with implicit consideration of just who constitutes the primary use communities.

A registry needs to be accurate, up-to-date and comprehensive in coverage in order to be effective and command respect, as well as meet specific requirements and functionality. The choice of data model and architecture (3) are critical in determining that these matters of quality can be met.

This registry and the provision of its basic services must be designed to survive for the long run, like its subject matter, digital preservation. This implies (4) a business model that is sustainable over the long, one aspect of which should be low cost.

(1) Metadata on e-journals

A system of persistent and internationally accepted identifiers is clearly a good thing for a registry. The inclusion of the **ISSN International Centre (ISSN-IC)** as partner in the **PEPRS** project is as well. The **ISSN-IC** co-ordinates the **ISSN Network** which manages the international standard numbering system for serials, of which e-journals are a proper subset.

It could be argued that any e-journal worth preserving ought to have an ISSN.

A registry that made use of the metadata hosted in the ISSN serials database would have a critical mass of serial titles for project purposes, likely representing a good majority of the world's scholarly publications, including open access journals.⁸

The total number of e-journals is unknown but could be said to be growing. Fortunately, in recent years the **ISSN-IC** has made e-journals a priority for inclusion in the **ISSN Register** and has already issued over 60,000 identifiers for e-serials.

continued on page 34

A workflow is being devised for e-journals that come to notice as not having been assigned an ISSN, and to cater for use of the print ISSN using the new linking identifier, the ISSN-L.⁹ Our initial intention was to follow ISSN rules and only include as e-journals those serials that were issued in digital format (i.e. “born” digital), and not “digitised journals” which were originally issued in print format, although this is now being actively reviewed for the purpose of this project.

Title-level metadata on serials is essential but it is the *article* that is the information object of desire. Libraries will want to know the extent of preserved content for a given title, in order therefore to know which articles are preserved. This is more complex and, as such, has been deferred to the second phase of the **PEPRS** project. Provisional thinking is to create four date fields for each e-journal: earliest and latest known date of issue in digital format; earliest and latest known date of issue archived.

(2) Metadata on preservation agencies and archiving action on each e-journal

Another key question is which archiving agencies to include in **PEPRS** project activity and over the longer term in the registry. The term “archiving” signals a potential widening of scope beyond that of digital preservation alone, to include “access continuity”: continuity of access to back content. This is triggered by a more recent UK report commissioned by **JISC** in which **Morrow** et al (2008)¹⁰ reviewed the policies and practices of six digital preservation agencies.¹¹ It noted that some agencies focused primarily upon long-term preservation of the scholarly record, while the main emphasis for others was on “perpetual access” — the latter phrase used to refer to “continuity of access” to back content in an e-journal after the cancellation of a current subscription (“post-cancellation”) or as back-up for short-term failure.

Dependence upon leased access to content hosted at remote servers beyond the academy threatens continuity of access for researchers and students via their library. Challenging the very reasons for a library, this has become acute in the near term as financial pressures upon budgets for library materials lead to cancellations of subscriptions.

The main areas of policy interest need to be resolved into agreed, standard fields of information. Examples include title identifiers such as ISSN and title, date ranges, status of preservation, and access conditions. The next step would be the development of a common vocabulary for entries to assist users of the registry service who will want to compare attributes of preservation actions and summary descriptions of the agencies themselves.

Initially the plan for the initial phase of the **PEPRS** project was to limit activity to three types of digital preservation agency: organizations operating at the international

against the grain people profile

Director, EDINA National Academic Data Centre
Causewayside House, 160 Causewayside, University of Edinburgh
Edinburgh, EH 9 1PR, Scotland, UK
Phone: +44 131 650 3301/2 • Fax: +44 131 650 3308
<p.burnhill@ed.ac.uk> *edina.ac.uk*

Peter Burnhill

BORN & LIVED: United Kingdom.

EARLY LIFE: Economics/statistician, research into schooling, university lecturer.

PROFESSIONAL CAREER AND ACTIVITIES: Edinburgh University Data Library (1984), EDINA (1995/96); Director & Methodologist.

HOW/WHERE DO I SEE THE INDUSTRY IN FIVE YEARS: I think we will all look back in five years time and accept that what seemed alternative and informal as means of scholarly communication has become recognised as mainstream record. That's the way with science fiction. 🐼

level (e.g., **CLOCKSS** and **Portico**); national libraries (e.g., **British Library**); and library consortia (e.g., **UK LOCKSS Alliance**). Were the scope of the registry to widen then the list might have to be revisited.

This and the diversity of use communities for the registry imply need and opportunity for cooperative inter-working, via interoperability, with third-party services providing information subscription status, likely organized on a territorial/nation-state basis rather than a global basis.

(3) Data model and architecture

The registry service needs to support machine-to-machine use as well as a Web-based user interface. Responsibility for specific fields of information is placed with the source best placed to deliver up-to-date information. A key feature of the data model is to establish dependence upon information sourced from the **ISSN Register** and from self-statement by the digital preservation (and archiving) agencies. This exploits the “always on” presumption about the Internet in order to ensure up-to-date report by the preservation agencies, and also to keep an historic record of the statements made.

There is likely to be a range of different types of user for the registry service, most of whose needs we hope to meet through a Web interface. However, especially with international use, there may be other communities of users to cater for. One way to address this is to give equal priority to indirect access: that is, to the provision of a programming interface (API) that would provide interoperability to third-party facilities geared to serve specific-use communities across multiple locales and languages.

(4) Business model and sustainability for the registry and its services

This registry and provision of its basic services must be for the long run, like its subject

matter, digital preservation. An important part of the **PEPRS** project therefore is to identify costs and propose a business model for the registry service. It may also be necessary to propose a form of governance.

Not surprisingly, the **JISC**-funded *Scoping Report* for this registry service touched on the matter of sustainability: “The archives themselves have to be sustainable over the long-term and to be of any use whatever, the registry must be equally long-lived.”¹² Discussion of this recommendation may seem premature, but the **PEPRS** project will be reviewed in 2009/2010 to assess whether the results of the project activity thus far and its business plan would justify the transition into service.

That might seem an appropriate open issue on which to end but perhaps this conclusion from the *Scoping Report* is more upbeat: “It seems to us that in order to gain the co-operation of the archiving organisations based around the world, which would be vital to its utility, the registry would have to be conceived as something which would serve the whole international scholarly community.”¹³ The *Report* continues that the registry should be managed and governed “in such a way as to secure and maintain trust of both the library community and publishers.”¹⁴

Request for Comment

PEPRS is a UK-funded project being carried out by a national academic data centre in partnership with an international standards body. In light of its potential to be international in scope and operation, and that any resultant registry service needs to exist over the long-run and to be of benefit across many sectors of the scholarly community, comments on issues raised, including governance and sustainability, are gratefully requested.¹⁵ 🐼

endnotes on page 36

MetaArchive: A Cooperative Approach to Distributed Digital Preservation



Endnotes

1. Jones, Maggie. "Archiving E-Journals Consultancy (Final Report)," October 2003, accessed January 16, 2009; available from http://www.jisc.ac.uk/uploaded_documents/ejournalsfinal.pdf; Internet.
2. Kenney, Anne R., Richard Entlich, Peter B. Hirtle, Nancy Y. McGovern, and Ellie L. Buckley. "E-Journal Archiving Metes and Bounds: A Survey of the Landscape," September, 2006, accessed January 16, 2009; available from <http://www.clir.org/PUBS/abstract/pub138abst.html>; Internet.
3. Sparks, Sue, Hugh Look, Adriene Muir and Mark Bide. "Scoping study for a registry of electronic journals that indicates where they are archived," **Rightscom and Loughborough University**, 2008; accessed January 2009), 33; available from <http://www.jisc.ac.uk/media/documents/programmes/preservation/ejournalregstudy.pdf>; Internet.
4. *ibid.*, 13.
5. SUNCAT includes serials' information from over 60 UK libraries, including the **British Library**, the **National Libraries of Scotland and Wales**, some of the largest Higher Education institutions in the UK and a number of specialist libraries. <http://www.sunecat.ac.uk/>.
6. <http://edina.ac.uk>. In addition to SUNCAT, EDINA also provides organisational support for the **UK LOCKSS Alliance**, <http://edina.ac.uk/lockss/> following project work at the **Digital Curation Centre**. The **University of Edinburgh** is a **CLOCKSS Archive Node**.
7. <http://www.issn.org/>; Internet.
8. The **Directory of Open Access Journals (DOAJ)** requires ISSN assignment.
9. Accessed on January 19, 2009; available from <http://www.issn.org/2-22637-What-is-an-ISSN-L.php>; Internet.
10. Terry Morrow, Neil Beagrie, Maggie Jones and Julia Chruszcz, Tee Em Consulting. "A Comparative Study of e-Journal Archiving Solutions," May 2008. http://www.jisc-collections.ac.uk/news_and_events/news_articles/ejournals_pr.aspx.
11. These six, **LOCKSS (Lots Of Copies Keep Stuff Safe)**, **CLOCKSS (Controlled LOCKSS)**, **Portico**, **e-Depot** (from the Dutch national library, Koninklijke Bibliotheek, KB), **OCLC Electronic Collections Online (ECO)**, and **British Library e-Journal Archiving Programme**, were selected on relevance to UK, with some relation to the 12 e-journal archiving initiatives included within the *CLIR Report* (op cit) as having met seven indicators of viability.
12. Kenney et al, op. cit., 3.
13. *ibid.*, 32.
14. *ibid.*, 16.
15. Comments should be sent to the author at <p.burnhill@ed.ac.uk>. A full version of this article is to appear in *Serials* in time for the UK Serials Group Meeting in April 2009.

by **Katherine Skinner** (Digital Projects Librarian, Emory Univ., 540 Asbury Cir., Atlanta, GA 30322; Phone: 404-783-2534; Fax: 404-727-0827) <katherine.skinner@emory.edu> <http://metaarchive.org>

and **Martin Halbert** (Director for Digital Innovations, Emory University, 540 Asbury Cir., Atlanta, GA 30322; Phone: 404-727-2204; Fax: 404-727-0827) <martin.halbert@emory.edu> <http://metaarchive.org>

What role will the Library take in digital preservation? On first glance, the question seems relatively easy to answer. As the library continues to transition from its centuries-long focus on print assets to a combination of print and digital resources, it will take an active role in the preservation of our digital cultural resources that is similar to that which it has long undertaken in the print realm.

Or will it?

Of late, many of us in the library field have become preoccupied with the concept of digital preservation — and rightly so. We wonder aloud about the forms that digital preservation will take, the amount it will cost, the rigor demanded in its implementation, and the feasibility of different organizational approaches to digital preservation.

But what does it mean to participate? How do we want to be involved? And what role(s) should we, as librarians and archivists, aspire to take in the realm of digital preservation?

Questions such as these led to the founding of the **MetaArchive Cooperative**, a collaborative network of institutions that have banded together to communally approach the challenges of preserving digital assets. The original six members founded this **Cooperative** due to their strong belief that libraries both could be and should be *actively engaged* in the creation and maintenance of their own digital preservation solution. They knew that alone, none of these institutions were likely to create and maintain — much less sustain — a robust digital preservation solution. However, they believed that if they approached the issue as a group and built a *shared* infrastructure, they could accomplish together what no one institution had the resources to achieve in isolation.

The MetaArchive Cooperative: A Shared Digital Preservation Infrastructure

The **MetaArchive Cooperative** (<http://MetaArchive.org>) formed to enable cultural memory organizations to effectively and mutually preserve their archival digital assets *for themselves*. **MetaArchive** began in 2004 as one of the original eight initiatives contracted by the **Library of Congress** under the **National Digital Information Infrastructure and Preservation Program (NDIIPP)**.¹ The venture was led by **Emory University** in collaboration with **Georgia Tech**, **University of Louisville**, **Virginia Tech**, **Auburn University**,

Florida State University, and the **Library of Congress**. The network established by this group was the first major effort to build and operate a private implementation of the open source **LOCKSS (for Lots of Copies Keep Stuff Safe)** software for digital preservation (<http://www.lockss.org>), an approach that has since been termed a **Private LOCKSS Network**, or **PLN**. The **MetaArchive PLN** is a distributed preservation infrastructure that meets the **OAIS Reference Model** standards for repositories.²

Technically speaking, the foundation of the network is the open source **LOCKSS** software developed at **Stanford University**, which enables a group of **LOCKSS** caches, or node servers, to work together across geographical space to replicate and preserve content.³ **MetaArchive** is the only **PLN** in operation thus far that does not depend on the **LOCKSS** team to administer the network; we run a separate cache manager (coded in collaboration with the **LOCKSS** team) to monitor our network. The **MetaArchive Cooperative** has created and layered additional modules on top of the **LOCKSS** framework to provide our members with administrative tools, including a conspectus database and the cache manager. The conspectus database enables members to capture collection-level metadata for preservation decisions and actions, and the cache manager serves as a monitoring tool for network-wide tracking and troubleshooting activities. We are in the process of packaging these open source software components for use by other **PLNs**, and plan to release this software through **SourceForge** next year.

The organizational framework that we have constructed has been as integral to our success as the technological platform upon which we have built our preservation services. After running the network for three years, we transitioned from a sponsored-funding-supported project to an independent, membership association in 2007, a transition that has been greatly assisted through the support of the **National Historical Publications and Records Commission**. As part of this work, we founded a 501c3, the **MetaArchive Services Group**, to administer the **Cooperative**. All of the components of the network we run are owned and maintained by our member institution. This decentralized apparatus enables the **Cooperative** and its services to be *independent*

continued on page 38