Preserving e-Journals: A review of the digital preservation landscape

Ken DiFiore, MLS, MS
Director, Library Relations
Evolution of Digital Content: Global Perspective
Evolution of Digital Content: Library Perspective


Compiled and Edited by
Martha Kyrillidou
Les Bland

Association of Research Libraries
Washington, DC
2009

Average E-Resource Expenditure as Percent of Total LME

Evolution of Digital Content: Library Perspective

- Accommodate patron’s demand for the convenience and accessibility of e-resources.
- Print processing and storage costs becoming prohibitive.
- Require floor space for new collaborative work environments.
- Maintain relevance: in the digital environment, networks and context are everything.
Evolution of Digital Content: Publisher Perspective

Print Typesetter
Evolution of Digital Content: Publisher Perspective

“Electronic” Typesetter

XML “source files” used to facilitate repurpose to multiple outlets.
Why Digital Preservation?

- Digital resources are now generally licensed rather than owned. Preservation and access are separated and the ownership of the preservation task is less clear.

- Multitude of electronic formats.

- Digital resources are surprisingly fragile and increasingly complex and innovative raising significant preservation challenges.
Influential Reactions from Library Community

The E-only Tipping Point for Journals

What’s Ahead in the Print-to-Electronic Transition Zone

Richard K. Johnson and Judy Luther

E-Journal Archiving Metes and Bounds:
A Survey of the Landscape

by Anne R. Kenney, Richard Entlich,
Peter B. Hirtle, Nancy Y. McGovern, and
Ellie L. Buckley

September 2006

Council on Library and Information Resources
Washington, D.C.
Influential Reactions from Library Community

1. Research and academic libraries and associated academic institutions must recognize that preservation of electronic journals is a kind of insurance.

2. Qualified preservation archives must provide a minimal set of well-defined services.

3. Libraries must invest in a qualified archiving solution.
Collaborative Solutions: LOCKSS

- The LOCKSS Program is an open-source, library-led digital preservation system built on the principle that “lots of copies keep stuff safe.”

- Allows libraries to take custody of and preserve access to the journals to which they subscribe.

- Analogous to libraries’ using their own buildings, shelves and staff to obtain, preserve and provide access to paper content.

- Initiated and managed at Sanford University Libraries, with a non-profit business model.
Collaborative Solutions: LOCKSS

- LOCKSS ingests content from publishers websites using a web crawler similar to those used by search engines.
- Preserves content by continually comparing the content it has collected with the same content collected by other LOCKSS Boxes, and repairing any differences.
- Delivers authoritative content to readers by acting as a web proxy, cache or via Metadata resolvers when the publisher’s website is not available.
Collaborative Solutions: CLOCKSS

- CLOCKSS is a geographically distributed dark archive with which to ensure the long-term survival of Web-based scholarly publications for the benefit of the greater global research community.

- Built on low-cost, open-source, award-winning LOCKSS technology, the CLOCKSS archive comprises a network of redundant nodes located at 12 major research libraries, into which e-content is ingested, copied, and preserved. CLOCKSS's decentralized, geographically disparate preservation model ensures that the digital assets of the community will survive intact.

- Operates as a non-profit business.
Collaborative Solutions: Portico

- Initiated by JSTOR, managed by ITHAKA.

- Portico is a centralized, replicated, and “dark” repository, utilizing a managed preservation methodology.

- Access is provided to participating libraries following a trigger event or post-cancellation access claim.

- Establishes and executes a preservation plan for each journal, which may include an initial migration from publisher specific article XML to the NLM archival standard.
Files Arrive

Files Are Processed

Files Preserved & Replicated

- Validating files against their format specifications
- Verifying checksums
- Extracting key technical and descriptive metadata to record in the preservation metadata
- Capturing all actions taken on the content and recording each as an event in the preservation metadata
- Creating a checksum for each file and recording it in the preservation metadata
- Creating a Portico archival information package
- Analyzing the structure of the content to identify all the files and determine whether all files received were expected.
- Transforming descriptive metadata or structured full-text to an archival format.
Launched in 1997, JSTOR helps libraries provide adequate space for an ever-increasing amount of published scholarship.

Digitizes and preserves printed scholarly journals and provides access through research and teaching platform.

Collaborates with Harvard University and the University of California to support maintenance of print journals in under JSTOR’s digital stewardship.
ITHAKA is a not-for-profit organization that helps the academic community use digital technologies to preserve the scholarly record and to advance research and teaching in sustainable ways.

- Journal content under the stewardship of JSTOR and Portico is treated with similar technological tools, techniques, and best practices.
“Light” vs “Dark” Archive

**Dark Archive**

- PORTICO
- CLOCKSS

- “Trigger Events” dictate access, with the exception of Portico, which in addition provides post-cancellation and audit access:
  - A. Publisher ceases business
  - B. Journal abandoned
  - C. Back issues eliminated

**Light Archive**

- JSTOR: Participation in archive collection ensures access
- LOCKSS: Publisher site no longer provides access.
Scholarly Publisher Participation

- There would be no digital preservation initiatives without publisher participation!
- Demonstrated ongoing willingness to collaborate with LOCKSS, CLOCKSS, and Portico.
Other Digital Preservation Service Providers

**Non-profit / Academic**
- **DPN**
- **ACADEMIC PRESERVATION TRUST**
- **CHRONOPOLIS**
- **HathiTrust**
- **DURACLOUD**

**Commercial**
- **ExLibris**
- **Preservica**
- **Arkivum**

- No specificity for electronic journals. Intellectual property rights guide the scope of preservation targets.
Evaluating Digital Preservation Service Providers
Evaluating Initiatives: Trustworthiness

- In 2006, CRL began certifying trustworthiness of repositories.

- Metrics based on Trusted Repository Archiving Checklist (TRAC):
  A. Organizational Infrastructure
  B. Digital Object Management
  C. Technologies, Technical Infrastructure and Security

- CRL consults a panel of advisers who represent the various sectors of its membership, ensuring that the certification process addresses the interests of the entire community.
Evaluating Initiatives: Trustworthiness

Portico Audit Report 2010

Release Date  Friday, January 1, 2010

Downloads  Portico Audit Report 2010

The Center for Research Libraries (CRL) conducted a preservation audit of Portico (www.portico.org) between April and October 2009 and, based on that audit, has certified Portico as a trustworthy digital repository. CRL found that Portico's services and operations basically conform to the requirements for a trusted digital repository. The CRL Certification Advisory Panel concluded that the practices and services described in Portico's public communications and published documentation are generally sound and appropriate to both the content being archived and the needs of the CRL community. Moreover, the CRL Certification Advisory Panel expects that in the future, Portico will continue to deliver content that is understandable and usable by its designated user community.

This finding is based upon a site visit and sampling of archives content, and upon the review of information gathered by CRL and its Certification Advisory Panel and documents and documentation provided by Portico. CRL's analysis was guided by the criteria included in the Trustworthy Repositories Audit and Certification checklist, and other metrics developed by CRL on the basis of its analyses of digital repositories.

CRL conducted its audit with reference to generally accepted best practices in the management of digital systems, the interests of its community of research libraries; and the practices and needs of scholarly researchers in the humanities, sciences and social sciences in the United States and Canada. The purpose of the audit was to obtain reasonable assurance that Portico provides, and is likely to continue to provide, services adequate to those needs without material flaws or defects and as described in Portico's public disclosures. The CRL audit provides a reasonable basis for these findings.

CRL has assigned Portico the following levels of certification (the numeric rating is based on a scale of 1 through 5, with 5 being the highest level, and 1 being the minimum certifiable level): [1]

<table>
<thead>
<tr>
<th>Category</th>
<th>Portico Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Organizational Infrastructure</td>
<td>3</td>
</tr>
<tr>
<td>Digital Object Management</td>
<td>4</td>
</tr>
<tr>
<td>Technologies, Technical Infrastructure, Security</td>
<td>4</td>
</tr>
</tbody>
</table>
Evaluating Initiatives: Trustworthiness

CLOCKSS Audit Report 2014

Release Date: Tuesday, July 1, 2014
Downloads: CLOCKSS Report 2014

The Center for Research Libraries (CRL) conducted a preservation audit of CLOCKSS (www.clockss.org) between September 2013 and May 2014, and on the basis of that audit certified CLOCKSS as a trustworthy digital repository of e-journal content. The CRL Certification Advisory Panel concluded that the practices and services described in CLOCKSS’ public communications and published documentation generally correspond to the operations of CLOCKSS and are appropriate to the e-journal content being archived and to the expressed needs of the CLOCKSS designated community. Moreover, the panel indicated its expectation that in the future, CLOCKSS will be able to deliver the content it preserves to appropriate third parties who are equipped to make it available for use by the designated community. CRL certification applies to the repository’s ability to preserve and manage digital content deposited by participating e-journal publishers as of May 2014.

The present report is based upon review by CRL and the members of its Certification Advisory Panel, of extensive documentation gathered by CRL independently from open sources and from third parties as well as data and documentation provided by CLOCKSS. The review also included a site visit by CRL audit personnel to the offices of the LOCKSS team in Redwood City, California. CRL’s evaluation of CLOCKSS and the information provided in this report reflect the policies, systems and procedures that were in place at CLOCKSS to manage e-journal content as of June 1, 2014.

On the basis of this evidence, the certification panel concluded that overall CLOCKSS can be recognized by its designated community as a trustworthy repository. However, in the course of the audit, the Certification Advisory Panel identified one issue that CLOCKSS will need to address to more fully satisfy the concerns of its research library constituents: the lack of a formal succession plan. In addition, two aspects of CLOCKSS operations became apparent that should be understood by stakeholders, as they may have a bearing on future CLOCKSS services. Those issues are described in the section titled, Detailed Audit Findings, with reference to the corresponding criteria in the TRAC checklist. CLOCKSS has agreed to address the succession plan issue and also to make certain disclosures to CRL periodically, as a condition of continued certification. Those ongoing requirements are outlined in Section C of this report.

CRL assessed CLOCKSS on each of the three categories of criteria specified in TRAC, and has assigned the level of certification below for each. The numeric rating used is based on a scale of 1 through 5, with 5 being the highest level, and 1 being the minimum certifiable level.

<table>
<thead>
<tr>
<th>TRAC Category</th>
<th>CLOCKSS rating</th>
<th>Optimum rating</th>
</tr>
</thead>
<tbody>
<tr>
<td>Organizational Infrastructure</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>Digital Object Management</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>Technologies, Technical Infrastructure, Security</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td><strong>TOTAL</strong></td>
<td><strong>13</strong></td>
<td><strong>15</strong></td>
</tr>
</tbody>
</table>

The basis for assigning these ratings is provided in Detailed Audit Findings.
Comparing & Contrasting Initiatives
Comparing Initiatives: What’s best for your library?

- Trustworthiness
- Journal Coverage
- Intellectual Property Rights
- Transparency
- Technological approach
- Sustainability
- Cost
Comparisons from Global Library Community

E-JOURNAL ARCHIVING FOR
UK HE LIBRARIES
A White Paper
Version 4.0 Final Draft 10/01/2011
Prepared by:
Charles Beagrie Limited
www.beagrie.com
funded by
JISC

ENSURING PERPETUAL ACCESS:
ESTABLISHING A FEDERATED STRATEGY ON
PERPETUAL ACCESS AND HOSTING OF
ELECTRONIC RESOURCES FOR GERMANY
Edited Final Report – February 2010
Prepared by:
Charles Beagrie Limited in association with Globales Informationstechnik GmbH
www.beagrie.com
A study funded by
The Alliance of German Science Organisations

Final Report of the 2CUL LOCKSS Assessment Team
Cornell University Library & Columbia University Library
Report Completed: March 2011
Public Release: October 2011

Introduction

Although LOCKSS is considered a successful digital preservation initiative, neither of the CSU’s orients that it fully understands the potential of the system for their own settings and collections. There is a range of practical issues that need to be explored in order to broaden this preservation system’s deployment within the CSU’s. In regard to digital content, LOCKSS is being deployed and the adoption of local practices for both CSU’s preservation frameworks. This study is seen as a high-level description to characterize the general landscape and identify further research questions.

The team worked with a continued timeline. November 2010-December 2011, and investigated the following questions:

1. To build a collection of preserved journals beyond the 3-5 journals already preserved via the LOCKSS protocol, what costs (in terms of space, storage, or similar) are incurred in order to preserve the journals in perpetuity?

2. The team conducted a study into the methodologies and tools used by the LOCKSS team to ensure that the digital content is preserved in perpetuity. The study included an analysis of the tools and methods used and the implications for future content preservation.

3. The team conducted a study into the methodologies and tools used by the LOCKSS team to ensure that the digital content is preserved in perpetuity. The study included an analysis of the tools and methods used and the implications for future content preservation.

4. What are the key challenges and opportunities for the continued deployment of LOCKSS within the CSU’s?

Evaluation of the JISC UK LOCKSS Pilot

May 2008

Pete Dalton and Dr Angela Conyers
Evidence Base
Research and Evaluation Services
Library and Learning Resources
Birmingham City University
Comparisons from Global Library Community

The Keepers Registry
Supporting long-term access to journal content

Discover who is looking after your e-journals

Search

Input a search term (e.g. free text, or a title or ISSN(s))

Search

The following organisations are the Keepers of digital content, working on your behalf to ensure long-term access to the scholarly and cultural record. They provide the registry with information on their archival holdings, ordered by most recent update (date of which is shown):

- Archaeology Data Service (09 Jul 2017)
- Global LOCKSS Network (08 Jul 2017)
- CLOCKSS Archive (07 Jul 2017)
- Cariniana Network (06 Jul 2017)
- HathiTrust (05 Jul 2017)
- Portico (05 Jul 2017)
- Scholars Portal (04 Jul 2017)
- British Library (03 Jul 2017)
- Library of Congress (03 Jul 2017)
- PKP PLN (01 Jul 2017)
- Swiss National Library (22 May 2017)
- National Science Library, Chinese Academy of Sciences (13 Feb 2017)
- e-Depot (24 Apr 2012)

The most up-to-date information may be available on the Keeper’s website.
**Snapshot of Portico’s Preservation Progress**

- **80 million** Archival units preserved
- **11%** growth in overall content in the archive in 2017

<table>
<thead>
<tr>
<th>Category</th>
<th>Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>Archival Units Preserved</td>
<td>80 million</td>
</tr>
<tr>
<td>Content Commitment</td>
<td>Total number of e-books and journals committed</td>
</tr>
<tr>
<td>E-books</td>
<td>802,000+</td>
</tr>
<tr>
<td>E-journals</td>
<td>26,000+</td>
</tr>
<tr>
<td>New Publishers Join YTD</td>
<td>47</td>
</tr>
<tr>
<td>Total Participating Publishers</td>
<td>447</td>
</tr>
</tbody>
</table>

- **3%** increase in e-journals preserved
- **3%** increase in e-books preserved
- **1%** increase in d-collection items preserved

---

**Notes:**
- The content committed to the archive includes a total of 802,000+ e-books and 26,000+ e-journals.
- 47 new publishers joined YTD.
- 447 total participating publishers.
<table>
<thead>
<tr>
<th>Journal</th>
<th>ISSN</th>
<th>Publisher</th>
</tr>
</thead>
<tbody>
<tr>
<td>2D Materials</td>
<td>20531583</td>
<td>Institute of Physics Publishing</td>
</tr>
<tr>
<td>3 Biotech</td>
<td>21905738, 2190572X</td>
<td>Springer</td>
</tr>
<tr>
<td>SD ON-LINE</td>
<td>10702192</td>
<td>Association for Computing Machinery</td>
</tr>
<tr>
<td>3D Printing and Additive Manuf.</td>
<td>23297670, 23297662</td>
<td>Mary Ann Liebert, Inc.</td>
</tr>
<tr>
<td>3D Research</td>
<td>20926731</td>
<td>Springer</td>
</tr>
<tr>
<td>3DTV Conference</td>
<td>21612021</td>
<td>IEEE, Inc.</td>
</tr>
<tr>
<td>4OR</td>
<td>16142411, 16194500</td>
<td>Springer</td>
</tr>
<tr>
<td>19th-Century Music</td>
<td>014502076, 15338606</td>
<td>University of California Press</td>
</tr>
<tr>
<td>21st Century Society</td>
<td>17450144, 17450152</td>
<td>Taylor &amp; Francis Group</td>
</tr>
<tr>
<td>24 Images</td>
<td>19235097, 07079389</td>
<td>Consortium Érudit</td>
</tr>
<tr>
<td>AORN Journal</td>
<td>000012092</td>
<td>Elsevier</td>
</tr>
<tr>
<td>A Current Bibliography on</td>
<td>23766662, 00113255</td>
<td>SAGE Publications</td>
</tr>
<tr>
<td>A Life in the Day</td>
<td>13666282</td>
<td>Emerald Group Publishing</td>
</tr>
<tr>
<td>A Practical Logic of Cognitive</td>
<td>10745075</td>
<td>Elsevier</td>
</tr>
<tr>
<td>A.A. Case Reports</td>
<td>23257237</td>
<td>Wolters Kluwer Health</td>
</tr>
<tr>
<td>A.M.A. Journal of Diseases of</td>
<td>15383628, 00966916</td>
<td>American Medical Association</td>
</tr>
<tr>
<td>a/b Auto/Biography Studies</td>
<td>00960975, 21517290</td>
<td>Project MUSE</td>
</tr>
<tr>
<td>AASN Advanced Critical Care</td>
<td>15597768</td>
<td>Wolters Kluwer Health</td>
</tr>
<tr>
<td>AACN Clinical Issues in Critical Care Nursing</td>
<td>10467467</td>
<td>Wolters Kluwer Health</td>
</tr>
<tr>
<td>AACN Clinical Issues: Advanced Practice in Acute and Critical Care</td>
<td>10790713</td>
<td>Wolters Kluwer Health</td>
</tr>
<tr>
<td>AADN in Practice</td>
<td>23251611, 23251603</td>
<td>SAGE Publications</td>
</tr>
<tr>
<td>AAOHN Journal</td>
<td>08910162</td>
<td>SLACK, Inc.</td>
</tr>
<tr>
<td>AAPS PharmSciTech</td>
<td>15309932</td>
<td>Springer</td>
</tr>
<tr>
<td>AASRI Procedia</td>
<td>22126716</td>
<td>Elsevier</td>
</tr>
<tr>
<td>Ab Imperio</td>
<td>21664072, 21649731</td>
<td>Project MUSE</td>
</tr>
<tr>
<td>Abacus</td>
<td>000012072, 14676281</td>
<td>John Wiley &amp; Sons, Inc.</td>
</tr>
</tbody>
</table>
### Portico’s Holdings Comparison Tool

#### ACTA HEMATOLOGICA

- **Publisher**: Karger
- **Identifier**: Content Set ID: ISSN_00015792
- **Completeness**: Years Preserved: 1993, 1996-2016 (v.1-131)
- **Portico Holdings**: 2010 - v. 1 (1-4), 2009/2010/2012 - v. 2 (1-6), 2009/2010 - v. 4 (1-6), 2009 - v. 5 (1-6), 2009 - v. 6 (1-6), 2009 - v. 7 (1-6), 2009 - v. 8 (1-6), 2009 - v. 9 (1-6), 2009 - v. 10 (1-6), 2009 - v. 11 (1-6), 2009 - v. 12 (1-6), 2009 - v. 13 (1-6), 2009 - v. 14 (1-6), 2009 - v. 15 (1-6), 2009 - v. 16 (1-6), 2009 - v. 17 (1-6), 2009 - v. 18 (1-6), 2009 - v. 19 (1-6), 2009 - v. 20 (1-6), 2009 - v. 21 (1-6), 2009 - v. 22 (1-6), 2009 - v. 23 (1-6), 2009 - v. 24 (1-5), 2009 - v. 25 (1-5)

#### Status

<table>
<thead>
<tr>
<th>Volume</th>
<th>Year(s)</th>
<th>Complete</th>
<th>Issue</th>
<th>Articles</th>
</tr>
</thead>
<tbody>
<tr>
<td>v.1</td>
<td>2010</td>
<td>Complete</td>
<td>n.1</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>n.2</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>n.3</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>n.4</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>n.5</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>n.6</td>
<td>8</td>
</tr>
<tr>
<td>v.2</td>
<td>2006/2010/2012</td>
<td>Complete</td>
<td>n.1</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>n.2</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>n.3</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>n.4</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>n.5</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>n.6</td>
<td>8</td>
</tr>
<tr>
<td>v.3</td>
<td>2009/2010</td>
<td>Complete</td>
<td>n.1</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>n.2</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>n.3-4</td>
<td>Complete</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>n.5</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>n.6</td>
<td>8</td>
</tr>
<tr>
<td>v.4</td>
<td>2006</td>
<td>Complete</td>
<td>n.1</td>
<td>7</td>
</tr>
<tr>
<td>v.5</td>
<td>2009</td>
<td>Complete</td>
<td>n.1</td>
<td>8</td>
</tr>
<tr>
<td>v.6</td>
<td>2009</td>
<td>Complete</td>
<td>n.1</td>
<td>8</td>
</tr>
<tr>
<td>v.7</td>
<td>2009</td>
<td>Complete</td>
<td>n.1</td>
<td>8</td>
</tr>
<tr>
<td>v.8</td>
<td>2000</td>
<td>Complete</td>
<td>n.1-2</td>
<td>8</td>
</tr>
<tr>
<td>v.9</td>
<td>2005</td>
<td>Complete</td>
<td>n.1</td>
<td>7</td>
</tr>
</tbody>
</table>

#### Notes

- v.131 (2014): No more issues
- v.132 (2014): No more issues
- v.133 (2014/2015): No more issues
- v.134 (2015): No more issues
- v.135 (2015/2016): No more issues
Portico’s Comparison with C|LOCKSS

<table>
<thead>
<tr>
<th></th>
<th>Portico</th>
<th>CLOCKSS</th>
<th>LOCKSS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total Titles</td>
<td>26,066</td>
<td>20,163</td>
<td>12,897</td>
</tr>
<tr>
<td>Titles Only Here</td>
<td>8,428</td>
<td>2,095</td>
<td>844</td>
</tr>
<tr>
<td>% of CrossRef</td>
<td>48%</td>
<td>38%</td>
<td>25%</td>
</tr>
<tr>
<td>Total Years (Volumes)</td>
<td>488,998</td>
<td>212,347</td>
<td>111,219</td>
</tr>
<tr>
<td>Participating Libraries</td>
<td>1,010</td>
<td>285</td>
<td>u/k</td>
</tr>
</tbody>
</table>
Portico is now actively engaging publishers to update their preservation agreements to allow access to their OA content to be opened, in case of a trigger event, not only to Portico members but to anyone around the world.
Why Participate in Digital Preservation Initiatives?
Value of Investment in e-Journal Digital Preservation

- To ensure your library’s secure and reliable transition from print to a reliance on electronic journals.
- Protects library’s investment in electronic journals.
- Facilitates deaccessioning of print journals, thereby freeing up space in the library for collaborative work areas.
What if we don’t participate...
Q&A

Ken DiFiore
Outreach Director

ken.difiore@portico.org
www.portico.org