Center for Research Libraries

GLOBAL RESOURCES NETWORK

Certification Report on CLOCKSS

Executive Summary

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 hereby certifies CLOCKSS as a trustworthy digital repository of e-journal content. The CRL Certification Advisory Panel has 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 expects 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 the use of 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 Section B, Detailed Audit Findings, below, 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.

About the Audit Participants

CLOCKSS

CLOCKSS (www.CLOCKSS.org) is a not-for-profit (501c3) organization, incorporated in the State of California. CLOCKSS operates as a joint venture based at Stanford University and supported by academic publishers and research libraries. A Board of Directors, whose members are drawn in equal numbers from the supporting libraries and participating publishers, is the organization's governing body. An Executive Director directs the work of CLOCKSS staff. The organization develops and maintains a geographically distributed, dark archive that preserves web-based scholarly publications. CLOCKSS uses the LOCKSS technology, developed at Stanford, to preserve e-journal and e-book content for publishers and the academic community to prevent loss of that content in the event that direct access from the publisher is discontinued for any reason. CLOCKSS also delivers archived content back to its original publisher on request in the event of data loss by the publisher.

CENTER FOR RESEARCH LIBRARIES

The Center for Research Libraries (CRL - www.crl.edu) is an international consortium of university, college, and independent research libraries. CRL supports advanced research and learning in the humanities, sciences, and social sciences by ensuring the survival, integrity, and accessibility of source materials vital to those disciplines. In order to enable its community to accelerate the shift to electronic-only resources in a cautious and responsible manner, CRL both preserves and maintains shared physical collections of materials and evaluates digital repositories of interest to its community.

Analysis of CLOCKSS documentation and operations was undertaken by CRL staff. Additional technical support for assessment of the CLOCKSS repository systems and architecture was provided by James A. Jacobs.

CRL CLOCKSS CERTIFICATION ADVISORY PANEL

To guide its CLOCKSS audit, CRL enlisted a panel of advisors representing the various sectors of the academic research libraries world. The Certification Advisory Panel included leaders in collection development, preservation, library administration, and digital information technology, and is so constituted as to ensure that the certification process addresses the interests of the entire CRL community.

THE MEMBERS OF CRL'S CLOCKSS CERTIFICATION ADVISORY PANEL WERE:

PERRY WILLETT (CHAIR)
Digital Preservation Services Manager
California Digital Library

WINSTON ATKINS
Preservation Officer
Duke University

PASCAL CALARCO
Associate University Librarian, Digital & Discovery Services
University of Waterloo

MALIACA OXNAM
Associate Librarian, Digital Content and Services (DCS)
University of Arizona

OYA Y. RIEGER
Associate University Librarian for Information Technologies
Cornell University

A. Audit and Assessment Methodology and Criteria

This assessment was undertaken to determine whether or not CLOCKSS meets the commitments it has made regarding the long-term preservation of e-journal content for the research community, and whether the repository's operations comply with established criteria for trusted digital repositories. The assessment included a review of information independently gathered by CRL from published and unpublished sources, a review of documents and documentation provided by CLOCKSS, and a site visit to test and verify certain repository processes and functions.

CRL conducted its audit with reference to:

- generally accepted best practices in the management of digital systems
- the interests of the CRL community of research libraries
- the practices and needs of scholarly researchers in the humanities, sciences, and social sciences in the United States and Canada
- the criteria enumerated in *Trustworthy Repositories Audit & Certification: Criteria and Checklist*[1]
- the criteria included in *Audit and Certification of Trustworthy Digital Repositories (TDR) checklist (ISO 16363)*[2]
- the Open Archive Information System reference model[3] (OAIS)
- other metrics developed by CRL in its analyses of digital repositories.

The primary metrics used by CRL in its assessments are those specified in the *Trustworthy Repositories Audit and Certification (TRAC)* checklist. TRAC was developed by a joint task force formed by the Research Libraries Group (RLG) and the National Archives and Records Administration in 2003 to provide criteria for use in identifying digital repositories capable of reliably storing, migrating, and providing long-term access to digital collections. TRAC represents best current practice and thinking about the organizational and technical infrastructure required for a digital repository to be considered trustworthy and thus worthy of investment by the research and research library communities. The approved ISO standard for Trustworthy Digital Repositories (ISO 16363) was also used in this audit. Because there is currently no ISO-approved mechanism for accrediting certifying bodies for the TDR standard, CRL's certification is to TRAC criteria.

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.

TRAC CATEGORY	CLOCKSS RATING	OPTIMUM RATING
Organizational Infrastructure	4	5
Digital Object Management	4	5
Technologies, Technical Infrastructure, Security	5	5
TOTAL	13	15

The basis for assigning these ratings is provided in Section B, Detailed Audit Findings, below.

It should be noted that CRL certification of CLOCKSS applies specifically to the repository's ability to preserve and manage in digital form e-journals contributed by publishers. CRL did not assess other types of content preserved by CLOCKSS.

1 TRAC - http://www.crl.edu/sites/default/files/attachments/pages/trac_0.pdf
2 TDR - http://public.ccsds.org/publications/archive/652x0m1.pdf
3 OAIS - http://public.ccsds.org/publications/archive/650x0m2.pdf

B. Detailed Audit Findings

The CLOCKSS Archive is a joint venture of publishers and libraries. Several major publishers of electronic journals, including Elsevier, Springer, Taylor & Francis, and Wiley-Blackwell, enable CLOCKSS to preserve the article contents of their journals on an ongoing basis. As of May 2014 CLOCKSS contained 5,771,160 articles from 13,135 titles, by 198 publishers.

Publishers provide their e-journal content to CLOCKSS for archiving. This is done in one of two ways: by allowing CLOCKSS to harvest that content directly from the publisher's website, or by file transfer. With harvest, CLOCKSS crawls the publisher's site and harvests the same content that the publisher makes available online to readers. A crawl generates a submission information package (SIP) consisting of the journal content and appropriate metadata. With file transfer, an FTP or "rsync" or other file transfer mechanism is used to transfer "packages" of content and metadata from the publisher to CLOCKSS.

With both harvested and transferred content, each SIP typically represents the articles published since the previous harvest or transfer. The unit archived by CLOCKSS typically contains all article content published by the publisher during a defined period of time (such as a year or a volume of a journal) plus files containing metadata related to that content.

Ingested content is then stored as the original bits on a global network of 12 "nodes," repositories maintained by participating universities, libraries, and other organizations, each of which has certain specified obligations to CLOCKSS. The nodes, located in the U.S. (5 nodes), Canada, the United Kingdom, Germany, Italy, Japan, Hong Kong, and Australia, are each obliged to store a complete version of the CLOCKSS Archive content. The nodes use LOCKSS technology to automatically and continually compare or "audit" their content against that held in the other nodes, and repair any differences.

In the event that access to the content through the publisher is disrupted for an extended period of time, CLOCKSS is authorized through its contracts with publishers to copy and transfer the content from the CLOCKSS Archive to selected host organizations. The host organizations agree to make the content available to the general public without charge under a Creative Commons license (or equivalent license). The University of Edinburgh and Stanford have agreed to serve as hosts and re-publish triggered content. It is important to note that the Creative Commons license permits anyone to re-publish such content and, indeed, some triggered content is in the Internet Archive (Annals of Clinical Psychiatry <http://web.archive.org/web/*/http://www.clockss.org/clockss/Annals_of_Clinical_Psychiatry>).[4]

These activities are governed by a written contract between CLOCKSS and each publisher. The contract grants to CLOCKSS archiving and certain specified re-publishing rights, and binds the publisher to providing to CLOCKSS specified content and accompanying metadata, and a specified level of monetary support on an annual basis.

In its audit, CRL determined that the CLOCKSS system operates as represented; appears to be generally well-designed and adequate to the preservation of the e-journal content currently archived; and is rigorously maintained. The governance of the effort is structured to ensure accountability to CLOCKSS' two major stakeholder communities: e-journal publishers and academic libraries. One of the strengths of CLOCKSS, in fact, is the deep engagement of the research library community in its planning and governance. This engagement is likely to ensure CLOCKSS' continued responsiveness to the needs of that community. The CLOCKSS funding model, moreover, is designed to enable the program to respond to changes in the amount, nature and value of the content archived.

4 The circumstances under which content can be "re-published" by CLOCKSS are specified in the standard contract between CLOCKSS and publishers, as when either: "(i) the owner of all rights to the Archived Content (including the copyrights) gives unconditional consent to the release of such Archived Content to the general public, or (ii) the Archived Content is determined in good faith by the Board to be unavailable from any publisher for at least six consecutive months."

The audit identified one issue that CLOCKSS will need to address in order to more fully satisfy the concerns of research library constituents: the lack of a formalized succession plan. The TRAC checklist specifies that a repository should have "an appropriate, formal succession plan, contingency plans, and/or escrow arrangements in place in case the repository ceases to operate or the governing or funding institution substantially changes its scope" (TRAC A1.2). At the time of the audit there was no designated successor organization for CLOCKSS. While discussions had occurred with both OCLC and Stanford University Libraries regarding serving as successor organizations, the details of such an arrangement had not yet been formalized. CLOCKSS has agreed to address the succession 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.)

In addition, two notable aspects of CLOCKSS operations became apparent in the audit that should be understood by current and prospective stakeholders. While not problematic enough to prevent certification, these matters could possibly have a bearing on future CLOCKSS services. The two notable aspects are described below with reference to the corresponding criteria in the TRAC checklist.

NOTABLE ASPECTS OF CLOCKSS

1. Repository has short- and long-term business-planning processes in place to sustain the repository over time. (TRAC A4.1)

   The CLOCKSS funding model is designed to enable the enterprise to respond to changes in the nature, value and amount of content archived. Each year, publishers pay a "means-based" annual fee, which is scaled to their total publishing revenue; plus a per-article fee, based on the amount of content archived that year. This price structure enables CLOCKSS to absorb the growing costs of content management to a certain extent. However, it is conceivable that as the cost of ingest and management of content inevitably increases with the amount and complexity of the content being managed those costs could require CLOCKSS to seek greater revenue from libraries and or publishers.

2. Repository has a documented process for testing understandability of the information content and bringing the information content up to the agreed level of understandability. (TRAC B2.10)

   CLOCKSS warrants that it will ensure that the journal articles in its archive, once ingested, will continue to be "understandable" at the level of understandability that they possessed at the time of ingest. That warranty is based on four assumptions:

   a) E-journal publishers create understandable, renderable content deliverable through web browsers; and, should problems with that content occur, readers will detect and report them, and publishers will correct.

   b) Web browsers will continue to be the primary rendering tool for e-journal content and will continue to render old web content as well as new web content over time. Formats that are not intended to be rendered by web browsers (such as Microsoft Office formats) are widely supported.

   c) The rendering of those files in the archive in discipline-specific formats that are not intended to be displayed in a web browser is considered by CLOCKSS "a problem for the specific field" and not something for which an archive can provide a generic solution.

   d) Emulation, rather than format migration, will be increasingly easy, robust and affordable and may be the preferred way to deliver content in an obsolete format if obsolescence ever occurs.

Assumption "a," that successful exposure of the actual journal content on the web is a guarantee of the renderability of that content, does not apply, however, to content ingested by CLOCKSS through file transfer, rather than direct web harvest. Yet in the view of the auditors, this strategy is technically reasonable and justifiable. CLOCKSS staff actively monitors work in the fields of digital preservation, format migration, and emulation to support this strategy. As evidence of that, CLOCKSS made minor changes to its own policy of dealing with potential file format obsolescence during the course of the audit.

The strategy is also prudent in terms of resource expenditure for a dark archive. Tracking formats over time and migrating them can be costly in terms of programming and development resources, computing time, data management, and disk storage. It is therefore reasonable to assume that dealing with what are likely to be a relatively small number of obsolete formats only once, at the time of a trigger event or at time of delivery from a re-publishing site, with the technologies available at that time, may be a wiser use of resources than constantly and repeatedly monitoring and migrating un-triggered content in a dark archive. The current state of technology suggests that these strategies will work now and may improve in the future.

OTHER FINDINGS

One additional area of concern is a practice that, although the norm among digital repositories, is the limitation of the right to re-publish triggered content. That is the lag time between a "trigger event" and the time at which CLOCKSS may republish triggered content by CLOCKSS without the publisher's permission. The lag time of up to six months specified in CLOCKSS' agreements with the publishers, although the norm with other repositories including Portico, is not likely to be acceptable in fields such as medicine, where a hiatus of such duration would have a greater impact on users than a comparable disruption in access to a journal in the humanities or social sciences. However, the lead time the CLOCKSS Archive currently requires for the technical process of triggering content is only 2 to 4 weeks, and CLOCKSS has demonstrated its ability to republish triggered content, with the agreement of the publisher, within that period. As reasonable over time, the archive should endeavor to tailor agreements with publishers to better accommodate use cases in all fields.

Re-publishing triggered content is not a core function of the CLOCKSS archive. Two institutions have agreed to serve as "host organizations" for such content: Stanford University Libraries and the University of Edinburgh's EDINA. The host organizations agree to "re-publish" the released content on the open web under a Creative Commons license that allows it to be re-hosted freely. It is then expected that the content will henceforth be maintained and made available by one or more additional organizations that have an interest in sustaining the material.

There will inevitably be a cost involved in the successful release and re-publishing of significant amounts of triggered content. Those expenses will have to be borne by the re-publishing host organizations, by the CLOCKSS community, or both. The cost will depend on the amount, complexity, and the nature of the use of that material, and could thus be trivial for small amounts of low-use content, or quite large if a trigger event, or series of such events, releases a huge number of articles from many popular journals. For that reason, it would be prudent for CLOCKSS management to develop detailed scenarios for how such services might be obtained and paid for.

It should also be noted here that CRL was not able to independently and comprehensively verify and monitor the presence and integrity of content in the CLOCKSS repository at a meaningful level of granularity. While such verification and monitoring is a challenge inherent in "dark" archives, which are unable to be accessed for such purposes, practices for such are emerging. CLOCKSS submits title- and volume-level metadata to both the Keeper's Registry and KBART. However, to improve transparency, and enable libraries to more confidently rely on CLOCKSS archiving, CLOCKSS has also agreed to provide issue-level metadata on its content to CRL, or to expose such metadata for CRL harvest periodically.

RATING

CRL assessed CLOCKSS on each of the three categories of criteria specified in TRAC and has assigned a level of certification for each. The numeric rating (below) is based on a scale of 1 through 5, with 5 being the highest level, and 1 being the minimum certifiable level. (The minimal certification rating of 1 is assigned in instances where a repository has inconsistencies or deficiencies in areas that might lead to minor defects of a systemic or pervasive nature, but where no major flaws are evident.)

TRAC CATEGORY	CLOCKSS RATING	OPTIMUM RATING
Organizational Infrastructure	4	5
Digital Object Management	4	5
Technologies, Technical Infrastructure, Security	5	5
TOTAL	13	15

C. Ongoing Requirements

The TRAC document notes that "attaining trusted status is not a one-time accomplishment—achieved and forgotten. To retain trusted status, a repository will need to undertake a regular cycle of audit and/or certification." To that end, CRL expects that CLOCKSS will also make certain disclosures on a regular basis. CRL and CLOCKSS have agreed that ongoing certification is contingent upon CLOCKSS making the following disclosures every three years:

- A detailed listing of new content added to the repository since certification;
- Description of any significant changes in repository system architecture or configuration, operating systems and/or critical software;
- New agreements and contracts with key depositors of content, content users, major funders or sources of revenue, and providers of critical repository services;
- New key policies regarding acquisition, management, and disposition of archived content and related files and metadata;
- Records of significant events (such as content migrations, system failures, loss or corruption of digital content) and significant changes in the characteristics of digital content ingested since the most recent audit; and of significant events and changes in the operations of the repository;
- The most recent three years of financial statements for the repository organization or service unit. The financial statements should indicate the categories and, where appropriate, sources of revenue and the level of same; the functional allocation of expenses; and changes in the financial position of the organization supporting the service unit;
- Revenue and expense projections by function, for the repository organization or service unit, for the next three years.

Certification is also contingent upon CLOCKSS agreement to a periodic, systematic sampling and/or inspection of its metadata for the repository's archived content by CRL, or by a third party designated by CRL and CLOCKSS jointly, using either a manual or an automated process, as determined by mutual agreement between CRL and CLOCKSS.