Center for Research Libraries

GLOBAL RESOURCES NETWORK

Certification Report on the HathiTrust Digital Repository

Executive Summary

The Center for Research Libraries (CRL) conducted a preservation audit of HathiTrust (www.HathiTrust.org) between November 2009 and December 2010, and on the basis of that audit certifies HathiTrust as a trustworthy digital repository. *The CRL Certification Advisory Panel has concluded that the practices and services described in HathiTrust public communications and published documentation are generally sound and appropriate to the content being archived and to the general needs of the CRL community*. Moreover the Panel expects that in the future, HathiTrust will continue to be able to deliver content that is understandable and usable by its designated community.

CRL certification applies to the repository's ability to preserve and manage digital files of books digitized by the University of Michigan, Google, and the Internet Archive, as well as the digital files generated from books digitized by other providers that conform to comparable standards.

This certification is based upon review by CRL and members of its Certification Advisory Panel of extensive documentation gathered by CRL as well as data and documentation provided by HathiTrust between November 2009 and December 2010; and a site visit held in May 2010. CRL's analysis was guided by the criteria included in the Trustworthy Repositories Audit and Certification Checklist (TRAC), and other metrics developed by CRL through its various digital repository assessment activities.

CRL conducted its audit with reference to generally accepted best practices in the management of digital systems; and with reference to 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 HathiTrust provides, and is likely to continue to provide, services adequate to those interests and needs without material flaws or defects and as described in HathiTrust's public disclosures. The CRL audit provides a reasonable basis for these findings.

CRL assigns HathiTrust 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]

CATEGORY	HATHITRUST RATING
A. Organizational Infrastructure	2
B. Digital Object Management	3
C. Technologies, Technical Infrastructure, Security	4

The ratings reflect the existence of robust systems and sound processes in most areas, in particular those areas addressed in TRAC category C and, to a lesser extent, in category B metrics; and the still emerging status of systems and processes addressed in category A metrics. In the course of the audit, the Certification Advisory Panel identified several issues that CRL urges HathiTrust to address to more fully satisfy the concerns of CRL libraries. Those issues are described in Section B, Detailed Audit Findings, below, and pertain to specific criteria in the TRAC checklist. HathiTrust has agreed to address these issues and, as a condition of continued certification, to make certain disclosures to CRL periodically. Those requirements for periodic disclosure are outlined in Section C of this report.

[1] A working version of the schema CRL uses in providing summary ratings of a repository's compliance with the TRAC criteria is available at: http://www.crl.edu/archiving-preservation/digital-archives/metrics-assessing-and-certifying/crl-ratings.

About the Audit Participants

HATHITRUST

The HathiTrust (www.HathiTrust.org) digital repository was established by the University of Michigan Libraries in October 2008. HathiTrust offers academic and research libraries the opportunity to combine resources to build and maintain a large scale repository. Participating members include research libraries in the United States and Europe. The majority of HathiTrust's content was digitized through the Google Books project, which has worked with libraries to scan books on library shelves. Additional content comes from the Internet Archive and from collections digitized by partner libraries. As of December 2010, HathiTrust had ingested approximately 7.5 million volumes. Repository costs are shared by the participating members. The actual repository systems are located at the University of Michigan in Ann Arbor and at Indiana University in Bloomington.

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 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 careful and responsible manner, CRL has a hybrid strategy of preserving and maintaining shared physical collections of materials and certifying digital repositories of interest to its community.

CRL analysis of HathiTrust documentation and operations was undertaken by Marie Waltz and other CRL staff. Additional technical support for the site visit and the assessment of HathiTrust repository systems and architecture was provided by James A. Jacobs, Data Services Librarian Emeritus, University of California, San Diego.

To guide its HathiTrust audit CRL enlisted a panel of advisors representing the various sectors of its membership. The Certification Advisory Panel includes 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.

MARTHA BROGAN (CHAIR)
Director of Collection Development & Management
University of Pennsylvania

WINSTON ATKINS
Preservation Officer
Duke University

BART HARLOE
Director of Libraries
St. Lawrence University

WILLIAM PAROD
Senior Repository Developer
Northwestern University Libraries

ANNE POTTIER
Associate University Librarian
McMaster University

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

PERRY WILLETT
Digital Preservation Services Manager
California Digital Library

A. Audit and Assessment Methodology and Criteria

This assessment was undertaken to determine whether or not HathiTrust meets the commitments it has made in regard to the long-term preservation of digital scholarly content for the academic community and whether the repository complies with established criteria for trusted digital repositories. The assessment included a site visit, a review of the information independently gathered by CRL from published and unpublished sources, and a review of documents and documentation provided by HathiTrust.

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 included in Trustworthy Repositories Audit & Certification: Criteria and Checklist[2]
- the Open Archive Information System reference model[3] (OAIS)
- other metrics developed by CRL through its analyses of digital repositories.

The general metrics used by CRL in such assessments are based on the Trustworthy Repositories Audit and Certification (TRAC) checklist, and on other metrics developed by CRL through its analyses of digital repositories. TRAC was developed by a joint task force, created 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 to be considered trustworthy and thus worthy of investment by the research and libraries communities.

CRL assessed HathiTrust 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:

TRAC CATEGORY	HATHITRUST RATING
Organizational Infrastructure	2
Digital Object Management	3
Technologies, Technical Infrastructure, Security	4

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

It should be noted that CRL certification of HathiTrust applies specifically to the repository's ability to preserve and manage digital files of books digitized by the University of Michigan, Google, and the Internet Archive, as well as the digital files generated from books digitized by other providers that conform to comparable standards. *CRL did not assess SDR procedures and processes for acquiring and managing more complex digital objects such as audio, video, archived Web sites, or other types of content.*

[2] TRAC - http://www.crl.edu/sites/default/files/attachments/pages/trac_0.pdf3 OAIS - http://public.ccsds.org/publications/archive/650x0b1.pdf
[3] OAIS - http://public.ccsds.org/publications/archive/650x0b1.pdf

B. Detailed Audit Findings

On the basis of the audit, CRL identified areas in which HathiTrust will need to improve processes or provide greater disclosure of information about those processes. These areas correspond to specific TRAC criteria or to features of the repository that members of the Certification Advisory Panel believe are important to the CRL community.

The specific areas identified for improvement are:

1. DEFINITION OF RIGHTS AND OWNERSHIP OF HATHITRUST ENTERPRISE ASSETS (TRAC CRITERIA A3.3, A3.7, AND A4.3)

"Repository commits to transparency and accountability in all actions supporting the operation and management of the repository, especially those that affect the preservation of digital content over time." (TRAC criterion A3.7)

HT is not a separate legal entity, and so cannot legally own the capital equipment, content, metadata, and other assets acquired or generated by the partnership. HathiTrust therefore must clearly define and establish where ownership and control of the repository's major assets and other property essential to the continued access and preservation of repository content reside.

The ownership of the individual objects within the repository is clearly specified in agreements with the depositors. However, it would be appropriate to clarify the ownership of the aggregate HathiTrust database; the rights and ownership of new content (such as derivative files, metadata, etc.); and rights to the non-content assets of the operation, including software and system tools to be developed in the future.

2. SUCCESSION OR DISPOSITION PLAN FOR HATHITRUST ASSETS (TRAC A1.2)

TRAC A1.2 requires "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."

Most 501c3 organizations provide in their bylaws for the disposition of property in the event of the organization's dissolution. HathiTrust should explicitly address the transferability or disposition of its assets in the event of discontinuation of repository operations. Under present circumstances the University of Michigan, as parent organization of the repository, would seem to be the owner of those assets and arbiter of such decisions, absent agreements to the contrary.

3. CLARIFY AND STRENGTHEN THE QUALITY ASSURANCE AND PRINT ARCHIVING COMPONENTS OF THE HATHITRUST PROGRAM (TRAC A3.8 AND B1.1, B1.8, B1.7 AND B2.4)

TRAC criterion A3.8 requires that the repository "commits to defining, collecting, tracking, and providing, on demand, its information integrity measurements." Criterion B1.1 requires that the "Repository identifies properties it will preserve for digital objects," and B1.7, 1.8, and 2.4 address the recording of information about rejected SIPs.

One explicit goal described in the HathiTrust mission statement is to "coordinate shared storage strategies among libraries, thus reducing long-term capital and operating costs of libraries associated with the storage and care of print collections." The repository should put in place and clarify its plan for achieving that goal, as the cost reduction described is a relevant metric of the value of HathiTrust and its services. The new HathiTrust pricing model, to be introduced in 2013, will directly correlate the overlap between the repository corpus and the print holdings of the participating libraries. This will increase pressure for participating libraries to divest of print volumes available through the repository.

The quality assurance measures for HathiTrust digital content do not yet support this goal. Inspection criteria and standards are in place for materials ingested from the Google Books project, but it is not clear what results when an object fails such inspection. It is also unclear to what level of quality review materials digitized by partner institutions

or those made available through entities such as the Internet Archive are subjected. This will be material to library decisions on whether to retain, conserve, or dispose of corresponding physical copies of books represented in the repository.

Currently, and despite significant efforts to identify and correct systemic problems in digitization, HathiTrust only attests to the integrity of the transferred file, and not to the completeness of the original digitization effort. This may impact institutions' workflow for print archiving and divestiture.

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 in addition to acting to remedy the issues identified above HathiTrust will also make certain disclosures on a regular basis. CRL and HathiTrust have agreed that ongoing certification is contingent upon HathiTrust making the following disclosures every two years:

- An item- or volume-level listing of new content added to the repository since prior certification. (Public disclosure through the current HathiFiles inventory would fulfill this requirement.)
- Description of any significant changes in repository system architecture or configurations, 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, such as server logs; 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 HathiTrust's agreement to a periodic, systematic sampling and inspection of the repository's archived content by CRL, or by a third party designated by CRL, using either a manual or automated process as determined by mutual agreement between CRL and HathiTrust.