Revised 8 April 2019

HathiTrust is intended to provide persistent and high-availability storage for deposited files. In order to facilitate this, the partnership uses a storage architecture with a rich set of features designed for fault tolerance and long-term data retention.

Central to the storage architecture is the use of two synchronized instances of storage with wide geographic separation (located in data centers in Ann Arbor, MI and Indianapolis, IN) and an encrypted tape backup with 6 months of previous-version retention (located in a third data center several miles from the Ann Arbor storage instance). All data centers meet the requirements for Uptime Institute Tier II classification. All storage is physically secure, locked in racks that are accessible only to authorized IT personnel.

The need for continuous integrity checking is fundamental to HathiTrust's data management strategy and underlies the choice of online (spinning magnetic disk) media for primary storage. Internally, each storage instance uses N+3 Reed-Solomon parity redundancy, which is analogous to but more fault-tolerant than conventional RAID 5 storage due to the additional parity redundancy. The storage system internally performs in-flight data integrity checks as well as periodic integrity checks of all at-rest data, and makes use of parity redundancy to permanently repair any errors encountered. External to the storage system, HathiTrust also conducts periodic validation of data with stored checksums to ensure that data has been ingested correctly and remains intact.

Storage equipment is typically refreshed every 4-5 years. The storage system is modular and virtualized, with files split into blocks that are distributed across nodes of a cluster and automatically redistributed as needed to balance storage utilization equally. Storage nodes that have reached retirement age may be removed from the cluster with an administrative command, and new nodes may be added, with all movement of data managed internally while employing the in-flight integrity checks described earlier. The remove and add processes neither disrupt services nor diminish the N+3 redundancy.

The following links provide more detailed information about our storage, backup, and disaster planning:

- A Preservation Infrastructure Built to Last: Preservation, Community, and HathiTrust
- Building a Future by Preserving Our Past: The Preservation Infrastructure of HathiTrust Digital Library
- HathiTrust is a Solution: The Foundations of a Disaster Recovery Plan for the Shared Digital Repository
- HathiTrust Trustworthy Repository Audit and Certification compliance

HathiTrust Digital Library Profile

A profile of the repository based on the Evaluation of Open-Source Electronic Publishing Systems (Cyzyk and Choudury, 2008) and the framework developed at Johns Hopkins

University as part of the Mellon-funded grant, A Technology Analysis of Repositories and Services (2006) is given below. Links to information about specific components of HathiTrust's technological infrastructure are included.

1) Institutional affiliation and other indicators of the viability of the project

Name of system	HathiTrust Digital Library
Current version of system	HathiTrust is comprised of multiple applications that are managed with version control and frequently deployed as features and bug fixes are done, but there is no meaningful overall version of those applications or of HathiTrust as a whole.
Tested version of system	See above
URL of project homepage	https://www.hathitrust.org
Institutional affiliation	Membership of academic and research libraries
Age of project	Launched October 2008
Notes on long-term viability of project	HathiTrust is a membership organization of academic and research libraries (see https://www.hathitrust.org/community for a list of members). It is supported with annual member fees from these institutions, with occasional grants or other temporary funding.
Degree of deployment	Repository is located at the University of Michigan in Ann Arbor, Michigan, with a full mirror site including load balancing and fail-over at Indiana University's Indianapolis campus.
Type of open-source license	HathiTrust as a whole does not have an open source license. Some applications and components are available via GitHub at https://github.com/hathitrust/ under a 3-Clause BSD license.
Licensing notes	---

Other documentation (Webliography)	Information about HathiTrust, including its mission and goals, governance, and objectives, as well as partnership information, papers and presentations, and documentation of rights management and preservation policies and procedures, APIs, accountability considerations and technical infrastructure are available at https://www.hathitrust.org/about.

2) Technical requirements, maintenance, scalability, and documented APIs

Local install or ASP?	There is one implementation of HathiTrust with mirroring in Indiana.
Operating system requirements	Linux
Hardware requirements	Commodity Intel-based servers
Application server requirements	N/A
Web server requirements	Apache, HAProxy
Primary programming language	Perl
Auxiliary programming language	Java, Ruby, JavaScript, PHP

Application framework	System and software development in HathiTrust is driven by the need to solve particular problems (as opposed to implementing specific software). This has resulted in a modular architecture where discrete systems fulfilling different OAIS functions (e.g., object ingest, storage, metadata management, indexing and dissemination) communicate and interoperate as an integrated whole. Disaggregation of the functional components of the repository allows agile response to problems that arise (e.g., issues with ingest, storage, or access systems are localized and may be addressed separately) and sharing of development responsibilities across several member institutions. Although many repository systems and services sit on central servers, the modular architecture and orientation toward open standards and open systems make it possible for member institutions to develop services and key pieces of repository functionality.
Database server requirements	MySQL
Other software requirements	Other software and tools: Solr; ZIP; JHOVE; Kakadu; NetPBM The University of Michigan Library would like to acknowledge the generous provision of a source code license by Kakadu Software which is instrumental in the creation, maintenance, and delivery of JPEG2000 images in HathiTrust. The US Federal Documents Registry depends upon Ruby, MongoDB, Blacklight, and Solr.
Required skills	Significant knowledge of Linux, Perl, Apache, MySQL, Ruby
Internal backup and restore functions	Backup and restore functionality is provided at a system level and consists of a) file system backup and b) database backup. Backup services are currently provided by IBM Spectrum Protect (formerly known as IBM Tivoli Storage Manager).
Scalability: Application	Applications are lightweight and served by multiple web servers; additional web servers can be added to increase application performance.
Scalability: Data	HathiTrust uses Isilon storage, which is a clustered storage system that scales to over 20TB in a single instance by adding new nodes to the storage cluster.
API: Code extensibility	APIs exist for retrieving content and bibliographic metadata from HathiTrust. Some code is available via GitHub at https://github.com/hathitrust/ and is open to suggested modifications.

API: Batch ingest	There is no public API as such for batch ingest. Batch ingest is handled automatically for material digitized by Google and Internet Archive. Specifications and tools are provided for partners to package material for deposit.
API: Batch ingest formats	ITU G4 TIFF, JPEG2000, and Unicode OCR, with accompanying METS or YAML metadata.
API: Batch export	The HathiTrust Data API is used to retrieve object packages, including image files and metadata, for individual volumes or batches of volumes from the repository. Specifications for the API are available at https://www.hathitrust.org/data_api.
API: Batch export formats	Formats stored in the repository are exported through the Data API.
API: Support for JSR 170	HathiTrust does not support JSR 170
API: Support for OAI harvesting	Public domain records in HathiTrust can be harvested via Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH). Details can be found at https://www.hathitrust.org/data.
API: Support for eduSource Communication Layer (ECL)	HathiTrust does not support ECL
API: Support for other Web services	HathiTrust makes limited metadata files for all volumes in HathiTrust available via the web for download (https://www.hathitrust.org/hathifiles). This metadata can be used to retrieve full bibliographic records from OCLC or the University of Michigan via Z39.50.

Security	
Access control	Access to items is determined by copyright status and is handled through the HathiTrust PageTurner application. A description of the PageTurner application is available at https://www.hathitrust.org/pageturner.

User management	HathiTrust manages lists of staff members at partner institutions with privileged access to repository content and lists of users ids for personalized services (e.g., Collection Builder).
Policy management	HathiTrust adheres to the information technology security policies of the University of Michigan Library, where it is hosted. The University Library participates in distributed organizational model where units across the University (of which it is one) have prime responsibility for planning and managing security within their units, coordinated by campus Information Assurance (IA).

Storage	
Add data	Content is added to HathiTrust via "feed", a batch ingest application that automatically queues and manages content digitized by Google, Internet Archive, and member institutions.
Access data	Content is accessed via the HathiTrust PageTurner application. Public domain volumes and works that rights holders have opened access to, are available in full view to anyone with a web browser. In-copyright works and those with undetermined copyright status are searchable only (the search application returns location information where query terms occur in a given volume). The HathiTrust Data API is another mode of accessing content in HathiTrust (https://www.hathitrust.org/data_api).
Remove data	The files that compose digital objects are contained in a directory in a file system. When objects are deleted due to rightsholder requests, content files are deleted and a tombstone record is made available in the user interface to indicate that the content once existed.
Manage metadata	Bibliographic metadata is managed by Zephir, a purpose-built metadata management application for HathiTrust hosted by the California Digital Library. Rights information is managed in a rights database (https://www.hathitrust.org/rights_management). Preservation, technical, and structural metadata are contained in a METS file for each object. Preservation metadata (PREMIS) is updated when actions occur on an object.

Aggregation	

Create aggregation	The Collection Builder application allows users to create their own aggregations of objects, regardless of how they are structured in the repository. Internally, objects are identified by content provider (who provided the content to HathiTrust), responsible entity (who has custodial responsibility for the content in HathiTrust), and digitization agent (who scanned the item) for reporting and aggregation purposes.
Remove aggregation	Personal collections of volumes created in Collection Builder can be deleted.
Change aggregation membership	Objects can be copied or moved from one Collection Builder collection to another. Updates to metadata in Zephir can change the internal aggregations to which an object belongs.
Find aggregation members	It is possible to facet search results in the HathiTrust catalog to limit results to objects from a particular content provider. It is possible to find and search Collection Builder collections through the web interface.

Other	
Locking	Ingest tracking ensures only one process is updating an object at a time via database locking.
Virtual object representation	TIFF and JPEG2000 images in the repository are dynamically converted to PNG format for viewing in the PageTurner.
Transactions	HathiTrust is configured to allow large-scale transactions on the content. Some of these that have taken place are the modification of METS and PREMIS in object packages across the repository. Objects are routinely zipped and unzipped for ingest purposes and display in the HathiTrust interface.

3) Submission, peer review management, and administrative functions

Support for multiple, discrete publications	As of April 2019, HathiTrust contains approximately 17 million items.
Multiple administrative roles	The ability to change repository code and content is managed with Unix permissions held by a very limited group of developers and administrators. There are multiple user roles that have elevated access to the repository for administrative purposes such as content and metadata quality control.
Administrative roles configurable	Developer and system administrator roles are configurable according to Unix permissions. Application roles are configurable via the application code.
Submission into system initiated by authors	N/A
Editorial workflow configurable per publication	N/A
Automated email alerts to authors	N/A
Automated email alerts to editors	N/A
Automated email alerts to reviewers	N/A
Style sheets, customizable look and feel per publication	HathiTrust maintains a consistent interface across all content, but certain collections can be branded according to user preferences. The University of Michigan Press collection is an example: https://babel.hathitrust.org/cgi/mb?a=listis;c=622231186.
Versioning	Versions of content are not kept in the repository. When content is modified, the old object is deleted and a new object added with the same identifier. This action is recorded in the PREMIS metadata.

Archiving	HathiTrust provides long-term preservation and curatorial services for deposited content. This includes repository administration, metadata management, content storage, and content migration.

4) Access, formats, and electronic commerce functions

Accessibility of system	The HathiTrust system and interface are designed to provide access to all digitized materials (regardless of copyright) for users with print disabilities, through the registered disability services officer(s) on their campuses, including users with low- to no-vision and learning disabilities. In addition to accessible interfaces for applications that make up HathiTrust (the catalog, Collection Builder, and PageTurner), a text-only interface exists for PageTurner that is optimized for the specific needs of users with print disabilities (including navigation keys, sections markers, OCRed text derived from images, and appropriate use of headings and labels). HathiTrust is additionally configured to grant full-text access to authorized users (to enhance usability with screen readers, digital Braille devices, etc.), regardless of a work's copyright status. HathiTrust is available freely on the web at https://www.hathitrust.org.
Accessibility of document output	---
Internationalization support	Unicode encoding (UTF-8 basic multilingual plane) is supported in repository applications (catalog, PageTurner, Collection Builder, large-scale search indexing) and associated databases.
Output in multiple document formats	HathiTrust delivers content in the user interface as page-images, OCR-text, or in PDF format.
Document formats supported	TIFF ITU G4, JPEG2000, UTF-8 text
Plug-in requirements	---
Usability notes	---
Citation linking	Each volume has a permanent URL, formed using the Handle service (https://handle.net/).

OpenURL resolver	Available via SFX.
RSS feed	Not available.
Digital rights management	HathiTrust performs an automated rights evaluation of incoming objects based on bibliographic data. These rights may be manually overwritten after copyright review has taken place, if bibliographic information is updated, or if rights holders open access to volumes. A Creative Commons License Declaration form for opening access to volumes is available at https://www.hathitrust.org/creative_commons_declaration_form. All rights information is stored in a rights database (https://www.hathitrust.org/rights_database).
Full-text search and retrieval	Full-text search of the entire repository via Solr
Federated searching	Federated search of catalog metadata via OCLC WorldcCat.
Authentication mechanisms	Authentication is used for two purposes in HathiTrust: personalization services (e.g., the Collection Builder application) and uses or services requiring authorization (staff uses such as access to works for copyright review, services for authorized users with print disabilities, and member download of full-book PDF files). Authentication is handled via SAML 2.0-compliant identity provision at the user's institution, with attributes passed to HathiTrust for authorization for access to members-only services. Non-members may authenticate via Google, Facebook, Twitter, etc., for personalization services only.
Subscription services	There are currently no subscription services in HathiTrust.
Electronic commerce functions	Where users can purchase a copy of a book, a link to "Buy a copy" is listed in PageTurner. Volumes contributed by the University of Michigan Press and Utah State University Press are available for print on demand from the press websites. Public domain volumes digitized from the University of Michigan and University of California are available for print on demand via Amazon.com.
Context-sensitive Help support	Feedback links to HathiTrust User Support are provided on the website. Users are routed to the help services as appropriate (e.g., copyright information, technical services corrections, help with metadata download, etc.).