HATHI TRUST DIGITAL LIBRARY

There is an elephant in the library

INTRODUCTION

The integration of mass book digitization efforts into the practice of academic library management reached an important milestone in November 2008 with the debut of HathiTrust, an academic library digital content repository. HathiTrust's origin was an effort led by the University of Michigan Library to provide a systematic and controlled means of storing, managing, and discovering the millions of files of scanned books created by its participation in the Google Books project. The name of the digital library was inspired by the Hindi word for elephant (hathi), an animal popularly admired for its long memory.

Hathi Trust is a not-for-profit collaborative of academic and research libraries preserving 17+ million digitized items HathiTrust offers reading access to the fullest extent allowable by U.S. copyright law, computational access to the entire corpus for scholarly research, and other emerging services based on the combined collection. HathiTrust members steward the collection — the largest set of digitized books managed by academic and research libraries — under the aims of scholarly, not corporate, interests.

Materials Indexed: Audio, Books, Maps, Music Recordings, Musical Scores, Video

Database Type: Electronic Book Collection, Electronic Journal Collection, Full Text Collection, Index

Interface Language: English

Materials Language: Multiple languages

Broad Category : Area Studies, Arts, Engineering, Ethnic & Gender Studies, Government Information, Humanities, Languages & Literatures, Multidisciplinary, Sciences, Social Sciences

MISSION AND GOALS

Mission

1. The mission of Hathi Trust is to contribute to research, scholarship, and the common good by collaboratively collecting, organizing, preserving, communicating, and sharing the record of human knowledge.

Goals

1. To build a reliable and increasingly comprehensive co-owned (jointly) and co-managed digital archive of library materials converted from the print collections of the member institutions.
2. To develop cost-effective and robust (strong) infrastructure for digital content of value to scholars and researchers, including a variety of formats and born-digital materials.
3. To develop partnerships and services that ensures preservation of the materials in HathiTrust and the entire print and digital scholarly record.
4. To build infrastructure that facilitates cost-effective and productive collaborations among partnering institutions to reduce the cost of securing campus intellectual assets.
5. To define and make available a set of services that supports research using the Hathi Trust corpus (Kosh).
6. To create a technical framework that allows for both central and distributed creation of tools and services.
7. To sustain the HathiTrust enterprise as a “public good” while at the same time defining a set of services that benefits member institutions
8. To reduce long-term capital and operating costs of storage and care of print collections through redoubled (much greater) efforts to coordinate shared storage strategies among libraries.

SERVICES AND PROGRAMS:

1. **HathiTrust Digital Library** preserves and provides lawful access to the 17 million digitized items.
2. **HathiTrust Research Center** offers services that support use of the HathiTrust corpus as a dataset for analysis via text and data mining research.
3. **Shared Print Program** develops a distributed, shared network of print collections with collective print retention.
4. **U.S. Federal Documents Program** expands access to and preserves U.S. federal publications.
5. **Copyright Review Program** review team finds and opens public domain materials in the U.S. and around the world
6. **Page turner mechanism:** HathiTrust supports an application for reading, downloading, and interacting with (e.g., zooming and rotating) texts and images in HathiTrust. The page turner application interfaces with mechanisms such as the Rights Database and Shibboleth (a mechanism for inter-institutional authentication) to provide appropriate access to materials, and integrates with services such as the Collection Builder, full text search, and the bibliographic catalog.

7. **OTHER PROJECTS AND GRANTS**

HathiTrust has participated in pilot projects and grants that have been completed.

- **mPach:** was conceived as a suite of tools to enable direct publishing of open access journals into Hathi Trust' s preservation and access environment.
- **Minnesota Digital Library Image Preservation**
- **Prototype Project (for videos)**
- **Grant Projects (Museum Program)**

FUNCTIONAL OBJECTIVES

Short-term

1. **Branding (overall initiative; individual libraries):** HathiTrust supports branding in the repository in a number of ways:

   The page turner prominently identifies the HathiTrust initiative;

A watermark on every page identifies the digitizing agent; and

A watermark on every page identifies the source library of the print material.

The source of the print material is included in feed of bibliographic identifiers so that institutions can import or update records with this information.

The page turner contains institution-specific branding, identifying to users at partners institutions that their institution is a member of HathiTrust.

2. **Format validation, migration and error-checking:** Format validation and error-checking is performed for all content that enters HathiTrust. Although, to date, no migration of content has been necessary to date, it believe that it have mitigated this need by choosing rich, flexible, standards-based formats. HathiTrust stores a variety of technical and digital preservation metadata along with each object in order to aid in migration should it become necessary. Strategies are in place to ensure and validate the integrity of HathiTrust materials on an ongoing basis.

3. **Development of APIs(application programming interface) that will allow partner libraries to access information and integrate it into local systems individually:** Several APIs have been released for this purpose. Two key examples are a bibliographic API (Bib API), which supports lookup and catalog integration, and a data API (Data API), which provides machine access to the underlying data in a digital object. Information on all modes of content and metadata distribution (including OAI and tab-delimited metadata files) can be found at http://www.hathitrust.org/data.

4. **Access mechanisms for persons with disabilities:** HathiTrust has deployed an accessible interface that uses descriptive labeling, key tabs, and other strategies to facilitate navigation and use by users with print disabilities (e.g., optimized for use with screen readers). HathiTrust has also deployed authorization mechanisms that permit users who are certified as having print disabilities to access the full text of public domain and in copyright volumes in HathiTrust. These mechanisms, which have been deployed at the University of Michigan, are sufficiently generalized to provide access at partner institutions pending agreement on entitlement attributes (to be used in connection with Shibboleth) and institutional policies. A CIC working

group chaired by Mark Sandler has initiated work to help address these needs.

5. **Public 'Discovery' Interface for HathiTrust:** HathiTrust released a temporary public version of a comprehensive bibliographic search application (i.e., a catalog) in April 2009 and has worked through a collective process to define a HathiTrust view in WorldCat. The WorldCat implementation of the HathiTrust catalog will be released as a pilot in November 2010.

6. **Ability to publish virtual collections:** HathiTrust has created a Collection Builder application that permits individuals to create public (i.e., shared) and private collections. Collection Builder uses Shibboleth authentication for users at partner institutions, but also permits authentication through the University of Michigan "friend" system so that unaffiliated users can create and maintain collections.

7. **Mechanism for direct ingest of non-Google content:** HathiTrust developed automated ingest mechanisms for book and journal content digitized by the Internet Archive in April 2010. A technical and policy framework for ingest of other digitized book and journal content (e.g., digitized by partner institutions) is being finalized currently. When this is complete, routine ingest of partner content will begin.

FUNCTIONAL OBJECTIVES – Long-term

1. **Compliance with required elements in the Trustworthy Repositories Audit and Certification (TRAC) criteria and checklist:** The Center for Research Libraries is conducting an independent assessment of the HathiTrust repository, based largely on the Trusted Repositories Audit and Certification (TRAC) criteria. The assessment is targeted to be complete by the end of 2010. Information about HathiTrust's compliance with TRAC can be found at http://www.hathitrust.org/standards

2. **Robust discovery mechanisms like full-text cross-repository searching:** An initial implementation of full-text search of the entire repository was released on November 19, 2009. The launch of this service represented significant research and development, much of which is

documented on the HathiTrust website at http://www.hathitrust.org/large_scale_search and http://www.hathitrust.org/blogs/large-scale-search.

3. **Development of an open service definition to make it possible for partner libraries to develop other secure access mechanisms and discovery tools:** HathiTrust has created a number of APIs for this purpose, as well as a collaborative development environment for partners to improve existing, and develop new applications.
4. **Support for formats beyond books and journals:** HathiTrust is investigating issues relating to the storage and delivery of electronic publications (in the ePub format in particular) and digital audio and image files (such as maps). Pilot projects in each of these areas are underway.

5. **Development of data mining tools for HathiTrust and use by HathiTrust of other analysis tools from other sources:** HathiTrust has engaged multiple strategies to support data mining in HathiTrust:

    1. Data Distribution: HathiTrust has made sample datasets available to researchers for computational processing and analysis. The purpose of the samples is to give researchers an idea of the structure of the repository ahead of broader distribution of the public domain in HathiTrust (planned for early 2011) and strategy 2 below.
    2. SEASR integration: The SEASR development team is in the process of integrating SEASR into HathiTrust as a proof of concept.

COLLECTIONS - HathiTrust Digital Library allows users to create lists or collections of items in the HathiTrust. These collections can be searched independently of the rest of the repository and are a great way to gather related resources together. Each collection can be kept private or made public so that other users can view them. HathiTrust has many different types of collections publicly available. Some featured collections could include Women Composers Collections, Records of the American Colonies, Islamic Manuscripts, and Ancestry and Genealogy.

TYPES OF COLLECTIONS –

Temporary Collections - Without logging in, you can create temporary collections. These collections have all the same functionality of permanent collections, but are only available until the end of the browser session.

Permanent Collections - If you are logged into an institutional account, you can create permanent collections. Permanent collections will be available any time that you are logged in and you can keep adding items to grow the collection. These collections can be shared via many different social media platforms and can be made public so they can be found on HathiTrust.

Creating Collections - Logged-in users can also create personal "collections" for private or public use by selecting records from search results and then grouping them into a collection. These collections can be saved for subsequent use and may be shared with and accessed by others.

Creating and Adding to Collection - If you want to create a permanent collection you will need to log in with your ID or you can create a temporary collection without logging in.

- Click on the "Collections" or "My Collections" option on the top of any HathiTrust page
- Click on "Create a New Collection"
- You will need to name the collection, give it a description, and if you are logged-in, you can choose where it is public or private.
- Click "add"

Ways to search HathiTrust:

- **Catalog Search**- search catalog by Title, Author, Publication Date
- **Full-text Search**- use keywords to search the full-text of all works in HathiTrust
- **Collection Builder**- search inside collection of materials that you or others create
- **Single-volume Search**- search inside a volume using keywords while viewing the work in the page-viewing application

- **Via UW Libraries**- all Full-view (not search-only) HathiTrust records are also discoverable via the UW Libraries catalog

BENEFITS

1. The primary benefits of partnership are cost-effective long-term preservation and access services for digitized content
2. Long-term commitments on digital content facilitate decision-making about digitization efforts and print collections management
3. For institutions with content to deposit, participation enables immediate preservation and access services, including bibliographic and full-text searching of the materials within the larger HathiTrust corpus, reading and download of content where available, and the ability to build public or private collections of materials
4. Whether depositing content or not, partners have full viewing and downloading abilities for public domain materials and materials for which we have received permissions, as well as specialized access to public domain and in-copyright materials for users who have print disabilities.

RAKHI SINGH