HathiTrust: Ten Years, 16 Million Volumes, and the Road Ahead

2018 LIBRARY TECHNOLOGY CONFERENCE
JOHN BUTLER, J-BUTL@UMN.EDU
UNIVERSITY OF MINNESOTA
Acknowledgements

Many thanks to Mike Furlough, Sandra McIntyre, Heather Christenson, Lizanne Payne, Angelina Zaytsev, and other HathiTrust staff for their significant contributions to this presentation.
Overview

- HathiTrust Overview
- Membership & Organization
- Collections
- Access
- HathiTrust Research Center
- Addressing Big Questions
- Significant Challenges / Opportunities
- What’s Next?
The Name

The meaning behind the name

- Hathi (hah-tee)—Hindi for elephant
- Big, strong
- Never forgets, wise
- Secure
- Trustworthy

Illustration of “Hathi” the elephant from 1895 edition of The Jungle Book found in HathiTrust.
HathiTrust Overview
Google Is Adding Major Libraries to Its Database

By JOHN MARKOFF
and EDWARD WYATT
Published: December 14, 2004

Google, the operator of the world’s most popular Internet search service, announced today that it had entered into agreements with some of the nation’s leading research libraries and Oxford University to begin converting their holdings into digital files that would be freely searchable over the Web.

It may be only a step on a long road toward the long-predicted global virtual library. But the collaboration of Google and research institutions that also include Harvard, the University of Michigan, Stanford and the New York Public Library is a major stride in an ambitious Internet effort by various parties. The goal is to expand the Web beyond its current valuable, if eclectic, body of material and create a digital card catalog and searchable library for the world’s books, scholarly papers and special collections.

Google - newly wealthy from its stock offering last summer - has agreed to underwrite the projects while also adding its own technical abilities to the task of scanning and digitizing tens of thousands of pages a day at each library.
1,000 Years or 7
Preservation + Access
Mission and Purpose

To contribute to research, scholarship, and the common good by collaboratively collecting, organizing, preserving, communicating, and sharing the record of human knowledge.

A trusted digital preservation service enabling the broadest possible access worldwide.

An organization with over 130 research libraries partnering to develop its programs.

A range of transformative programs enabled by working at a very large scale.
# HathiTrust’s Portfolio of Work

<table>
<thead>
<tr>
<th>Collection Development</th>
<th>Preservation</th>
<th>Use</th>
<th>Rights Management</th>
<th>Collection Management</th>
<th>Computational Research</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mass Digitization</td>
<td>TRAC Certification</td>
<td>Discovery</td>
<td>Investigation &amp; Determination</td>
<td>Holdings Analysis</td>
<td>HathiTrust Research Center</td>
</tr>
<tr>
<td>Member Digitization</td>
<td>Integrity Monitoring</td>
<td>Access</td>
<td>Licensing</td>
<td>Shared Print Retentions</td>
<td>Derived Data Releases</td>
</tr>
<tr>
<td>Born Digital</td>
<td>Format Consistency &amp; Migration</td>
<td></td>
<td></td>
<td></td>
<td>Enhancements to the Corpus</td>
</tr>
</tbody>
</table>
HathiTrust Today -- by the numbers

- 16,170,172 total volumes
- 7,904,299 book titles
- 437,039 serial titles
- >1,000,000 U.S. Federal Gov’t Publications
- 5.66 billion pages
- ~2.5 trillion words indexed / tokens computable
- 725 terabytes
- 191 miles
- 6,055,009 vols (37%) open for reading (public domain & CC-licensed)

The collection includes (mostly) published materials in bound form, (mostly) digitized from library collections.
Digitization Sources for HathiTrust Collection

Google: 94.82%
Internet Archive: 3.53%
Member digitized: 1.66%

As of December 31, 2017
# Google Books-HathiTrust Comparison

<table>
<thead>
<tr>
<th></th>
<th>Google Book Search</th>
<th>HathiTrust</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Volumes</strong></td>
<td>&gt;20M? volumes</td>
<td>16.1M volumes, includes Google, IA content, more</td>
</tr>
<tr>
<td><strong>Search</strong></td>
<td>Full-text*</td>
<td>Bibliographic and Full-text</td>
</tr>
<tr>
<td><strong>Data Access</strong></td>
<td>Only through the web interface</td>
<td>Numerous APIs and growing; in WorldCat</td>
</tr>
<tr>
<td><strong>Metadata</strong></td>
<td>“It’s all data” / Google’s black box</td>
<td>12 types: MARC, METS, PREMIS, Rights, etc.</td>
</tr>
<tr>
<td><strong>Enumerations</strong></td>
<td>Each volume = edition</td>
<td>Clearly presented; clarity around “parts-to-whole”</td>
</tr>
<tr>
<td><strong>Rights Management</strong></td>
<td>Google’s black box</td>
<td>Detailed rights mgmt. system; verification work</td>
</tr>
<tr>
<td><strong>Preservation</strong></td>
<td>No long-term assurances</td>
<td>TRAC certified</td>
</tr>
<tr>
<td><strong>User Experience</strong></td>
<td>General audience; largely effective</td>
<td>Oriented towards academic users; HT Research Center</td>
</tr>
</tbody>
</table>

* Full-text searches for “minnesota” in GBS and HT yield 10.4M and 1.2M results, respectively.
Membership & Organization
<table>
<thead>
<tr>
<th>HathiTrust Members.....</th>
</tr>
</thead>
<tbody>
<tr>
<td>Allegheny College</td>
</tr>
<tr>
<td>American University of Beirut</td>
</tr>
<tr>
<td>Arizona State University</td>
</tr>
<tr>
<td>Auburn University</td>
</tr>
<tr>
<td>Baylor University</td>
</tr>
<tr>
<td>Boston College</td>
</tr>
<tr>
<td>Boston University</td>
</tr>
<tr>
<td>Brandeis University</td>
</tr>
<tr>
<td>Brown University</td>
</tr>
<tr>
<td>Bryn Mawr College</td>
</tr>
<tr>
<td>Bucknell University</td>
</tr>
<tr>
<td>Carnegie Mellon University</td>
</tr>
<tr>
<td>Case Western Reserve</td>
</tr>
<tr>
<td><strong>Carleton College</strong></td>
</tr>
<tr>
<td>Claremont Colleges</td>
</tr>
<tr>
<td>Colby College</td>
</tr>
<tr>
<td>Columbia University</td>
</tr>
<tr>
<td>Cornell University</td>
</tr>
<tr>
<td>Dartmouth College</td>
</tr>
<tr>
<td>DePaul University</td>
</tr>
<tr>
<td>Dickinson College</td>
</tr>
<tr>
<td>Duke University</td>
</tr>
<tr>
<td>Emory University</td>
</tr>
<tr>
<td>Getty Research Institute</td>
</tr>
<tr>
<td>George Mason University</td>
</tr>
<tr>
<td>Georgetown University</td>
</tr>
<tr>
<td>Graduate College of the City</td>
</tr>
<tr>
<td>University of New York</td>
</tr>
<tr>
<td>Harvard University Library</td>
</tr>
<tr>
<td>Haverford College</td>
</tr>
<tr>
<td>Indiana University</td>
</tr>
<tr>
<td>Iowa State University</td>
</tr>
<tr>
<td>Johns Hopkins University</td>
</tr>
<tr>
<td>Kansas State University</td>
</tr>
<tr>
<td>Lafayette College</td>
</tr>
<tr>
<td>Library of Congress</td>
</tr>
<tr>
<td><strong>Macalester College</strong></td>
</tr>
<tr>
<td>Massachusetts Institute of Technology</td>
</tr>
<tr>
<td>McGill University`</td>
</tr>
<tr>
<td>Michigan State University</td>
</tr>
<tr>
<td>Montana State University</td>
</tr>
<tr>
<td>Mount Holyoke College</td>
</tr>
<tr>
<td>New Mexico State University</td>
</tr>
<tr>
<td>New York Public Library</td>
</tr>
<tr>
<td>New York University</td>
</tr>
<tr>
<td>North Carolina Central University</td>
</tr>
<tr>
<td>North Carolina State University</td>
</tr>
<tr>
<td>Northeastern University</td>
</tr>
<tr>
<td>Northwestern University</td>
</tr>
<tr>
<td>Oklahoma State University</td>
</tr>
<tr>
<td>Ohio State University</td>
</tr>
<tr>
<td>Pennsylvania State University</td>
</tr>
<tr>
<td>Princeton University</td>
</tr>
<tr>
<td>Purdue University</td>
</tr>
<tr>
<td>Rutgers University</td>
</tr>
<tr>
<td>Smith College</td>
</tr>
<tr>
<td>Southern Methodist University</td>
</tr>
<tr>
<td>Stanford University</td>
</tr>
<tr>
<td>State University System of Florida</td>
</tr>
<tr>
<td>Swarthmore College</td>
</tr>
<tr>
<td>Syracuse University</td>
</tr>
<tr>
<td>SUNY Buffalo</td>
</tr>
<tr>
<td>Temple University</td>
</tr>
<tr>
<td>Texas A&amp;M University</td>
</tr>
<tr>
<td>Texas Christian University</td>
</tr>
<tr>
<td>Texas Tech University</td>
</tr>
<tr>
<td>Tufts University</td>
</tr>
<tr>
<td>Tulane University</td>
</tr>
</tbody>
</table>
...More HathiTrust Members

<table>
<thead>
<tr>
<th>Union College</th>
<th>University of Illinois Chicago</th>
</tr>
</thead>
<tbody>
<tr>
<td>Universidad Complutense de Madrid</td>
<td>University of Illinois at Urbana</td>
</tr>
<tr>
<td>University of Alabama</td>
<td>Champaign</td>
</tr>
<tr>
<td>University of Alberta</td>
<td>The University of Iowa</td>
</tr>
<tr>
<td>University of Arizona</td>
<td>University of Kansas</td>
</tr>
<tr>
<td>University of British Columbia</td>
<td>University of Maryland</td>
</tr>
<tr>
<td>University of Buffalo</td>
<td>University of Mass. Amherst</td>
</tr>
<tr>
<td>University of Calgary</td>
<td>University of Miami</td>
</tr>
<tr>
<td>University of California</td>
<td>University of Michigan</td>
</tr>
<tr>
<td>Berkeley</td>
<td>University of Minnesota</td>
</tr>
<tr>
<td>Davis</td>
<td>University of Mississippi</td>
</tr>
<tr>
<td>Irvine</td>
<td>University of Missouri</td>
</tr>
<tr>
<td>Los Angeles</td>
<td>University of Nebraska-Lincoln</td>
</tr>
<tr>
<td>Merced</td>
<td>University of Nevada-Las Vegas</td>
</tr>
<tr>
<td>Riverside</td>
<td>University of New Mexico</td>
</tr>
<tr>
<td>San Diego</td>
<td>University of North Carolina at</td>
</tr>
<tr>
<td>San Francisco</td>
<td>Chapel Hill</td>
</tr>
<tr>
<td>Santa Barbara</td>
<td>University of Notre Dame</td>
</tr>
<tr>
<td>Santa Cruz</td>
<td>University of Oklahoma</td>
</tr>
<tr>
<td>California Digital Library</td>
<td>University of Oregon</td>
</tr>
<tr>
<td>The University of Chicago</td>
<td>University of Pennsylvania</td>
</tr>
<tr>
<td>University of Connecticut</td>
<td>University of Pittsburgh</td>
</tr>
<tr>
<td>University of Delaware</td>
<td>University of Queensland</td>
</tr>
<tr>
<td>University of Houston</td>
<td></td>
</tr>
<tr>
<td>University of Rochester</td>
<td>University of Tennessee, Knoxville</td>
</tr>
<tr>
<td>University of Texas</td>
<td>University of Utah</td>
</tr>
<tr>
<td>University of Vermont</td>
<td>University of Virginia</td>
</tr>
<tr>
<td>University of Virginia</td>
<td>University of Washington</td>
</tr>
<tr>
<td>University of Wisconsin-Madison</td>
<td>University System of Georgia</td>
</tr>
<tr>
<td>University of Wyoming</td>
<td>Utah State University</td>
</tr>
<tr>
<td>Vanderbilt University</td>
<td>Wake Forest University</td>
</tr>
<tr>
<td>Virginia Commonwealth University</td>
<td>Washington University</td>
</tr>
<tr>
<td>Virginia Tech</td>
<td>Washington State University</td>
</tr>
<tr>
<td>Wake Forest University</td>
<td>Wesleyan University</td>
</tr>
<tr>
<td>Washington University</td>
<td>West Virginia University</td>
</tr>
<tr>
<td>Washington State University</td>
<td>Williams College</td>
</tr>
<tr>
<td>Wesleyan University</td>
<td>Wichita State University</td>
</tr>
<tr>
<td>Williams College</td>
<td>Yale University Library</td>
</tr>
</tbody>
</table>
Expectations of members

Required

◦ Submission of holdings data on physical collections.
◦ Implementation of SAML-based authentication system for access services.
◦ Annual fees

Not required, but encouraged:

◦ Deposit of collections
◦ Participations in working groups, governance
Membership

Membership available to academic/research libraries.
  ◦ All members have a specific user community that they support, e.g., university libraries.

Member fees support 100% of programs and operations.

2018 fees begin at about $9,500 US/year.
  ◦ All members pay an equal share of cost for open content.
  ◦ Members pay a proportional share for in copyright materials
    ◦ Based on the overlap between physical collection/HathiTrust.

Membership is not synonymous with “subscription.”
Focus is on cooperative efforts and cooperative benefits.
Cooperative Work

We draw upon distributed expertise among members

<table>
<thead>
<tr>
<th></th>
<th>Michigan</th>
<th>Indiana</th>
<th>Illinois</th>
<th>California</th>
</tr>
</thead>
<tbody>
<tr>
<td>Administration</td>
<td>✔</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Preservation &amp; Access Repository</td>
<td>✔  ✔</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Research Center</td>
<td>✔  ✔</td>
<td>✔</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Metadata Management (Zephir)</td>
<td></td>
<td></td>
<td>✔</td>
<td>✔</td>
</tr>
</tbody>
</table>
Members Govern HathiTrust

Board of Governors

Program Steering Committee

Executive Director

Committees and Working Groups

Operations
Collections
10-Year HathiTrust Growth

View Status/Copyright Status in HathiTrust Collection

- Limited View: 62.76%
- Full View: 37.24%

- Public Domain: 17.88%
- US Fed Docs: 6.23%
- Public Domain in the US: 12.92%
- Open Access/Creative Commons: 0.22%
Distribution by Pub Date/Rights Status in HathiTrust

- PD/OPEN
- IC/LIMITED
In 2011, the HathiTrust membership voted to:

Facilitate collective action to create a comprehensive digital corpus of U.S. federal publications including those issued by GPO and other federal agencies.
US Federal Documents Program

Focus:
- Expanded coverage & enhanced access to U.S. federal documents
- First deliberate HathiTrust collection development initiative

Near term activities:
- Developing a registry of US Federal Government Documents
  - Document holdings records of 57 institutions
  - [https://www.hathitrust.org/usdocs_registry](https://www.hathitrust.org/usdocs_registry)
- Digitize! Focus first on known and cataloged materials
  - Gap analysis driven, focused on print, post-1976 materials
- Improve discoverability/findability of collection
  - [https://is.gd/HathiFedDocs](https://is.gd/HathiFedDocs)
US Federal Documents Program

Number of Items in HathiTrust Identified as U.S. Federal Government Documents: 1,116,763

- Full View: 1,000,675
- Limited View: 116,088

Collection has been built mostly via mass digitization, with contributions from more than 50 HathiTrust member libraries
Shared Print Monographs Program

https://www.hathitrust.org/shared_print_program

Focus & Goals

◦ Ensure preservation of print and digital collections
◦ Catalyze national/continental collective management of collections
◦ Commit to retain print holdings that mirror book titles in the HathiTrust digital collection
◦ Maintain a lendable print collection distributed among HathiTrust members
◦ Build on existing arrangements

Original proposal, task force charge, & preliminary recommendations:
◦ https://www.hathitrust.org/print_monograph_archiving
HathiTrust Shared Print Retention Libraries (Phase 1)

Arizona State University
Brandeis University
Brown University
Bryn Mawr College
Claremont Colleges
Colby College
Columbia University
Duke University
Emory University
Georgia Tech University
Getty Research Institute
Harvard University
Indiana University
Iowa State University
Johns Hopkins University
Lafayette College
Massachusetts Institute of Technology
McGill University
New York Public Library
Northwestern University
Ohio State University
Princeton University
Swarthmore College
Tufts University
University of Alberta
University of Calgary
University of California, Merced
University of California, San Diego
University of California, Santa Cruz
University of California Northern Regional Library Facility (NRLF)
University of California Southern Regional Library Facility (SRLF)
University of Chicago
University of Delaware
University of Florida
University of Illinois, Urbana-Champaign
University of Iowa
University of Michigan
University of Minnesota
University of Missouri
University of Notre Dame
University of Pennsylvania
University of Queensland
University of Texas at Austin
University of Virginia
University of Washington
University of Wisconsin-Madison
Washington University in St. Louis
Yale University

49 libraries!
Proposed HathiTrust Shared Print Commitments: Phase 1 (2017)

49 Retention Libraries proposed over 16 million commitments

- 256 million print monographs in HathiTrust member collections
- 145 million print monographs in Retention Library collections
- 58 million of those match HathiTrust
- 16 million print monographs proposed for retention
- 4.8 million distinct OCLC numbers proposed for retention
HathiTrust Shared Print Phase 1
Summer 2016 – December 2017 (18 months)

- Monographs matching HathiTrust
- 49 libraries
- 4.8 million monograph titles
- 16 million monograph volumes
- 25-year retention commitment
- No validation required
- Lend with local ILL

Build on Shared Print Metadata: System Development to Support Shared Print Functions

- Disclose retentions in OCLC WorldCat
- Compare library holdings to HT retentions
- Seek replacements for lost/damaged retention volumes
- Explore resource-sharing options

Overlap and Uniqueness

Based on work presented by John Wilkin in “HathiTrust and Print Storage: Building around a digital core”; http://www.hathitrust.org/documents/HathiTrust-CIC-201105.ppt
HathiTrust Digital Surrogates for Print?

Supplant

Artifactual Value

Resource sharing

Fidelity, Quality & completeness

Full viewability of digital copy

Preservation -- enduring access

Discovery, full-text search

HT Status

Shared Print Fulfillment?

Contractual Restrictions?

Production, Remediation R&D

37%

Supplement

✓

✓
Access
Search HathiTrust's digital library

Search words about or within the items

Advanced full-text search | Search tips

FULL-TEXT  CATALOG

Search

Should I search catalog or full-text?

HathiTrust is a partnership of academic & research institutions, offering a collection of millions of titles digitized from libraries around the world.

WHAT CAN YOU DO WITH HATHITRUST?

BROWSE COLLECTIONS
Explore user-created featured collections.

READ BOOKS ONLINE
Read millions of titles online — like this one!

READ BOOKS ON THE GO
Take the library's books anywhere with our mobile website.

DOWNLOAD BOOKS* & CREATE COLLECTIONS
*requires institutional login
Access in a Nutshell

Anybody anywhere
- Full text search of entire collection (via web)
- Read public domain and open access works (via web)
- Build and share customized collections

Members only
- Download public domain and open access works.
- Replacement access for lost and damaged print copies (in US).
- Access for users who are blind or with print disabilities (where law allows).
Usage of the HathiTrust Collection in 2017

Averaging 22 Million Hits/Month

<table>
<thead>
<tr>
<th>How users arrive</th>
<th>% of Total Sessions</th>
<th>Bounce Rate</th>
<th>Pages / Session</th>
<th>Avg Session Duration</th>
</tr>
</thead>
<tbody>
<tr>
<td>From search engines</td>
<td>45.77%</td>
<td>43.91%</td>
<td>14.10</td>
<td>0:05:32</td>
</tr>
<tr>
<td>Referred from another website</td>
<td>33.05%</td>
<td>13.17%</td>
<td>21.34</td>
<td>0:08:57</td>
</tr>
<tr>
<td>Direct traffic</td>
<td>19.79%</td>
<td>23.79%</td>
<td>16.56</td>
<td>0:08:36</td>
</tr>
<tr>
<td>Via social media links</td>
<td>1.40%</td>
<td>12.10%</td>
<td>16.12</td>
<td>0:05:15</td>
</tr>
<tr>
<td>From email links</td>
<td>0.00%</td>
<td>44.66%</td>
<td>13.64</td>
<td>0:04:21</td>
</tr>
</tbody>
</table>

Usage of the HathiTrust Collection in 2017

Usage of the HathiTrust Collection in 2017

Return visits for Members

Usage of the HathiTrust Collection in 2017

HathiTrust Research Center

BOOKS AS BIG DATA
HathiTrust Research Center

- Enables computational text analysis of works in the HathiTrust Digital Library (HTDL) to facilitate non-profit research and educational uses of the collection.

- HTRC operates under a *non-consumptive research* paradigm:
  - Makes available the collection for computational analysis, while remaining clearly within the bounds of the fair use rights courts have recognized as applying to text analysis.
“Non-consumptive” Research:

No action or set of actions on part of users, either acting alone or in cooperation with other users over duration of one or multiple sessions can result in sufficient information gathered from collection of copyrighted works to reassemble pages from collection.

Definition disallows collusion between users, or accumulation of material over time.
HATHI TRUST CORPUS

CORPUS FEATURE SETS

OTHER TEXT, E.G., DICTIONARIES, WIKI, TWITTER

HTRC @IU @UIUC

TEXT MINING TOOLS

Complexity hiding interface

BIG RED II

BLUE WATERS
The HathiTrust Research Center: Services and Infrastructure

- Persistent and sustainable structure to enable original and cutting edge *non-consumptive* research
- Developed collaboratively by Indiana University and University of Illinois.
  - Additional funding from HathiTrust and foundations
- Analytics Portal
  - [https://analytics.hathitrust.org/](https://analytics.hathitrust.org/)
- Advanced Collaborative Support Programs
  - [https://www.hathitrust.org/hathitrust-research-center-awards-three-acs-projects](https://www.hathitrust.org/hathitrust-research-center-awards-three-acs-projects)
- Dataset distribution:
  - [https://www.hathitrust.org/datasets](https://www.hathitrust.org/datasets)
HathiTrust Research Center Analytics

Supports large-scale computational analysis of the works in the HathiTrust Digital Library to facilitate non-profit and educational research.

https://analytics.hathitrust.org/

Featured Services

Extracted Features
An unrestricted dataset of metadata and word counts for each page in the HathiTrust Digital Library. Download and explore on your own machine.

Text Analysis Algorithms
Web-based, click-and-run tools that perform computational text analysis on worksets, which are user-created collections of volumes. No programming required.

Data Capsules
Secure virtual environments for non-consumptive text analysis, where researchers can implement their own data analysis and visualization tools.
Dr. Alexopoulos, an economist, is using the vast historical record contained in the HathiTrust to study the diffusion of various technologies over time. By tracking word usage trends of 1,214 technology-related terms identified by Alexopoulos, such as the steam engine, her research based on HathiTrust book content has the potential to overturn accepted theories about the economic and societal impacts of a technology.

- 1,012,633 volumes analyzed.
- Over 22 hours of processing using a 32-node cluster on Indiana University’s high-performance supercomputer, Big Red II.
- Each node had 32 cores and 64 GB of RAM.
Women Were Better Represented in Victorian Novels Than Modern Ones

Big data shows that women used to be omnipresent in fiction. Then men got in the way

- U of Illinois English Prof. Ted Underwood and U of Cal-Berkeley Information Science Prof. David Bamman
- Algorithm analyzed the characters and authors of 104,000 novels (1703 to 2009) in HathiTrust.
- Findings: a paradox. As rigid gender roles seemed to dissipate moving into the 20th C., indicating more equality between the sexes, the number of women characters— and proportion of women authors—decreased.
- Published in the journal Cultural Analytics

By Kat Eschner
SMITHSONIAN.COM
FEBRUARY 14, 2018

Source: https://www.smithsonianmag.com/arts-culture/what-big-data-can-tell-us-about-women-and-novels-180968153/
Digging Deeper, Reaching Further
Libraries Empowering Users to Mine the HathiTrust Digital Library Resources

http://teach.htrc.illinois.edu

- Funded by 3-year IMLS Laura Bush 21st Century Librarian grant award (award #RE-00-15-0112-15)

GOALS:
- Arm librarians with instructional content and tool skills in digital scholarship and digital humanities;
- Empower librarians to become active research partners on digital projects at their institutions;
- Enable librarians to build foundations for digital scholarship centers and services
Addressing Big Questions
About copyright....

HathiTrust policies are primarily based on US law...
- Exceptions for “fair use”
- Exceptions for print disabled
- Exceptions for preservation
- Potentially other exceptions to investigate

We respect copyright law in other jurisdictions.
- But we aren’t able to support local copyright laws as easily as we can US laws.
10-year Growth of HathiTrust Collections
(millions of volumes)

Authors Guild v. HathiTrust, 755 F.3d 87 (2d Cir. 2014)
Landmark Court Opinions

Authors Guild, Inc. v. HathiTrust

Authors Guild v. HathiTrust, 755 F.3d 87 (2d Cir. 2014), is a United States copyright decision finding search and accessibility uses of digitized books to be fair use.

The Authors Guild, other author organizations, and individual authors claimed that the HathiTrust Digital Library had infringed their copyrights through its use of books scanned by Google. A federal court ruled against the plaintiffs in October 2012, finding that HathiTrust's use was permissible under fair use. The plaintiffs appealed the decision to the Second Circuit, and were rebuffed in 2014. In an opinion by Barrington Daniels Parker, Jr., the Second Circuit largely affirmed the lower court's findings of fair use for accessibility and search, remanding only to consider whether the plaintiffs had standing to sue about library preservation copies. The remaining claims were settled on January 6, 2015.

Background

The HathiTrust Digital Library (HDL) is a spin-off of the Google Books Library Project. It was founded in 2008 by the Committee on Institutional Cooperation and the University of California system. The collections of these university libraries were digitized by Google and then combined by HDL. The digitization by Google has been the subject of a separate lawsuit.

HDL's main objective is the long-term preservation of the collection. Member libraries may order replacement copies of works if "(1) the member already owned an original copy; (2) the member's original copy is lost, destroyed, or stolen, and (3) a replacement copy is unobtainable at a fair price." The HDL main functionality is full-text search. When search results are found in works in the public domain, the work is displayed online, and so are works for which...
Collective Action: Copyright Review

Systematic manual review of copyright registrations to determine status of portions of the HathiTrust Collection, Supported generously by IMLS. Winner of the 2016 L. Ray Patterson Award.

<table>
<thead>
<tr>
<th>Project</th>
<th>Reviewed</th>
<th>Out of Copyright</th>
</tr>
</thead>
<tbody>
<tr>
<td>CRMS US Pub in US 1923-1963</td>
<td>375,576</td>
<td>203,172 (54.1%)</td>
</tr>
<tr>
<td>(includes US State Documents)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CRMS World Pub in UK (1875-1944), Canada and Australia (1894-1964)</td>
<td>312,149</td>
<td>159,195 (51%)</td>
</tr>
</tbody>
</table>
Service for print-disabled users

Provides eligible users with access to any item in the HathiTrust collection, regardless of copyright status.

Eligibility is determined by the member institution following their own established practices.

Access for the user is managed by a service provider on campus.
HathiTrust at U-M, NFB to make 14M+ books accessible to blind and print-disabled users

June 29, 2016

ANN ARBOR—More than 14 million digital books will soon be made available to blind and print-disabled users, thanks to a new collaboration involving the National Federation of the Blind and the HathiTrust Digital Library, a digital repository hosted at the University of Michigan.

When launched, the program will dramatically increase the availability of books for users who are blind or print-disabled. According to the NFB, currently less than 5 percent of all published works are estimated to be available to the blind, most of which are popular titles.

NFB President Mark Riccobono says the effort will be an important advancement that will specifically benefit print-disabled students and scholars within the academic community.
Significant Opportunities / Challenges

- Strategic Partnerships in the Larger Digital Ecosystem
- Support for New Forms of Scholarship
- Post Mass-Digitization Collection Development and Growth
- Organization Growth
  - Diversification of members – scale, nationality, research institution type
  - Balancing Services to End Users and to Members
- Massive Digital Library Challenges
  - Duplicates
  - Object quality
  - Large-scale search
  - Metadata management – 12 different types of HT metadata
  - Expanding from ‘digitized books’ to ‘born digital text’
  - Harnessing “compute” to help address internal and end-user needs
What’s next?
Six stages of HathiTrust

2002-2006: Prehistory
2006-2008: Moving towards launch
2008-2011: Rapid start-up and organizing
2012-2014: New governance and leadership
2014-2017: Settling in and starting up
2017-: Taking stock and looking out
What is different now?

**HathiTrust**
- Demonstrated exemplar of collective action
- Legal challenges have ended, but some questions remain
- Membership diversification
- Organizational maturity (but still adolescent?)
- Governance is addressing a wider range of challenges

**Digital Library Ecosystem**
- Mass digitization is assumed and non-controversial
- Cooperation and collaboration at scale is proven but still hard
- Large-scale data management is a generalized problem
Strategic Directions

To address significant challenges libraries cannot independently confront to advance innovative forms of research, pedagogy, and public engagement.

*Empower—Enhance—Transform*

- Integral role in advancing research, teaching, and learning
- Enhanced discovery
- Keep collection focus on text-based materials for several years to come.
- Focus on end user experience and services
- More intentional collection development and aggregation
- Expanded shared print services
- More clearly delineated services for member libraries and their users.
THANK YOU and QUESTIONS?

John Butler
Chair, HathiTrust Program Steering Committee
j-butl@umn.edu