Big Questions, Big Data, and HathiTrust

2017 NEW YORK LIBRARY ASSOCIATION
SARATOGA SPRINGS, NY
MIKE FURLOUGH, EXECUTIVE DIRECTOR
10 NOVEMBER
Today’s Plan

What is HathiTrust?

What are our preservation and access services?

How does HathiTrust collectively address large-scale challenges?
  ◦ Copyright “indeterminacy”
  ◦ Print management and preservation
  ◦ Access to US Federal Documents
  ◦ New forms of scholarship

Where are we going next?
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.
<table>
<thead>
<tr>
<th>Allegheny College</th>
<th>George Mason University</th>
<th>North Carolina State University</th>
</tr>
</thead>
<tbody>
<tr>
<td>American University of Beirut</td>
<td>Georgetown University</td>
<td>Northeastern University</td>
</tr>
<tr>
<td>Arizona State University</td>
<td>Graduate College of the City University of New York</td>
<td>Northwestern University</td>
</tr>
<tr>
<td>Auburn University</td>
<td>Harvard University Library</td>
<td>Oklahoma State University</td>
</tr>
<tr>
<td>Baylor University</td>
<td>Haverford College</td>
<td>Ohio State University</td>
</tr>
<tr>
<td>Boston College</td>
<td>Indiana University</td>
<td>Pennsylvania State University</td>
</tr>
<tr>
<td>Boston University</td>
<td>Iowa State University</td>
<td>Princeton University</td>
</tr>
<tr>
<td>Brandeis University</td>
<td>Johns Hopkins University</td>
<td>Purdue University</td>
</tr>
<tr>
<td>Brown University</td>
<td>Kansas State University</td>
<td>Rutgers University</td>
</tr>
<tr>
<td>Bryn Mawr College</td>
<td>Lafayette College</td>
<td>Smith College</td>
</tr>
<tr>
<td>Bucknell University</td>
<td>Library of Congress</td>
<td>Southern Methodist University</td>
</tr>
<tr>
<td>Carnegie Mellon University</td>
<td>Macalester College</td>
<td>Stanford University</td>
</tr>
<tr>
<td>Case Western Reserve</td>
<td>Massachusetts Institute of Technology</td>
<td>State University System of Florida</td>
</tr>
<tr>
<td>Claremont Colleges</td>
<td>McGill University</td>
<td>Swarthmore College</td>
</tr>
<tr>
<td>Colby College</td>
<td>Michigan State University</td>
<td>Syracuse University</td>
</tr>
<tr>
<td>Columbia University</td>
<td>Montana State University</td>
<td>SUNY Buffalo</td>
</tr>
<tr>
<td>Cornell University</td>
<td>Mount Holyoke College</td>
<td>Temple University</td>
</tr>
<tr>
<td>Dartmouth College</td>
<td>New Mexico State University</td>
<td>Texas A&amp;M University</td>
</tr>
<tr>
<td>DePaul University</td>
<td>New York Public Library</td>
<td>Texas Christian University</td>
</tr>
<tr>
<td>Dickinson College</td>
<td>New York University</td>
<td>Texas Tech University</td>
</tr>
<tr>
<td>Duke University</td>
<td>North Carolina Central University</td>
<td>Tufts University</td>
</tr>
<tr>
<td>Emory University</td>
<td></td>
<td>Tulane University</td>
</tr>
<tr>
<td>Getty Research Institute</td>
<td></td>
<td>Union College</td>
</tr>
</tbody>
</table>
...More HathiTrust Members

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

Membership available to academic/research libraries.

Member fees support 100% of operations.

2018 fees begin at about $9,500.

- 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.”
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>
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>Discovery • Catalog • Full Text • Discovery services</td>
<td>Access • Print Disabled Services • Differs for members</td>
<td></td>
<td>Enhancements to the Corpus</td>
</tr>
</tbody>
</table>
HathiTrust Collections Today

15.9 million total digitized items (volumes)
  ◦ 7.8 million book titles
  ◦ 430,000 serial titles
  ◦ Over 1 million US federal government documents

5.96 million items open for reading
  ◦ Includes public domain & CC-licensed items

The collection includes (mostly) published materials in bound form, (mostly) **digitized from library collections.**
Growth of HathiTrust Collections (millions of volumes)
Titles by Language

- English: 50%
- German: 9%
- French: 7%
- Spanish: 7%
- Russian: 3%
- Chinese: 3%
- Japanese: 3%
- Italian: 2%
- Portuguese: 1%
- Arabic: 1%
- 451 other languages: 13%

10 November 2017
Volumes by View/Rights Status

Limited View: 62.75%
Full View: 37.25%

Public Domain: 17.97%
US Fed Docs: 6.02%
Public Domain US: 13.05%
Open Access/Creative Commons: <0.01%
Access in a Nutshell

Anybody anywhere

- Full text search of entire collection (via web)
- Text and data mining
  - In copyright data mining is in pilot mode (via HTRC)
  - Services require additional registration
- Read public domain and open access works (via web)

Members only

- Download public domain and open access works
- Replacement access for lost and damaged print copies.
- Access for users who are blind or with print disabilities
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.
Addressing Big Questions
Collective Action: Copyright Review

Copyright Review Management System

- 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</td>
<td>375,576</td>
<td>203,172 (54.1%)</td>
</tr>
<tr>
<td>Pub in US 1923-1963</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(includes US State Documents)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CRMS World</td>
<td>312,149</td>
<td>159,195 (51%)</td>
</tr>
<tr>
<td>Pub in UK (1875-1944), Canada and Australia (1894-1964)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Shared Print Monographs 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](https://www.hathitrust.org/print_monograph_archiving)
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 Retention Libraries

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

49 libraries!
US Federal Documents Program

Focus:
- expanded coverage & enhanced access to U.S. federal documents.

Near term activities:
- Developing a registry of US Federal Government Documents
  - [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)
Selected Documents Priorities

Acquire comprehensive runs of essential titles:
- Congressional Serial Set
- Congressional Record [bound]
- Statistical Abstract of the United States
- Economic Report of the President
- Foreign Relations of the United States
- United States Reports
- Large statistical yearbooks, such as Agricultural Statistics, etc.
- Agency annual reports
- All hearings

Acquire key agencies:
- EPA
- NOAA
- NASA
- Central Intelligence Agency
- Bureau of Indian Affairs

Acquire agencies within:
- Department of Transportation
- Department of the Interior
- Department of Agriculture
“Non-consumptive” Research: The HathiTrust Research Center

HathiTrust Research Center
- 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)
Tracking Technology Diffusion Through Time in the HathiTrust Corpus
Michelle Alexopoulos, University of Toronto

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.

Literary Geography at Scale
Matthew Wilkens, University of Notre Dame

With the help of natural language processing, Dr. Wilkens will extract and geocode place names in nearly eleven million volumes from the HathiTrust, including twentieth and twenty-first century texts. This project to geolocate world literature, supported by a 2014-2015 American Council for Learned Societies Digital Innovation Fellowship, is one of the largest humanities text-mining projects to date.

Example Advanced Collaborative Support Projects

HTRC Use Case: Collaboration between Scholars and the HTRC

Collaborating with scholars
Research Datasets

Downloadable, non-consumptive book data.

HTRC Extracted Features Dataset

Page-level features from 13.7 million volumes [v.1.0]

Description

The HTRC Extracted Features Dataset v.1.0 is comprised of page-level features for 13.7 volumes in the HathiTrust Digital Library. This version contains non-consumptive features for both public-domain and in-copyright books.

Features include part-of-speech tagged term token counts, header/footer identification, marginal character counts, and much more.

A full explanation of the dataset’s features, motivation, and creation is available at the EF Dataset documentation page

Download the data

All 13.7 million files as well as custom subsets of the EF data are accessible using rsync, as described in the documentation.

A sample is available for download through your browser – sample.zip – as well as thematic collections: DocSouth (87 volumes), EEBO (355 volumes), ECCO (505 volumes).

Attributes

Digging Deeper, Reaching Further: Libraries Empowering Users to Mine the HathiTrust Digital Library Resources

Modern academic librarianship demands reconfigured skillsets and expertise to meet the rapidly evolving, data-driven needs of today’s students, faculty, and researchers as they pursue digital scholarship and produce immense amounts of research data. These rapid developments in user needs and research services speak to an urgent need to provide training for academic librarians to support digital scholarship. We seek to address this need in “Digging Deeper, Reaching Further: Libraries Empowering Users to Mine the HathiTrust Digital Library Resources,” a three-year project funded by a Laura Bush 21st Century Librarian grant award from the IMLS. This project will leverage tools and data from HathiTrust Research Center.
HathiTrust Research Center

Register now for the

2018 HathiTrust Research Center UnCamp

January 25 and 26 2018
University of California Berkeley

For details
Google “HTRC UnCamp 2018”
What next?
Six stages of HathiTrust

2002-2006: Prehistory
2006-2008: Moving towards launch
2008-2011: Rapid start-up and organizing
2012-14: New governance and leadership
2014-17: Settling in and starting up
2017-: Taking stock and looking out
What is different now?

The World – because we changed it.

HathiTrust
- Membership diversification
- Organizational maturity (but still adolescent?)
- Governance is addressing a wider range of challenges

Ecosystem
- Mass digitization is assumed and non-controversial
- Legal challenges have ended, but questions remain
- Cooperation and collaboration at scale is proven but still hard
- Data management is a generalized problem
2015 Collections Survey: Findings

Don’t mess up what you already do well!

Keep building the collections and do so faster.

No strong impetus to expand collecting focus.

Quality is important!

DECISION: Keep collection focus on ”books” and “book-like” materials for several years; subject to revision.
Charting Strategic Directions

Collections Priorities Survey (2015)
Planning to plan (2016)
Board/Program Steering Committee Retreat (June 2017)
Member outreach and engagement (August-October)
Member Meeting (November 2 2017)
Wrap up and finalization (November - March)
Some likely strategic priorities

Grow—Enhance—Impact

- More intentional collection aggregation.
- Enhanced discovery.
- Expanded shared print services.
- More clearly delineated services for member libraries and their users.

Focus: make libraries more powerful and more effective through cooperative programs and service.
THANK YOU and QUESTIONS?

Mike Furlough

furlough@hathitrust.org